Validation of Continuously Learning AI/ML Systems in Medical Devices – A Scenario-based Analysis

Haimerl Martin[1]
[1] Hochschule Furtwangen University
Martin.Haimerl@hs-furtwangen.de

Abstract. This paper discusses the use of continuously learning AI/ML based medical devices, i.e. devices which optimize their performance during the product’s lifetime. For such devices, a regulatory strategy was recently proposed by the US Food & Drug Administration (FDA). The paper analyzes the options this approach provides as well as potential shortcomings it may pose. In particular, it studies the proposed concept of automated validation for these devices. In this analysis, the assessment of the relationship between technical parameters and clinical effects is a main focus. This includes the association to potential risks as well the dependencies between the algorithmic outcomes and the clinical environment. Additionally, potential issues w.r.t. bias, explainability, and fairness of the algorithms are addressed. The paper uses application scenarios, where ML based devices are utilized in the intensive care unit (ICU). In summary, ML based medical devices and especially continuously learning devices still possess considerable challenges which should be addressed thoroughly. Regarding appropriate regulatory strategies, a deliberate approach is recommended which prioritizes the collection of sufficient experience with ML based devices over amplifying their use in a rather uncontrolled fashion.

Keywords: Machine Learning, Continuously learning systems, SaMD – Software as a Medical Device, Regulatory requirements, Automated validation

1 Introduction

Medicine is a highly empirical discipline, where important aspects have to be demonstrated using adequate data and sound evaluations. This is one of the core requirements, which were emphasized during the development of the Medical Device Regulation (MDR) of the European Union (EU) [1]. This applies to all medical devices, including mechanical and electrical devices as well as software systems. Also, the US Food & Drug Administration (FDA) recently set a focus on the discussions about using data for demonstrating the safety and efficacy of medical devices [2]. Beside pure approval steps, they foster the use of data for optimization of the products, as nowadays data can be acquired more and more, using modern IT technology. In particular, they pursue the use of real world evidence, i.e. data that is collected through the lifetime of a device, for demonstrating improved outcomes. [2]

Such approaches require the use of sophisticated data analysis techniques. Beside classical statistics, artificial intelligence (AI) and machine learning (ML) are considered to be powerful techniques for this purpose. Currently, they gain more and more attention. These techniques allow to detect dependencies in complex situations, where inputs and/or outputs of a problem have high-dimensional parameter spaces. This can e.g. be the case when extensive data is collected from diverse clinical studies or also treatment protocols from local sites. Furthermore, AI/ML based techniques may be used in the devices themselves. For example, devices may be developed which are considered to improve complex diagnostic tasks or find individualized treatment options for specific medical conditions (see e.g. [3, 4] for an overview). For some applications, it already has been demonstrated that ML algorithms are able to outperform human experts with respect to specific success rates (e.g. [5, 6]). In this paper, it will be discussed how ML based techniques can be brought onto the market including an analysis of appropriate regulatory requirements. For this purpose, the main focus lies on ML based devices applied in the intensive care unit (ICU) as e.g. proposed in [7, 8].

The need for specific regulatory requirements comes from the observation, that AI/ML based techniques pose specific risks which need to be considered and handled appropriately. For example, AI/ML based methods are more challenging w.r.t. bias effects, reduced transparency, vulnerability to cybersecurity attacks, or general ethical issues (see e.g. [9, 10]). In particular cases, ML based techniques may lead to noticeably critical results, as it has been shown for the IBM Watson for Oncology device. In [11], it was reported that the direct use of the system in particular clinical environments resulted in critical treatment suggestions. The characteristics of ML based systems led to various discussions about their reliability in the clinical context. It requires to find appropriate ways to guarantee their safety and performance. (cf. [12]) This applies to the field of medicine / medical devices as well as AI/ML based techniques in general. The latter was e.g. approached by the EU in their *Ethics guidelines for trustworthy AI* [9].

Driven by this overall development, the FDA started a discussion regarding an extended use of ML algorithms in SaMD (software as a medical device) with a focus in quicker release cycles. In [13], it pursued the development of a specific process which makes it easier to bring ML based devices onto the market and also to update them during their lifecycle. Current regulations for medical devices, e.g. in US or EU, do not provide specific guidelines for ML based devices. In particular, this applies to systems which continuously collect data in order to improve the performance of the device. Current regulations focus on a fixed status of the device, which may only be adapted in a minor extent after the release. Usually, a new release or clearance by the authority is required, when the clinical performance of a device is modified. But continuously learning systems exactly want to do such improvement steps using additional real-world data from daily applications without extra approvals (see fig. 1).

