




# CRISP-PCCP – A Development Methodology Supporting FDA Approval for Machine Learning Enabled Medical Devices

Ludwig Pechmann<sup>1</sup>, Yannik Potdevin<sup>2</sup>, Kai Brehmer<sup>3</sup>, Dirk Nowotka<sup>2</sup>, and  
Martin Leucker<sup>1,4</sup>

<sup>1</sup> UniTransferKlinik Lübeck GmbH, Lübeck, Germany

<sup>2</sup> Dependable Systems Group, Kiel University, Kiel, Germany

<sup>3</sup> Institute for Electrical Engineering in Medicine, University of Lübeck, Lübeck,  
Germany

<sup>4</sup> Institute for Software Engineering and Programming Languages, University of  
Lübeck, Lübeck, Germany

**Abstract.** The U.S. Food and Drug Administration (FDA) is the regulatory body that ensures the safety, efficacy, and security of medical devices and software in the healthcare sector in the U.S. However, its guidelines and regulations often set a global benchmark, influencing medical device standards in Europe and other regions. The FDA recently published a draft guidance, the Predetermined Change Control Plan (PCCP), aiming to support medical device manufacturers with the release of continual learning Machine Learning-Enabled Device Software Functions (ML-DSF). Such ML-DSFs are intended to change after initial market approval. We present a systematic process to support the implementation of the PCCP. Building upon the Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance methodology (CRISP-ML(Q)), we present an approach that a manufacturer may use to identify and evaluate the impact of anticipated changes to ML-DSF. Our process also indicates a forecast, whether the anticipated change would be accepted by the FDA as a part of the PCCP.

**Keywords:** FDA, Predetermined Change Control Plan (PCCP), CRISP-ML(Q), Machine Learning-Enabled Device Software Function (ML-DSF)

## 1 Introduction

Medical devices are safety critical systems that are typically subject to strict regulatory restrictions. In the U.S., these are imposed by the Food and Drug Administration (FDA), which approves the medical product prior to market release. As part of this, software functions of such systems need sophisticated testing. Depending on the change of software after market approval, a re-approval becomes necessary. In the EU, the Medical Device Regulation (MDR) is

mandatory and defines a rigorous certification process. However, for this paper, we concentrate on the development of the FDA.

Today, more and more such software functions are developed using machine learning techniques. Often, such learning Machine Learning-Enabled Device Software Functions (ML-DSF) are developed iteratively, meaning software changes are continually applied, either by manually triggered re-learning of functions or improving the system automatically whenever new data is available.

ML applications that do not change after market approval can be approved under current FDA regulations. However, any manual update would require re-approval. Systems that update automatically could not even get approval. This is because continually learning systems may adapt to changes in the environment or input on their own. Such a system would change dynamically without manufacturer supervision after initial FDA approval. As a result, the device in its updated state no longer conforms to what was originally tested and approved. This is contrary to current regulations, which require that any medical device is used in its approved form, with no subsequent changes that could affect its functionality. However, this hinders the usage of the benefits of continual learning systems which may improve their quality in a continual fashion.

To address this problem, the FDA initiated discussions with corresponding stakeholders and proposed a draft guidance to support the use of continual learning machine learning software as medical devices, hereby assisting medical device manufacturers with the development and the approval process of such systems. The most recent guidance is the Predetermined Change Control Plan (PCCP)[27]. Briefly explained, the FDA expects the medical device manufacturers to state at application for initial approval, what changes to the Machine Learning-Enabled Device Software Function (ML-DSF) are expected to occur during the lifetime of the medical device and how this change would affect the overall device. The anticipated changes are compiled into the PCCP. It is emphasized that the FDA expects to define the PCCP at initial approval, potentially long before the expected changes may occur. The intention is that if an acceptable anticipated change occurs and the manufacturer acts according to what he or she stated in the PCCP, a modification to the ML part of the ML-DSF may be distributed *without* requiring re-approval by the FDA [27]. Otherwise, that is if a change has not been anticipated, or if the steps to address the changes turn out to be not suitable or sufficient, a completely new approval by the FDA may be needed. Note that in this case, distributing the modified medical device would constitute adulteration and misbranding.

