Johner Institut

Villa Rheinburg
Reichenaustr. 1

78467 Konstanz, Germany

Guideline for Manufacturers of Medical Devices using Machine Learning

2019-11-22

Compiled by:

Prof. Dr. Christian Johner
Johner Institute
christian.johner@johner-institute.com

Christoph Molnar
LMU Munich
christoph.molnar@gmail.com

AI Guideline for Medical Devices

A) Meta information

1. Objective of the guideline

The objective of this guideline is to provide medical device manufacturers and notified bodies instructions and to provide them with a concrete checklist to

- understand what the expectations of the notified bodies are,
- to promote step-by-step implementation of safety of medical devices, that implement artificial intelligence methods, in particular machine learning,
- to compensate for the lack of a harmonized standard (in the interim) to the greatest extent possible.

The guideline is **not** meant to serves as a training manual or guideline to achieve the safety of AI based medical devices. It is to be a guideline for its review.

The annex lists the recitals which led to the development of this guideline.

2. Scope of applicability and target group.

This guideline is only applicable to medical devices that use AI methods, in particular machine learning. The guideline applies in particular to

- Manufacturers of these products
- Their service providers (such as engineering providers)
- People and organizations that must assess the safety of these products, such as auditors, authorities and notified bodies.

3. Instructions for use

a) Structure of the guideline

This guideline follows the thought that the safety of AI based medical products can only be achieved through a process-oriented approach, whereby all relevant processes and phases of the life cycle must be considered such as:

1. Research and development
2. Data management
3. Post-market surveillance

Accordingly, the guideline does not set forth specific requirements for the products, but for the processes. It contains the following chapters:

- General requirements
- Requirements for product development
    - Intended use
    - Software requirement specification
    - Data management

    - Model development
    - Product development
    - Product release
- Requirements for phases following development

b) Binding character of the guideline

This guideline is neither a legal requirement nor a harmonized standard. Accordingly, there is no differentiation between normative and informative elements.

Much more, the guideline brings together best practices to best describe the required "state-of-the-art".

Some of these best practices are not applicable in all situations, for all products or for all methods of machine learning. The manufacturers should at least justify non-obvious exclusions.

c) Use of the guideline

Creating and reviewing specifications

The manufacturers should first use the guideline to review the completeness of the specifications (process and work instructions, checklists, etc.). These tasks are normally assumed by the following roles:

- Process-related persons, in particular head of development
- Quality manager and quality management deputy
- Regulatory affairs manager

Assessing products and QM system

Then the people responsible for the following tasks should use the guideline:

- Reviewing the conformity of products with the underlying safety and performance requirements
- Assessing the conformity of the technical documentation with the regulatory requirements
- Assessing the efficacy of the internal quality management system (e.g. for design reviews or audits)

The following roles are normally responsible for these tasks:

- Quality managers
- External and internal auditors (including notified bodies)
- Internal and external reviewers of technical documentation (including notified bodies and authorities)
- Testers
- Data scientists
- Clinical affairs specialists
- Regulatory affairs manager
- Risk manager

d) Structure of the guideline

The guideline is a grouped list of review criteria according to the aforementioned chapter. Each list element contains the following attributes:

- ID
- Review criteria, that can be simply and binarily assessed as met or not met
- Comments

The comments may contain:

- Note or reference to regulatory requirement
- Recommendations for implementation of reference to additional explanations
- Tips for auditors such as how the fulfillment of the criteria can be assessed
- Notes on binding nature and applicability and limitations

4. Authors and rights of use

The following authors created this guideline:

- Prof. Dr. Christian Johner (Johner Institute)
- Christoph Molnar (LMU Munich)
- Dr. Andreas Purde, Dr. Abtin Rad (TÜV SÜD)
- Prof. Dr. Christian Dierks (Dierks + Company)
- Stefan Bunk (CTO) and Sven Piechottka (Government & Regulatory Affairs) (Merantix)

The guideline is published under the (Creative Commons License) of type BY-NC-SA. This requires a list of names of authors ("Christian Johner, Christoph Molnar et al.") and allows third parties to build on this work, e.g. to correct it; however only for non-commercial purposes.

The license permits using the product for commercial consultancy purposes including audits and reviews. It is prohibited, however, to use this work itself in an unchanged or changed version for commercial purposes, e.g. in the form of sale as a brochure.

5. Document handling, document identification

This document is managed via the version management system git or the GitHub platform. Only the documents listed in this repository are valid.

The version history, including any authors may be found in the document history.