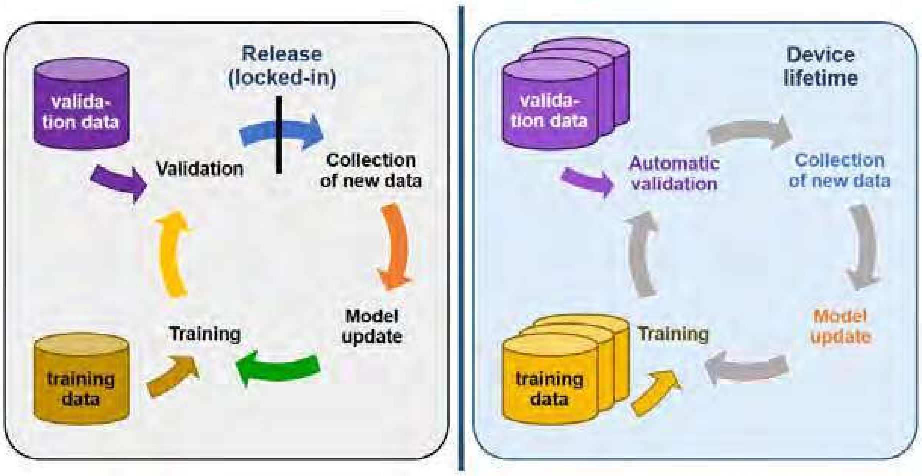

Fig. 1. Basic approaches for AI/ML based medical devices. Left side: classical approach, where the status of the software has to be fixed after the release / approval stage. Right side: continuously learning system where data is collected during the lifetime of the device without a separated release / approval step. In this case, an automatic validation step has to guarantee proper safety and efficacy.

In [13], the FDA made suggestions how this could be addressed. It proposed the definition of so called SaMD Pre-Specifications (SPS) and an Algorithm Change Protocol (ACP), which are considered to represent major tools for dealing with modifications of the ML based system during its lifetime. Within the SPS, the manufacturer has to define the anticipated changes which are considered to be allowed during the automatic update process. In addition, the ACP defines the particular steps which have to be implemented to realize the SPS specifications. See [13] for more information about SPS and ACP. But the details are not yet well elaborated by the FDA at the moment. The FDA requested for suggestions with respect to this.

In particular, these tools serve as a basis for performing an automated validation of the updates. The applicability of this approach depends on the risk of the SaMD. In [13], the FDA uses the risk categories from the International Medical Device Regulators Forum (IMDRF) [14]. This includes the categories *State of healthcare situation or condition* (*critical* vs. *serious* vs. *non-critical*) and *Significance of information provided by SaMD to healthcare decision* (*treat or diagnose* vs. *drive clinical management* vs. *inform clinical management*) as the basic attributes. According to [13], the regulatory requirements for the management of ML based systems are considered to depend on this classification as well as the particular changes which may take place during the lifetime of the device. The FDA categorizes them as changes in *performance*, *inputs*, and *intended use*. Such anticipated changes have to be defined in the SPS in advance.

The main purpose of the present paper is to discuss the validity of the described FDA approach for enabling continuously learning systems. Therefore, it uses a scenario based technique to analyze whether validation in terms of SPS and ACP can be considered adequate tools. The scenarios represent applications of ML based devices in the ICU. It checks its consistency with other important regulatory requirements and analyzes pitfalls which may jeopardize the safety of the devices. Additionally, it discusses whether more general requirements can be sufficiently addressed in the scenarios, as e.g. proposed in ethical guidelines for AI based systems like [9, 10]. This is not considered as a comprehensive analysis of the topics, but as an addition to current discussions about risks and ethical issues, as they are e.g. discussed in [10, 12].

Finally, the paper proposes own suggestions to address the regulation of continuously learning ML based systems. Again, this is not considered to be a full regulatory strategy, but a proposal of particular requirements, which may overcome some of the current limitations of the approach discussed in [13]. The overall aim of this paper is to contribute to a better understanding of the options and challenges of AI/ML based devices on the one hand and to enable the development of best practices and appropriate regulatory strategies, in the future.

2 Methods

Within this paper, the analysis of the FDA approach proposed in [13] is performed using specific reference scenarios from ICU applications, which are particularly taken from [13] itself. The focus lies on ML based devices which allow continuous updates of the model according to data collected during the lifetime of the device. In this context, SPS and ACP are considered as crucial steps which allow an automated validation of the device based on specified measures. In particular, the requirements and limitations of such an automated validation are analyzed and discussed, including the following topics / questions.