The released versions are labeled via one day as such in the repository. Versions without a tag are documents in the draft stage.

B) General requirements

1. Processes

The manufacturers should cover all aspects listed below either in the procedural instructions or in the relevant plans to ensure that the safety of the product is systematically guaranteed. Normally, the following standard operating procedures or plans are affected:

- Development
- Risk management
- Data management
- Verification or validation (if not part of development)
- Post-market surveillance and vigilance

- Service, installation, decommissioning
- Customer communication
- Management review (ISO 13485:2016 requires consideration of “applicable new or revised regulatory requirements”.)

If the manufacturer outsources processes, the requirements apply accordingly. Examples would be a (software) development service provider or contract research organization to be required to consider the relevant chapters of this guideline.

2. Competencies

The manufacturers must ensure and prove that they have sufficient competencies to ensure the relevant safety and performance of the products according to the state of the art. This proof is often gained most easily through internal or external training.

Manufacturers can use the competency of external resources.

Requirements	**Comments**
The manufacturer has created a list of all roles that are directly or indirectly concerned with AI [1]	
The manufacturer has determined the competencies for each role in relation to AI [2]	Examples of competencies: Machine learning, explainable AI, medicine (for relevant domains), clinical and usability validation
The manufacturer has appropriate records for the training, further education and competencies that allow for the conclusion that the persons actually have these competencies	
The (software) development plans lay out the product-specific competencies (beyond or deviating)	Requirements of ISO 13485:2016

3. Documentation

The manufacturers should keep evidence that they have followed the relevant requirements of this guideline. There are no specific requirements for documentation and “objective evidence”.

In Europe at least, there is no obligation to create a specific document that summarizes the activities especially for AI. Manufacturers can integrate these aspects into existing documents such as QM documents (e.g. standard operating procedures, work instructions) and the technical documentation (e.g. software files, risk management files, clinical evaluation, summative evaluation of the usability).

[1] Examples are: Data Scientists, Developers, Testers, Regulatory Affairs and Quality Mangers, Service and Support Employees, Product Managers, Medical Device Consultants, Physicians
[2] Competencies (level if understanding, capability to perform tasks) should be listed, not primary subjects

C) Requirements for product development

1. Intended use and stakeholder requirements

a) Intended use

Requirements	Comments
The manufacturer has determined for which medical purpose (diagnosis, therapy, monitoring predictions) the medical product should support.	The intended use / purpose should not be mistaken for the description of functionality (e.g. calculation of scores).
The manufacturer has characterized the patients to be diagnosed, treated or monitored with the medical product. This characterization includes indications, contraindications and associated diseases.	This characterization is also included in IEC 62366-1. Patients may also simultaneously be users of the product.
The manufacturer has set forth on which body locations the product will be used or from which body location the data originate.	Also called for in IEC 62366-1.
The intended use specifies the goal of machine learning.	Classification and regression, clustering, similarity search and recommender systems are the typical goals of methods of machine learning.

b) Intended users, intended use context

Requirements	Comments
The manufacturer has characterized the intended users, e.g. using demographic features (age, gender), regarding the training, experience in medical domains, regarding technical knowledge, physical and mental limitations, linguistic skills and cultural background.	If the manufacturer does not foresee any limitations regarding these attributes, it must document this.
The manufacturer has characterized the intended use environment, e.g. using physical properties (brightness, volume, temperature, contamination, moisture), using the social environment (stress, shift work, frequently changing colleagues) as well as further parameters (such as wearing gloves)	Also included in IEC 62366-1.

c) Stakeholder requirements

Requirements	Comments

The manufacturer has operationalized the goals listed in the intended use with quantitative values [3].	It is not unusual that these values are supplemented and revised during the course of development.
The manufacturer has set forth the runtime environment of the product regarding hardware (screen size, screen resolution, storage, network connection etc.) and software (e.g. operating system, browser, runtime environments such as Java runtime environment or .NET).	For apps, this characterization must be done for the app and for the server part.
The manufacturer has specified the data interfaces using the levels of the interoperability model and set forth the formats and for images, their specific properties (size, resolution, color coding).	This is required pursuant to IEC 62304 chapter 5.2.2.
The manufacturer has specified the requirements for input data.	Input data can be dependent on the method of data generation e.g. recording procedure, technical parameters such as magnetic field strength, number of deflection electrodes, direction and environmental conditions of recordings, manufacturer, medical product etc.
The manufacture has set forth all markets and all regulatory requirements relevant to these.	Show this list.
The manufacturer has determined whether the system should learn further after it goes to market. If this is the case, the manufacturer has shown whether this continual training will occur globally/centrally or decentralized, e.g. per product or per hospital.	

d) Input for risk management and clinical evaluation

Requirements	**Comments**
The manufacturer has listed alternative methods and assessed them with regard to benefits, safety and performance.	Discussion of the state-of-the-art is a requirement of MEDDEV 2.7/1 and MDR/IVDR.
The manufacturer has compared the aforementioned quantitative values with the relevant values of alternative methods.	Manufacturers should create a tabular overview.