- Is automated validation reasonable for these cases? What are limitations / potential pitfalls of such an approach when applied in the particular clinical context?
- Which additional risks could apply to AI/ML based SaMD, in general, which go beyond the existing discussions in the literature as e.g. presented in [9, 10, 12]?
- How should such issues be taken into account in the future?
  What could be appropriate measures / best practices to achieve reliability?

The following exemplary scenarios are used for this purpose.

- Base Scenario ICU: ML based Intensive Care Unit (ICU) monitoring system where the detection of critical situations (e.g. regarding physiological instability, potential myocardial infarcts or sepsis) is addressed by using ML. Using auditory alarms, the ICU staff is informed to initiate appropriate measures to treat the patients in these situations. This scenario addresses a 'critical healthcare situation or condition' and is considered to 'drive clinical management' (according to the risk classification used in [13]).
- Modification "locked": ICU scenario as presented above, where the release of the monitoring system is done according to a locked state of the algorithm.
- Modification "cont-learn": ICU scenario as presented above, where the detection of alarm situations is continuously improved according to data acquired during daily routine, including adaptation of performance to sub-populations and/or characteristics of the local environment. In this case, SCS and ACP have to define standard measures like success rates of alarms/detection and requirements for the management of data, update of the algorithm, and labeling. More details of such requirements are discussed later. This scenario was presented as scenario 1A in [13] with minor modifications.

3 Analysis

This section provides the basic analysis of the scenarios according to the particular aspects addressed in this paper. It addresses the topics automated validation, man-machine interaction, explainability, bias effects, and confounding, fairness and non-discrimination as well as corrective actions to systematic deficiencies.

Basic Considerations about Automated Validation

According to standard regulatory requirements [1, 15, 16], validation is a core step in the development and for the release of medical devices. According to [17], a change in performance of a device (including an algorithm in a SaMD) as well as a change in particular risks (e.g. new risks, but also new risk assessment or new measures) usually triggers a new premarket notification (510(k)) for most of the devices which get onto the market in the US. Thus, such situations require an FDA review for clearance of the device. For SaMD, this requires to include an analytical evaluation, i.e. correct processing of input data to generate accurate, reliable, and precise output data. Additionally, a clinical validation as well as the demonstration of a valid clinical association need to be provided. [18] This is intended to show that the outputs of the device appropriately work in the clinical environment, i.e. have a valid association regarding the targeted clinical condition and achieve the intended purpose in the context of clinical care. [18]
Thus, based on the current standards, a device with continuously changing performance usually requires a thorough analysis regarding its validity. This is one of the main points, where [13] proposes to establish a new approach for the "cont-learn" cases. As already mentioned, SPS and ACP basically have to be considered as tools for automated validation in this context. Within this new approach, the manual validation step is replaced by an automated process with only reduced or even no additional control by a human observer. Thus, it may work as an automated of fully automatic, closed loop validation approach. The question is whether this change can be considered as an appropriate alternative. In the following, this question is addressed using the ICU scenario with a main focus on the "cont-learn" case. Some of the aspects also apply to the "locked" cases. But the impact is considered to be higher in the "cont-learn" situation, since the validation step has to be performed in an automated fashion. Human oversight, which is usually considered important, is not included here during the particular updates.

Analysis of Validation Steps

Within the ICU scenario, the validation step has to ensure that the alarm rates stay on a sufficiently high level, regarding standard factors like specificity, sensitivity, area under curve (AUC), etc. Basically, these are technical parameters which can be analyzed according to an analytical evaluation as discussed above. (see also [18]) This could also be applied to situations, where continuous updates are made during the lifecycle of the device, i.e. in the “cont-learn”. However, there are some limitations of the approach. On the one hand, it has to be ensured, that this analysis is sound and reliable, i.e. it is not compromised according to statistical effects like bias or other deficiencies in the data. On the other hand, it has to be ensured that the success rates really have a valid clinical association and can be used as a sole criterion for measuring the clinical impact. Thus, the relationship between pure success rates and clinical effects has to be evaluated thoroughly and there may be some major limitations.

One major question in the ICU scenario is, whether better success rates really guarantee a higher or at least sufficient level of clinical benefit. This is not innately given. For example, a higher success rate of the alarms may still have a negative effect when the ICU staff relies more and more on the alarms and subsequently reduces attention. Thus, it may be the case that the initiation of appropriate treatment steps may be compromised even though the actually occurring alarms seem to be more reliable.