[3] **Example**: Purpose: The software supports radiologists in diagnosing cancers using CT images of the head. Quantitative value: 95% of radiologists working with software detect the cancer.

The manufacturer has justified why machine learning is superior to the other methods and thus justified the associated risks.	
The manufacturer has created a list of risks specifically associated with the use of the method of machine learning.	Is part of risk management file
The manufacture has analyzed the risks arising if persons other than the specified users use the product.	
The manufacturer has analyzed the risks arising through use in an environment different than that specified.	
The manufacturer has analyzed the risks arising from inputs not in the specified format.	
The manufacturer has analyzed the risks arising from data that was not generated under the specified conditions.	
The manufacturer has assessed the risks if the system is used for another patient populations than that specified.	

2. Software requirements

a) Functionality and performance

Requirements	**Comments**
The manufacturer has derived traceable quantitative quality criteria and requirements for the software and/or the algorithm from the intended use [4].	This traceability is shown particularly well with a traceability matrix.
The manufacturer has considered the following quantitative quality criteria: For classification problems accuracy (mean or balanced accuracy), positive predictive value (precision), specificity and sensitivity: for regression problems mean absolute error and mean square error.	For unbalanced data, meaning if labels occur at very different frequencies, balanced instead of mean accuracy must be used. The selection of quality criteria strongly depends on intended use.
The manufacturer has specified the expected value ranges.	

[4] Examples: **Example 1**: The stakeholder requirement states that 95% of radiologists must be able to detect a cancer with the product. The requirement of the algorithm states that it must display a sensitivity of 97%. **Example 2**: The stakeholder requirements state that arterial calcification must be able to be detected at a sensitivity of 92%. The requirements of the algorithm state that it must be able to exactly predict the strength of the plaques in the blood to 0.2 mm.

The manufacturer has specified the requirements regarding repeatability and reproducibility of requirements.	This is particularly relevant with "Continuous Learning Systems".
The manufacturer has determined how the system behaves if the inputs do not meet the specified requirements[5].	This is an aspect of robustness, which must be specified pursuant to ISO 25010 and IEC 62304 chapter 5.2.
The manufacturer has determined which self-tests the system must perform and how it behaves if this is not successful.	This is particularly relevant for "Continuous Learning Systems".
The manufacturer has determined how fast the system must create the outputs.	This determination may be done depending on the size and amount of data.
The manufacturer has specified the availability of the medical device.	This is an aspect of robustness and must be specified pursuant to ISO 25010 and IEC 62304 chapter 5.2.

b) User Interface

Requirements	**Comments**
The manufacturer has specified what the user interface must display if the conditions are not met[6], to operate the system safely (e.g. invalid or unexpected inputs).	
The manufacturer has specified what the user interface must display if the output does not meet the specified quality criteria.	
The manufacturer has determined whether there is a need for instructions for use and training materials.	The MDR / IVDR allow exceptions from the obligation.

c) Additional software requirements

Requirements	**Comments**
The manufacturer has set forth which requirements the system must fulfill to detect system errors.	Could be an audit log or a monitoring port.
The manufacturer has checked that the patients are not exposed to decisions through the specified system that are exclusively based on automatic data processing.	Requirements of Art. 22 of the GDPR.

[5] Examples: incomplete data sets, lack of data sets, wrong data format, excessive data quantities, data outside of specified value ranges, wrong temporal sequence of data.
[6] Examples: incomplete data sets, lack of data sets, wrong data format, excessive data quantities, data outside of specified value ranges, wrong temporal sequence of data.