In particular, this may apply in situations where the algorithms are adapted to local settings, like in the “cont-learn” scenario. Here, the ML based system is intended to be optimized to sub-populations in the local environment or to specific treatment preferences at the local site. According to habituation effects, the staff’s expectations get aligned to the algorithm’s behavior to a certain degree after a period of time. But when the algorithm changes or an employee from another hospital or department takes over duties in the local unit, the reliability of the alarms may be affected. In these cases, it is not clear whether the expectations are well aligned with the current status of the algorithm – either in the positive or negative direction. Since the data updates of the device are intended to improve its performance w.r.t. detection rates, it is clear that significant effects on user interaction may happen. Under some circumstances, the overall outcome in terms of the clinical effect may be impaired.

Evaluation of such risks have to be addressed during validation. It is questionable whether this can be performed by using an automatic validation approach which focuses on alarm rates but does not include an assessment of the associated risks. At least a clear relationship between these two aspects has to be demonstrated in advance. It is also unclear, whether this could be achieved by assessment of pure technical parameters which are defined in advance as required by the SPS and ACP. Usually, ML based systems are trained to a specific scenario. They provide a specific solution for this particular problem. But they do not have a more general intelligence and reasoning about potential risks, which were not under consideration at that point of time. Such a more general intelligence can only be provided when using human oversight.

Consideration of Risks and Man-Machine Interaction

In general, it is not clear whether technical aspects like alarms lead to valid reactions by the users. In technical terms, alarm rates are basically related to the probability of occurrence of specific hazardous situations. But they do not address a full assessment of occurrence of harm. However, this is pivotal for risk assessment in medical devices, in particular for risks related to potential use errors. This is considered to be one of the main reasons why a change in risk parameters triggers a new premarket approval in the US according to [17]. Also, the MDR [1] sets high requirements to address the final clinical impact and not only technical parameters.

Basically, the example emphasizes the importance to consider the interaction between man and machine, or in this case, the algorithm and its clinical environment. This is addressed in the usability standards for medical devices, e.g. ISO 62366 [19]. For this reason, the ISO 62366 requires that the final (summative) usability evaluation is performed using the final version of the device (in this case, the algorithm) or an equivalent version. This is in conflict with the FDA proposal which allows to perform this assessment based on previous versions. At most, a predetermined relationship between technical parameters (alarm rates) and clinical effects (in particular, use related risks) can be obtained. For usage of ML based devices, it remains crucial to consider the interaction between the device and the clinical environment as there usually are important interrelationships.

Comprehensiveness of Included Inputs and Outcomes

The outcome of an ML based algorithm always depends on the data it gets provided. Whenever an input parameter is omitted, which is clinically relevant, the resulting outcome of the ML based system is limited. In the presented scenarios, the pure alarm rates may not be the only clinically relevant outcomes. Even though, such parameters are usually the main focus regarding the quality of algorithms, e.g. in publications about ML based techniques. This is due to the fact, that such quality measures are commonly considered the best available objective parameters, which allow a comparison of different techniques.
This even more applies to other ML based techniques which are also very popular in the scientific community, like segmentation tasks in medical image analysis. Here the standard quality measures are general distance metrics, i.e. differences between segmented areas. [20] They usually do not include specific clinical aspects like the accuracy in specific risk areas, e.g. important blood vessels or nerves. But such aspects are key factors to ensure the safety of a clinical procedure in many applications. Again, only technical parameters are typically in focus. The association to the clinical effects is not assessed accordingly. This situation is depicted in fig. 2 for the ICU as well as image segmentation cases.

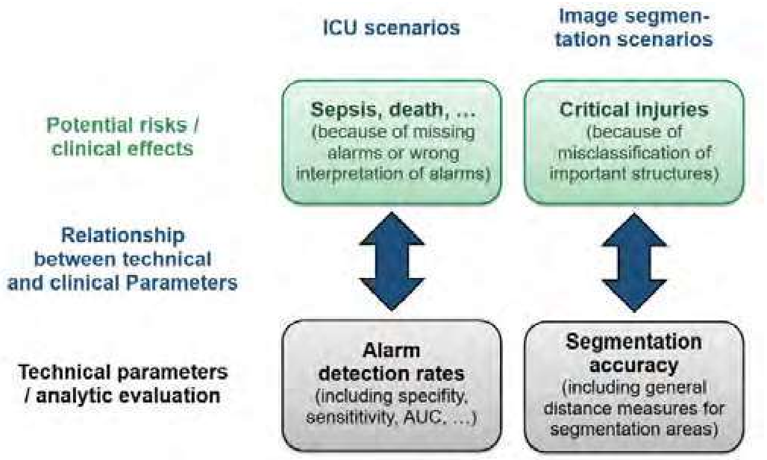
ICU scenarios
Image segmentation scenarios
Potential risks / clinical effects
Sepsis, death, ... (because of missing alarms or wrong interpretation of alarms)
Critical injuries (because of misclassification of important structures)
Relationship between technical and clinical Parameters
Technical parameters / analytic evaluation
Alarm detection rates (including specifity, sensititivity, AUC, ...)
Segmentation accuracy (including general distance measures for segmentation areas)

Fig. 2. Relationship between technical parameters / analytical evaluation (as usually considered in many publications about ML based techniques) on the one hand and potential risks / clinical effects (finally relevant for release of medical devices) on the other hand (using ICU and image segmentation scenarios).