d) Risk management and clinical evaluation

Requirements	Comments
The manufacturer has assessed the risks arising if the inputs do not meet the specified requirements[7].	
The manufacturer has derived the quantitative quality criteria using the state of the art.	The manufacturer must list the quality criteria for alternative technologies and methods and be able to argue if the medical product is not superior to alternatives with regard to quality criteria[8].
The manufacturer has set the gold standard and justified its choice, with which the quality criteria can be reviewed.	
The manufacturer has analyzed the risks arising if the outputs do not meet the specified quality criteria.	
The manufacturer has assessed the consequences if the system provides socially unacceptable outputs (e.g. discriminatory).	These "consequences" are not necessarily risks in terms of ISO 14971.
The manufacturer has assessed the risk arising if the system is unavailable.	
If the manufacturer uses self-tests, it has shown which of the specific quality criteria will be reviewed and which risks are managed by this.	
With Continuous Learning Systems, the manufacturer has considered the option of resetting the system to a known status.	Check risk table.
With Continuous Learning Systems, the manufacturer has shown quantitatively why the risk-benefit analysis is better than for non-continuously learning systems.	

[7] Examples: incomplete data sets, lack of data sets, wrong data format, excessive data quantities, data outside of specified value ranges, wrong temporal sequence of data.
[8] The state-of-the-art of technology is not necessarily consistent with the state of science and thus the gold standard nor with the "Ground Truth". This means that the system requirements are lower than with a gold standard respectively the "Ground Truth". This would be the case in particular if the latter require an invasive or very cost-intensive procedure.

3. Data Management

Data generally have to be understood as training, validation and test data. Each type has to fulfill specific requirements. If not specified differently, however, the usage of the term "data" relates in the following to all three types.

a) Data collection

Requirements	Comments
The manufacturer has set the number of data sets and given a reason why this is sufficient[9].	
The manufacturer has specified which data is required per data set to train the algorithm.	
The manufacturer has characterized the inclusion an exclusion criteria of patient data using relevant attributes[10].	
The manufacturer has specified the technical inclusion and exclusion criteria for data[11].	
The manufacturer has described the procedure by which it ensures that data sets that do not meet the inclusion criteria or should be excluded are actually excluded.	Procedure includes a software supported assessment. This software must be validated.
The manufacturer has described the collected data using descriptive statistics[12].	The "Dataset Nutrition Label" is an recommended option.
The manufacturer has justified where the data are collected and why these are representative for the target population. As reasonable, these have been compared to scientific publications and to registers.	
The manufacturer has listed factors and discussed factors that could cause a bias in the data.	

[9] A specification for the number of data is hardly possible. This depends on the “signal-noise-ratio” among other things. For example, for one data set, the percentage of relevant genes and the strength and frequency of the predicted effects affect the number. For data to be classified, the number of the data sets with the rare class (e.g. the prevalence of diseases) is decisive.

[10] e.g. demographic data (age, gender), physical parameters (height, weight), diseases, vital parameters, lab parameters, presence of additional tests, case history.

[11] Examples: **Example 1**: Patients who must be ruled out due to a heart pacemaker or lung surgery because the images cannot be analyzed or could lead to erroneous classification. **Example 2**: Formats and technical parameters such as image sizes, resolution, brightness and contrasts, color coding, compression, recording equipment, recording method (e.g. CT versus MRI), with or without contrast agent, zoom. **Example 3**: Completeness of meta-data.

[12] Usually, the calculation of distributions (histograms), mean / average values, quartiles, possibly “joint distribution of features". The correlation of data among that data should be examined. Additional examples are found in the publication by Sarah Holland et al. (such as in table 1)

The manufacturer has analyzed which influences the type and location of data collection have on the data[13].	
The manufacturer has established a method by which data are anonymized or pseudonymized before testing and training.	
The manufacturer has examined and excluded the possibility of a "label leakage"[14].	
The manufacturer that uses surveys has justified the selection of the surveys, the time of survey and possibly the method for their assessment, in particular if no standardized survey exists.	

b) Labeling of data

Requirements	**Comments**
The manufacturer using "supervised learning" has derived the labels from the intended use and justified this selection.	
The manufacturer using "supervised learning" has determined a procedure for labeling if no labels were present in the data.	
This procedure specifies quantitative classification criteria for labeling. The selection of these criteria has been justified by the manufacturer[15].	If the "Ground Truth" is not selectedC.3.b.2, because it is too expensive or invasive, this must also be justified.
This procedure specifies the requirements for the number, training and competency for the people responsible for labeling.	
This procedure sets forth how the competencies of the persons responsible for labeling is tested.	This can be done by the labeling of selected data sets.
This procedure sets forth how the persons responsible for labeling are trained and how the success of this training is evaluated.	

[13] Examples: Influence of various measurement devices, surveys, policies (such as a clinic only takes lab parameters only in emergencies, another routinely. Frequency and reason examined with the patient), type of clinic (e.g. small hospital, from which all serious cases must be referred versus university hospital > survivor basis), self-selection bias (e.g. patients with various pre-existing conditions usually go to a hospital rather than a medical practice), type of study (prospective versus retrospective)

[14] These are data in which non-causal information are found in the data via the label, e.g. in the sorting (e.g. first the data of healthy persons, then of ill persons), in the hospital (from one the severe cases originate), in images (e.g. for skin cancer, one must always see a ruler). An additional example would be multiple CT images of a patient, in which the model learns using the patient and not the disease. This could happen if a rib fracture can be seen in addition to the cancer on multiple images.

[15] If, for example, patients have to be classified as healthy and sick, the manufacturer must derive the criteria specifically for the intended use, when a patient is to be classified as healthy and when as sick.

This procedure sets forth how the correctness of the label is systematically reviewed. The selection of this justification has been documented by the manufacturer.	The manufacturer can provide identical data sets of multiple persons and assess the consistency of the results.
This procedure sets forth, how the monitoring occurs, that the persons responsible for labeling are continually fit and willing to perform the labelling [16].	This can be done with datasets with already known labels that are inserted unnoticed by the person during labeling.

c) Procedure for (pre-)processing of data

Requirements	**Comments**
The manufacturer has set a procedure that describes the pre-processing of the data	
This procedure describes the individual processing steps such as conversion, transformation, aggregation, normalization, format conversion, calculation of feature, conversion of numerical data into categories.	A graphic representation creates a rapid overview. The conversion of numerical to categorical values requires a justification.
The procedure describes how the correctness of the interim steps and the final results are assessed[17]. These evaluations are done risk-based.	This is consistent with the requirements of ISO 13485:2016 chapter 4.1.6. The risk management file must contain these analyses.
This procedure specifies how values with various measurement scales or units are detected and processed.	
This procedure specifies how values are detected and processed that have been collected with various measurement methods.	
This procedure specifies how values or metadata with the same names (such as in column headers) are detected and processed.	This, however, depends on the ML method (e.g. tabular data, image data) and cannot be demanded as a general best practice.
This procedure specifies how missing values within data sets are detected and processed. The manufacturer gives a rationale for the decision[18].	Make sure that the rationale differentiates between “missing at random” and “missing not at random”[19].

[16] The labeling of dozens of data sets is arduous. A payment per data set can cause an inappropriate motivation.
[17] Options include software tests and redundant or alternative calculations such as with Excel.
[18] The options for processing include deleting the data set, replacement by the average value of other data sets, new value “missing” (for categorical values).
[19] An example for "missing not at random” is lab values that are too high that are cut off.

This procedure specifies how outliers are detected and processed[20]. The manufacturer gives a rationale for the decision[21].	Show example of a date / feature. This, however, depends on the ML method (e.g. tabular data, image data) and cannot be demanded as a general best practice.
This procedure specifies how unusable data sets are detected and handled[22]. The determination was justified by the manufacturer.	Request example of a date / feature.

d) Documentation and version control

Requirements	Comments
The manufacturer has described the "funnel" that allows detection of how much data originates from which data source (e.g. clinics) and at which processing step how many data sets have fallen away for which reason.	
The manufacturer has described the collected data using descriptive statistics[23].	The "Dataset Nutrition Label" is recommended.
The manufacturer has documented all software for data processing including the libraries used and listed under version control.	

4. Model development

a) Preparation

Requirements	Comments
The manufacturer has justified the selection of the features considered during training.	
The manufacturer has described the dependency of the features among each other.	A Directed Acyclic Graph (DAG) helps in visualization. This, however, depends on the ML method and cannot be demanded as a general best practice.
The manufacturer has documented and justified the ratio that it divides up the data into training, validation and test data.	

[20] The options for processing include deleting the data set, correcting the value, setting the value to a set value (min/max).
[21] This justification is more important in the regression method than with tree-based methods.
[22] Examples are x-rays of poor quality or patients who do not meet the inclusion criteria.
[23] Usually, the calculation of distributions (histograms), mean / average values, quartiles, possibly "joint distribution of features". The correlation of data among that data should be examined. Additional examples are found in the publication by Sarah Holland et al. (such as in table 1)

The manufacturer has documented the stratification it uses to divide up the data in to training, validation and test data[24].	
The manufacturer has documented how it ensures that multiple data sets for an object are in the same "bucket" (training, validation and test data).	
The manufacturer has documented how it ensures that the development team has no access to the test data.	
The manufacturer has described when it recodes the data specifically for the model or specifically for the library[25].	