Additionally, the validity of an outcome in medical treatments depends on many factors. Regarding input data, multiple parameters from a patient's individual history may be important for deciding about a particular diagnosis or treatment. A surgeon usually has access to a multitude of data and also side conditions (like socio-economic aspects) which should be included in an individual diagnosis or treatment decision. His general intelligence and background knowledge allow him to include a variety of individual aspects, which have to be considered for a specific case-based decision. In contrary, ML based algorithms rely on a more standardized structure of input data and are only trained for a specific purpose. They lack a more general intelligence, which allows them to react in very specific situations. Even more, ML based algorithms need to generalize and thus to mask out very specific conditions, which could by fatal in some cases.

In [13], the FDA presents some examples where changes of the inputs in an ML based SaMD are included. It is surprising, that the FDA considers some of them as candidates for a continuous learning system, which does not need an additional review, when a tailored SPS/ACP is available.

Lack of Explainability and its Impact on Clinical Practice

Such discrepancies between technical outcomes and clinical effects also apply to situations like the ICU scenario, which only informs or drives clinical management. Often users rely on automatically provided decisions, even when they are informed that this only is a proposal. Again, this is a matter of man-machine interaction. This gets even worse due to the lack of explainability which ML based algorithms typically have. [9, 21] When surgeons or more general users (e.g, ICU staff) detect situations which require a diverging treatment because of very specific individual conditions, they should overrule the algorithm. But users will often be confused by the outcome of the algorithm and do not have a clear idea how they should treat conflicting results between the algorithm's suggestions and their own belief. As long as the ML based decision is not transparent to the user, they will not be able to merge these two directions. The IBM Watson example, referenced in the introduction shows, that this actually is an issue [11].

This may be even more serious, when the users (i.e. healthcare professionals) fear litigation because they did not trust the algorithm. In a situation, where the algorithm's outcome finally turns out to be true, they may be sued because of this documented deviation. Because of such issues, the EU General Data Protection Regulation (GFPR) [22] requires that the users get autonomy regarding their decisions and transparency about the mechanisms underlying the algorithm's outcome. [23] This may be less relevant for the patients, who usually have only limited medical knowledge. They will probably also not understand the medical decisions in conventional cases. But it is highly relevant for responsible healthcare professionals. They require to get basic insights how the decision emerged, as they finally are in charge of the treatment. This demonstrates that methods regarding the explainability of ML based techniques are important. Fortunately, this currently gets a very active field. [21, 24] This need for explainability applies to locked algorithms as well as situations where continuous learning is applied.

Bias Effects due to Imbalance of Data

Based on their own data-driven nature, ML based techniques highly depend on a very high quality of data which are provided for learning and validation. In particular, this is important for the analytical evaluation of the ML algorithms. One of the major aspects are bias effects due to unbalanced input data. For example, in [25] a substantially different detection rate between white and colored people was recognized due to unbalanced data. Beside ethical considerations, this demonstrates dependencies of the outcome quality on sub-populations, which may be critical in some cases. Even though, the FDA proposal [13] currently does not consequently include specific requirements for assessing bias factors or imbalance of data.

However, high quality requirements for data management are crucial for ML based devices. In particular, this applies to the ICU "cont-learn" cases. There have to be very specific protocols that guarantee that new data and updates of the algorithms are highly reliable w.r.t. bias effects. Most of the currently used ML based algorithms fall under the category of supervised learning. Thus, they require accurate and clinically sound labeling of the data. During the data collection, it has to be ensured how this labeling is performed and how the data can be fed back into the system in a “cont-learn” scenario. Additionally, the data needs to stay balanced – whatever this means in a situation where adaptions to sub-populations and/or local environments are intended for optimization. It is unclear, whether and how this could be achieved by staff who is only operating with the system but possibly does not know potential algorithmic pitfalls.