b) Training

Requirements	Comments
The manufacturer performs model training, tuning of hyperparameters and model selection exclusively with the training and validation data (using cross-validation).	
The manufacturer has documented and justified the choice of the hyperparameters[26].	
The manufacturer has documented and justified the choice of epochs[27].	Where possible, display learning curves.
The manufacturer has determined, documented the quality metrics to which it wants to optimize the model and justified it based on the intended use.	The selection of these quality metrics is specific to the intended use.
The manufacturer has trained multiple models with multiple hyperparameters (including simpler and interpretable models).	

c) Evaluation

Requirements	Comments

[24] For data with rare features or labels, it may be necessary to distribute the data not just at random.
[25] Examples of this are normalization, selection of class labels (e.g. 0 or 1), selection of column names, distribution of categorical values over multiple columns.
[26] Examples: Loss function, optimizer, learning rate, number of epochs
[27] It might be helpful to illustrate the dependency between the quality of the model on the one hand and number of epochs on the other hand e.g. using learning curves. These learning curves, however, exist for neuronal networks and boosting procedures, for example, but not for models with numerical solution (e.g. linear regression) or for a single tree.

The manufacturer has documented the quality metrics for the various models, such as for a binary classification using a confusion table.	This documentation should not include only the values that the manufacturer has used to optimize the model.
The manufacturer has not only globally assessed and documented the quality metrics for the various models, but also separately for various features.	
The manufacturer has examined the data sets that were particularly good and particularly badly predicted.	We recommend a residual analysis in which the errors are listed via the feature values.
The manufacturer has examined the data sets in which the model is particularly secure and particularly insecure[28].	
The manufacturer has justified the ultimate selection of the model using the quality criteria and intended use and in particular shown if simpler and interpretable models were not used.	
The manufacturer has considered (in particular for tabular data sets) to show for individual data sets the feature that the model particularly determined in the decision[29].	This, however, depends on the ML method and cannot be demanded as a general best practice.
The manufacturer has considered to evaluate how and how strongly individual features had to change for the model to come to another prediction.	This is referred to as "Counterfactuals". This, however, depends on the ML method and cannot be demanded as a general best practice.
The manufacturer has analyzed/visualized the dependency (strength, direction) of the prediction of the feature values[30].	This, however, depends on the ML method and cannot be demanded as a general best practice.
The manufacturer has considered (in particular for tabular data sets) to evaluate / visualize the dependency (magnitude, direction) of predictions on feature values	This, however, depends on the ML method and cannot be demanded as a general best practice.
The manufacturer has considered to synthesize data sets that activate the model particularly strong[31].	This, however, depends on the ML method and cannot be demanded as a general best practice.

[28] For classification tasks, the model is particularly insecure with probabilities around 0.5.
[29] Approaches include LIME (Local Interpretable Model-agnostic Explanations), Beta (Black Box Explanations through Transparent Approximations), LRP (Layer-wise Relevance Propagation) and Feature Summary Statistics (incl. Feature Importlands and Feature Interaction.
[30] Examples of Sharpley-Values, ICE-Plots, Partial Dependency Plots (PDP)
[31] For examples see http://yosinski.com/deepvis

The manufacturer has approximated the model using a simplified surrogate model such as a decision tree.	This, however, depends on the ML method and cannot be demanded as a general best practice.

d) Documentation

Requirements	Comments
The manufacturer has the model[32] and/or the training code under version and configuration control.	
The manufacturer can reproduce test and validation results.	This can prompt for version and configuration control of data, test results and assessments.
The manufacturer has the SOUP (libraries and frameworks) under version and configuration control.	
The manufacturer has documented the architecture of the model, the model itself including its hyperparameters.	
The manufacturer has described when it worked with a "pretrained model" and shown why this "pre-training" is suitable for the task.	
The manufacturer has documented the quality of the model based on the quality metrics.	This quality metrics relate to the testing with the test data.
The manufacturer has documented (in particular when using tabular data) the limits (such as feature values) within which the model has achieved the quality metrics.	This, however, depends on the ML method and cannot be demanded as a general best practice.

5. Product development

a) Software development

Requirements	Comments
The manufacturer has performed the required activities pursuant to IEC 62304 and documented them.	Notes for auditors[33]
If the manufacturer has implemented the model in another programming language or for another runtime environment, it has created a plan that repeats the activities pursuant to chapter 4.	
The manufacturer tests the performance (response times, resource consumption) on the target hardware (e.g. browser, mobile device).	
The manufacturer has described how to verify all SOUP or OTS components.	

[32] Trained models can be serialized.

[33] The manufacturers should adhere to the normal best practices such as adherence to coding guidelines, review of code by code reviews using defined criteria, testing to code with unit tests with a defined coverage, etc.

b) Accompanying materials

Requirements	Comments
The instructions for use clearly identify the version of the product.	If possible, indicate the UDI
The instructions for use describe the intended use of the product including the expected medical benefit.	
The instructions for use specifiy the intended patient population using indications, contraindications and if relevant using other additional parameters such as age, gender, accompanying diseases or availability of information.	
The instructions for use explicitly list the patients / data / use case for which the product may not be used.	
The instructions for use document the requirements of the input data (including formats, resolutions, value ranges, etc.).	
The instruction for use specify the intended primary and secondary users pursuant to intended use.	
The instructions for use describe the other conditions applicable to the product (e.g. runtime environment. use environment).	
The instructions for use describe the residual risks.	
The instructions for use indicate the data with which the model was trained.	This is related both to the patient collective and to the features used.
The instructions for use describe the model and algorithms.	
The instructions for use name the quality metrics.	
The instructions for use list the factors that could have a negative effect on the quality metrics.	
The instructions for use specify whether the product is further trained during use.	
The instructions for use describe how updates occur.	
The instructions for use contain references to additional literature.	
The instructions for use contain references to licensing rights.	
The instructions for use identify the manufacturer and lists channels for posing questions.	
The instructions for use list possible ethical problems.	
The instructions for use contain the URL under which the most current versions of the instruction of use can be found.	

c) Usability validation

Requirements	Comments
The manufacturer assesses whether the users understand the instructions for use.	
The manufacturer assesses whether users blindly trust or mistrust the results of the product during usability validation.	
The manufacturer assesses whether the users correctly detect and understand the results during usability validation.	

d) Clinical evaluation

Requirements	Comments
The manufacturer assesses whether the promised medical benefit is achieved with the quality parameters.	
The manufacturer assesses whether the promised medical benefit is achieved is consistent with the state of the art.	

6. Product release

Requirements	Comments
The manufacturer has documented the model using the criteria listed in chapter 5.b).	
The manufacturer has assessed and documented the risks as acceptable in risk management and that all of the activities specified in the risk management plan were performed.	Notes for auditors[34]
The manufacturer has shown in a "Software as a Medical Device Pre-Specifications "(SPS) report which types of changes it anticipates for systems that it wishes to market in the USA[35].	
The manufacturer has shown in Algorithm Change Protocol (ACP) how it will perform these changes for systems that it wishes to market in the USA[36].	
The manufacturer has created a Post-Market Surveillance Plan, see below.	

D) Requirements for phases following development

1. Production, Distribution, Installation

Requirement	Comments

[34] Using examples, check that the efficacy of risk management measures was tested so that there is a traceability of risks for risk control measures.
[35] Changes may affect the intended use, the input data and the clinical and analytical performance.
[36] The approach must, for example, address handling data, re-training, the performance and the updates.

The manufacturer has described how it ensures that only exactly the intended artefacts (files) in exactly the intended version of the product or as a product are delivered	This is configuration management. Also relevant to downloads or AppStores
The manufacturer has described how the persons responsible for installation know which is the most current version and how mistakes in installation can be ruled out	This is only relevant to stand-alone software. A SOP or work instruction would be expected here
The manufacturer has described how one ensures during installation that the requirements specified in the accompanying material are actually fulfilled (see above)	A SOP or work instruction would be expected here
The manufacturer has established procedures that ensure that it can communicate with the operators and users of its product in a timely manner	

2. Post-Market Surveillance

Requirements	**Comments**
The manufacturer has created a Post-Market Surveillance (PMS) Plan	
The manufacturer has specified the data it wishes to collect and analyze in this PMS plan.	
The manufacturer has specified in the PMS plan the quality criteria and threshold values that it considers necessary for handling of in particular a re-evaluation of the risk-benefit analysis.	
The manufacturer has analyzed when determining these threshold values which feedback loops the threshold values can influence[37].	
The manufacturer has analyzed when determining these threshold values which self-fulfilling prophecies the threshold values can influence[38].	
In the PMS plan, the manufacturer described how it collects and analyzes information on adverse medical effects.	
In the PMS plan, the manufacturer described which information on (adverse) behavioral changes or (predictable) misuse is collected and analyzed[39].	
In the PMS plan, the manufacturer described how it collects and analyzes information on additional “adverse effects” [40].	