In the ICU scenario, many data points probably need to be recorded by the system itself. Thus, a precise and reliable recording scheme has to be established which automatically avoids imbalance of data on the one hand and fusion with manual labelings on the other hand. Basically, the SPS and ACP (proposed in [13]) are tools to achieve this. The question is whether this is possible in a reliable fashion using automated processes. A complete closed loop validation approach seems to be questionable, especially when the assessment of clinical impact has to be included. Thus, the integration of humans including adequate healthcare professionals as well as ML/AI experts with sufficient statistical knowledge seems reasonable. At least, bias assessment steps should be included. As already mentioned, this is not addressed in [13] in a dedicated way.

Confounding Factors and Dependence on Clinical Environment

Further on, the outcomes may be compromised by side effects in the data. It may be the case, that the main reason for a specific outcome of the algorithm is not a relevant clinical parameter but a specific data artifact, i.e. some confounding factor. In the ICU case, it could be the case, that the ICU staff reacts early to a potentially critical situation and e.g. gives specific medication in advance to prevent upcoming problems. The physiological reaction of the patient can then be visible in the data as some kind of artifact. During its learning phase, the algorithm may recognize the critical situation not based on a deeper clinical reason, but on detecting the physiological reaction pattern. This may cause serious problems as shown subsequently. In the presented scenario, the definition of clinical situation and the pattern can be deeply coupled by design, since the labeling of the data by the ICU staff and the administration of the medication will probably be done in combination at the particular site. This may increase the probability of such effects.

Usually, confounding factors are hard to determine. Even when they can be detected, they are hard to be communicated and managed in an appropriate way. How should healthcare professionals react, when they get such potentially misleading information (see discussion about liability). This further limits the explanatory power of ML based systems. When confounders are not detected, they may have unpredictable outcomes w.r.t. the clinical effects. For example, consider the following case. In the ICU scenario, an ML based algorithm gets trained in a way that it basically detects the medication artifact as described above during the learning phase. In the next step, this algorithm is used in clinical practice and the ICU staff relies on the outcome of the algorithm. Then, on the one hand, the medication artifact is not visible unless the ICU staff administers the medication. On the other hand, the algorithm does not recognize the pattern and thus does not provide an alarm. Subsequently, the ICU staff does no act appropriately to manage the critical situation.

In particular, such confounders may be more likely in situations where a strong dependence between the outcome of the algorithm and the clinical treatment exists. Further examples of such effects were discussed in [7] for ICU scenarios. The occurrence of confounders may be a bit less probable in pure diagnostic cases without influence of the diagnostic task onto the generation of data. But even here, such confounding factors may occur. The discussion in [10] provided examples where confounders may occur in diagnostic cases e.g. because of rulers placed for measurements on radiographs. In most of the publications about ML based techniques, such side effects are not discussed (or only in a limited fashion). In many papers, the main focus is the technical evaluation and not the clinical environment and the interrelation between technical parameters and clinical effects.

Fairness and Non-Discrimination

Additional important aspects which are amply discussed in the context of AI/ML based systems are discrimination and fairness (see e.g. [10]). In particular, the EU puts a high priority of their future AI/ML strategy on fairness requirements [9]. Fairness is often closely related to bias effects. But it goes beyond to more general ethical questions, e.g. regarding the natural tendency of ML based systems to favor specific subgroups. For example, the ICU scenario “cont-learn” is intended to optimize w.r.t. to specifics of sub-populations and local characteristics, i.e. it tries to make the outcome better for specific groups. Based on such optimization, other groups (e.g. minorities, underrepresented groups) which are not well represented may be discriminated in some sense. This is not a statistical but a systematic effect.

Superiority of a medical device for a specific subgroup (e.g. gender, social environment, etc.) is not uncommon. For example, some diagnosis steps, implants, or treatments achieve deviating success rates when applied to women in comparison to men. This also applies to differences between adults and children. When assessing bias in clinical outcome in ML based devices, it will probably often be unclear whether this is due to imbalance of data or a true clinical difference between the groups. Does an ML based algorithm has to adjust the treatment of a subgroup to a higher level, e.g. a better medication, to achieve comparable results, when the analysis recognized worse results for this subgroup? Another example could be a situation where the particular group does not have the financial capabilities to afford the high-level treatment. This could e.g. be the case in a developing country or in subgroups with a lower insurance level. In these cases, the inclusion of socio-economical parameters into the analysis seems to be unavoidable. Subsequently, this compromises the notion of fairness as basic principle in some way.