[37] Examples for these feedback loops: **Example 1**: A travel recommendation app sends targeted advertising depending on feature (last trip). This influences travel behavior. **Example 2**: An algorithm provides prognoses. Therefore, the physician will treat the patients better or earlier...
[38] Example 1 (criminalistics)
[39] Example: Radiologists rely on the software and don't look at the images anymore, so they overlook findings.
[40] Examples would be ethical challenges such as the YouTube algorithm, which achieves the goal, maximizing the click count or use duration, but promoting violence and conspiracy videos.

The manufacturer has described in the PMS plan how it collects information to be able to analyze whether the data in the field is consistent with the expected data or training data[41].	Note for auditorsD.2.6
In the PMS plan, the manufacturer has described how and how often it wants to collect information on whether the product still meets the state of the art.	Note for auditors[42]
In the PMS plan, the manufacturer has described how and how often it wants to collect information on whether the "Ground Truth" or the gold standard are still up to date.	
In the PMS plan, the manufacturer has described how and how often changes pursuant to the Algorithm Change Protocol (ACP) and within the "SaMD Pre-Specifications" (SPS) are made.	

E) Annexes

2. Additional literature

a) Laws

- Medical Device Regulation MDR
- In-vitro Diagnostic Device Regulation IVDR

b) Standards and Best Practice Guides

- **IEC 62304/AMD1**, Medical device software – Software life cycle processes
- **IEC 82304-1**, Health software – Part 1: General requirements for product safety
- FDA Guidance Documents on Machine Learning

b) Industry literature, textbooks

- Christoph Molnar: Interpretable Machine Learning
- Patrick Hall: Machine Learning Interpretability with H2O Driverless AI
- Patrick Hall: On the Art and Science of Machine Learning Explanations
- Johner Institute: Video training on machine learning for medical products

3. Recitals

1. Manufacturers are increasingly developing medical products that use the process of artificial intelligence, in particular machine learning. Many of these procedures are still very new, and lack best practices. This creates new risks for patients, users and third parties.
2. The EU directives (MDR, IVDR) explicitly require the safety of the products in the relevant annexes I. Bu concrete requirements for these classes of products are completely lacking.

[41] One speaks here of a distribution shift or data drift.
[42] For example, have the manufacturer explain the process on how it is systematically informed of new developments in machine learning, and how it assesses these developments and reacts to them.

Therefore, both the manufacturers and the notified bodies and authorities lack concrete guidelines on how to evaluate the safety of the products.

3. Contrary to most other fundamental requirements, no standards on the subject of AI are harmonized. Therefore, there is no canonical catalog of requirements that reflects the recognized state of the art of technology.
4. The FDA has started to formulate requirements on using Continuous Learning Systems, CLS. These specifications are unsuitable to sufficiently set requirements for the products and processes as early as the product development stage.
5. The safety of medical products must be considered in all phases of the product life cycle processes. A limitation to testing is insufficient. This fact must be in line with best practices and guidelines.
6. One hopes that standards for the safety of AI-based medical products will be developed and harmonized. This will still take years. Therefore, we need a guideline (only) in this interim phase.
7. This guideline should be available very soon (by July 2019) to be able to quickly serve the manufacturers as an orientation and enable them to act immediately. The high speed of development makes compromises regarding harmonization with the most parties possible inevitable.
8. The technological advancement in the area of artificial intelligence is immense. New procedures and technologies are continuously being published. On the one hand, a guideline should be a specific as possible. On the other, it cannot be so specifically targeted toward one procedure or technology, to achieve a sensible “shelf life”. Therefore, a guideline must address general concepts. However, it cannot claim to be complete.
9. Such a guideline must take into consideration the specifics of medical products, which includes the principles of patient safety (safety) and a risk-based approach. In a concrete case, selected actions for information security ("controls”) will be in conflict with the fundamental requirements. For this reason, there can be no set list of “controls” for medical products. The manufacturer's intended use of the product is critical.
10. The simple intelligibility and practicability is essential to the desired positive influence of a guideline on the safety of AI-based medical products. Therefore, there must be the least abstract or “high level” requirements possible but “binary decisive” test criteria.
11. Do increase practicability, the authors have avoided collating many requirements to the greatest extent possible. Rather, they have limited themselves to those that they consider particularly relevant and implementable.
12. And to promote distribution and the level of familiarity, the guideline must be available and remain available at no cost.
13. The guideline should be available in German and English.