This is nothing genuine to ML based devices. But in the case of ML based systems with a high degree of automation, the responsibility for the individual treatment decision more and more shifts from the health care professional to the device. It is implicitly defined in the ML algorithm. In comparison to human reasoning, which allows some weaknesses in terms of individual adjustments of general rules, ML based algorithms are rather deterministic / unique in their outcome. For a fixed input, they have one dedicated outcome (when we neglect statistical algorithms which may allow minor deviations). Differences of opinions and room for individual decisions are main aspects of ethics. Thus, it remains unclear how fairness can be defined and implemented at all when considering ML based systems. This is even more challenging as socio-economical aspects (even more than clinical aspects) are usually not included in the data and analysis of ML based techniques in medicine. Additionally, they are hard to assess and implement in a fair way, especially when using automated validation processes.

Corrective Actions regarding Systematic Deficiencies

Another disadvantage of ML based devices is the limited opportunities to fix systematic deficiencies in the outcome of the algorithm. Let us assume that during the lifetime of the ICU monitoring system a systematic deviation of the intended outcome was detected, e.g. in the context of post-market surveillance or due to an increased number of serious adverse events. According to standard rules, a proper preventive respectively corrective action has to be taken by the manufacturer. In conventional software devices, the error simple should be eliminated, i.e. some sort of bug fixing has to be performed. For ML based devices it is less clear, how bug fixing should work especially when the systematic deficiency is deeply hidden in the data and/or ML model. In these cases, there usually is no clear reason for the deficiency. Subsequently, the deficiency cannot be resolved in a straightforward way using standard bug fixing. There is no dedicated route to find the deeper reasons and to perform changes which could cure the deficiencies, e.g. by providing additional data or changing the ML model. Even more, other side effects may easily occur, when data and model are changed manually by intent to fix the issue.

4 Discussion and Outlook

In summary, there are many open questions, which are not yet clarified. There still is little experience how ML based systems work in clinical practice and which concrete risks may occur. Thus, the FDA's commitment to foster the discussion about ML based SaMD is necessary and appreciated by many stakeholders as the feedback docket [26] for [13] shows. However, it is a bit surprising that the FDA proposes to substantially reduce its very high standards in [13] at this point of time. In particular, it is questionable whether an adequate validation can be achieved by using a fully automatic approach as proposed in [13]. ML based devices are usually optimized according to very specific goals. They can only account for the specific conditions that are reflected in the data and the used optimization / quality criteria. They do not include side conditions and a more general reasoning about potential risks in a complex environment. But this is important for medical devices.

For this reason, a more deliberate path would be suited, from the author's perspective. In a first step, more experience should be gained w.r.t. to the use of ML based devices in clinical practice. Thus, continuous learning should not be a first hand option. First, it should be demonstrated that a device works in clinical practice before a continuous learning approach should be possible. This could also be justified from a regulatory point-of-view. The automated validation process itself should be considered as a feature of the device. It should be considered as part of the design transfer which enables safe use of the device during its lifecycle. As part of the design transfer, it should be validated itself. Thus, it has to be demonstrated that this automated validation process, e.g. in terms of the SPS and ACP, works in a real clinical environment. Ideally, this would have been demonstrated during the application of the device in clinical practice.

Thus, one reasonable approach for a regulatory strategy could be to reduce or prohibit the options for enabling automatic validation in a first release / clearance of the device. During the lifetime, direct clinical data could be acquired to demonstrate a better insight into the reliability and limitations of the automatic validation / continuous learning approach. In particular, the relation between technical parameters and clinical effects could be assessed on a broader and more stable basis. Based on this evidence in real clinical environments, the automated validation feature could then be cleared in a second round.

Otherwise, the validity of the automated validation approach would have to be demonstrated in a comprehensive setting during the development phase. In principle, this is possible when enough data is available which truly reflects a comprehensive set of situations. As discussed in this paper, there are many aspects which do not render this approach impossible but very challenging. In particular, this applies to the clinical effects and the interdependency between the users and clinical environment on the one hand and the device, including the ML algorithm, data management, etc., on the other hand. This also includes not only variation in the status and needs of the individual patient but also the local clinical environment and potentially also the socioeconomic setting.

Following a consequent process validation approach, it would have to be demonstrated that the algorithm reacts in a valid and predictable way no matter which training data have been provided, which environment have to be addressed, and which local adjustments have been applied. This also needs to include deficient data and inputs in some way. In [20], it has been shown that the variation of outcomes can be substantial, even w.r.t. rather simple technical parameters. In [20], this was analyzed for scientific contests ("challenges") where renowned scientific groups supervised the quality of the submitted ML algorithms. This demonstrates the challenges validation steps for ML based systems still include, even w.r.t. technical evaluation.

For these reasons, it seems adequate to pursue the regulatory strategy in a more deliberate way. This includes the restriction of the "cont-learn" cases as proposed. This also includes a better classification scheme, where automated or fully automatic validation is possible. Currently, the proposal in [13] does not provide clear rules when continuous learning is allowed. It does not really address a dedicated risk-based approach that defines which options and limitations are applicable. For some options, like the change of the inputs, it should be reviewed, whether automatic validation is a natural option. Additionally, the dependency between technical parameters and clinical effects as well as risks should get more attention. In particular, the grade of interrelationship between the clinical actions and the learning task should be considered.

In general, the discussions about ML based medical devices are very important. These techniques provide valuable opportunities for improvements in fields like medical technologies, where evidence based on high quality data is crucial. This applies to the overall development of medicine as well as to the development of sophisticated ML based medical devices. This also includes the assessment of treatment options and success of particular devices during their lifetime. Data-driven strategies will be important for ensuring high-level standards in the future. They may also strengthen regulatory oversight in the long term by amplifying the necessity of post-market activities. This seems to be one of the promises the FDA envisions according to their concepts of "total product lifecycle quality (TPLC)" and "organizational excellence" [13]. Also, the MDR strengthens the requirements for data-driven strategies in the pre- as well as post-market phase. But it should not shift the priorities for a basically proven-quality-in-advance (ex-ante) to a primarily ex-post regulation, which boils down to a trial-and-error oriented approach in the extreme. Thus, we should aim at a good compromise between pushing these valuable and innovative options on the one hand and potential challenges and deficiencies on the other hand.

References

1. European Parliament and Council. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices – Medical Device Regulation (MDR). 2017
2. Food & Drug Administration (FDA). Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices. Guidance for Industry and Food and Drug Administration Staff. FDA (Aug 31, 2017)
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. (2019) Jan;25(1):44-56
4. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ. Digit Med. (2018) 14;1:5.
5. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. (2017) 2;542(7639):115-118
6. Wang, X. et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017) 3462–3471

7. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, Oermann EK. An attention based deep learning model of clinical events in the intensive care unit. PLoS One. (2019) 13;14(2):e0211057.
8. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. (2018) 24(11):1716-1720
9. The European Commission's High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI. European Commission. (Apr 8, 2019).
10. Kelly, C.J., Karthikesalingam, A., Suleyman, M. et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. (2019) 17, 195
11. Ross C, Swetlitz I. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. Stat News https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafeincorrect-treatments/ (2018)
12. Wenzel M, Wiegand T. Towards International Standards for the Evaluation of Artificial Intelligence for Health. ITU Kaleidoscope 2019: ICT for Health: Networks, standards and innovation, Atlanta, USA. (2019)
13. Food & Drug Administration (FDA). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). FDA (2019). https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf
14. International Medical Device Regulators Forum (IMDRF) – SaMD Working Group. "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. FDA (2014)
15. International Electrotechnical Commission. IEC 62304:2015. Medical device software - Software life-cycle processes. International Electrotechnical Commission. (2015)
16. Food & Drug Administration (FDA). General Principles of Software Validation. Final Guidance for Industry and FDA Staff. FDA (Jan 11, 2002)
17. Food & Drug Administration (FDA). Deciding When to Submit a 510(k) for a Change to an Existing Device. Guidance for Industry and Food and Drug Administration Staff. FDA (Oct 25, 2017)
18. Food & Drug Administration (FDA). Software as a Medical Device (SAMD): Clinical Evaluation. Guidance for Industry and Food and Drug Administration Staff. FDA (Dec 8, 2017)
19. International Electrotechnical Commission. IEC 62366-1:2015 - Part 1: Application of usability engineering to medical devices. International Electrotechnical Commission (2015)
20. Maier-Hein L et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun. (2018) 6;9(1):5217.
21. Holzinger A, Biemann C, Pattichis CS. What do we need to build explainable AI systems for the medical domain? (2017) arXiv:1712:09923; https://arxiv.org/abs/1712.09923.
22. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation - GDPR). 2016
23. Schönberger D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications, I. J. Law and Information Technology. (2019) 27(2):171–203
24. Samek W, Wiegand T, Müller K-R. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. (2017) arXiv:1708.08296. http://arxiv.org/abs/1708.08296.
25. Ward-Peterson M, et al. Association between race/ethnicity and survival of melanoma patients in the United States over 3 decades. Medicine. (2016) 95:e3315.
26. Food & Drug Administration (FDA). Docket for Feedback – Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). FDA. https://www.regulations.gov/docket?D=FDA-2019-N-1185