To support the development of PCCP, this paper proposes CRISP-PCCP as a new methodology to systematically identify effects and implications of changes during the release process of an ML-DSF. It is inspired by the Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance methodology (CRISP-ML(Q)) [19]. In simple words, CRISP-PCCP consists of several steps that are iterated. The first step of CRISP-PCCP is to provide a sufficiently precise description of the anticipated device change.

Next, one must walk through the phases and sub-phases (generic tasks) of an ML processing pipeline and determine for each generic task, whether it is affected by the anticipated device change. If so, additional generic questions for further details must be answered. In this way, a deeper understanding of the potential changes is obtained, and a subsequent risk re-evaluation must be performed to determine their potential impact on the overall system safety, effectiveness, and compliance with regulatory standards. This last step is called the change impact analysis. Following the proposed methodology not only sharpens the conception of the anticipated device change and its consequences, but also provides an estimation of whether an anticipated device change will be accepted by the FDA as part of the PCCP.

The rest of the paper is organized as follows: In Section 2, the increasing importance of continuously learning AI/ML products is highlighted and the need for a forward-looking change control plan for AI/ML models in the medical device industry is underlined. Then we continue in Section 3 with a brief overview of the regulatory context. In Section 4 we describe the process, which uses the CRISP-ML(Q) as its foundation: The risk-based approach for manufacturer to identify changes along the ML life cycle and to predetermine their impact during field usage. Further we estimate FDA acceptance within each phase of the process. Finally, we discuss our experience and findings with the development and usage of the approach on real world projects in Section 5 and conclude with steps to improve the approach in Section 6.

## 2 Machine Learning in Medical Devices

Machine Learning (ML) is an area of computer science dedicated to developing systems that can execute tasks usually associated with cognitive processes. ML systems can analyze data, recognize patterns, make decisions, and adapt to evolving situations without explicit programming. Within the medical domain, ML has transformed medicine by enhancing diagnostics, personalizing treatments, streamlining drug discovery, and enabling predictive analytic [11]. However, it also presents challenges related to data privacy, regulation, and ethical use [14,2]. The ethical issues referred to include concerns about bias and fairness in ML algorithms. Bias can occur when an ML system generates skewed or prejudiced results due to flawed assumptions in the algorithm or biased data inputs. This in turn can lead to unfair treatment of individuals or groups [14].

In comparison to classical V-Model driven software development, where requirements engineering and testing are typically performed at well-defined stages in sequence [9], the development of ML models is performed slightly different. The typical ML development process involves developing a model that is taken through a series of iterative steps that require constant adaptation and learning as the model interacts with data. At the center of this process are the ML algorithms that serve as a blueprint and dictate how the system should achieve its goal. The model acts as an instance of the algorithm and dynamically adapts through training iterations to approximate a target function

that is initially unknown<sup>5</sup>. This iterative nature of development and focus on adapting models through learning makes the traditional software development life cycle less suitable to the nuanced and evolving landscape of ML development. [5,13]

ML approaches may be distinguished whether they are static or continual: A static model is trained offline. The model will be trained till it reaches a defined goal in predicting certain features. After the training phase, the model will be used without further changes. However, static models often perform well on similar data but could perform poorly in scenarios that are rare in the training process. Also, they prevent the ability to learn from post-approval, real-world data, and thus cannot improve over time in the same way as adaptive systems.

Dynamic models, also known as continual learning ML models, are trained online. The model will also be trained till it reaches a defined goal in predicting certain features. But after the training phase, the data that is continually processed by the system, is also used to update the model [28]. Continual learning ML algorithms are designed to update and improve themselves as their input data, environments, and/or targets change. This ability to adapt to changing data has the potential to create more advanced Machine Learning-Enabled Device Software Function (ML-DSF) that would allow to improve the performance [28,15].

However, it poses risks that need to be addressed, such as the introduction of new errors, system performance deterioration, if the newly integrated data are biased, and the risk that new information could interfere with what the model has already learned. Therefore, it is important to carefully manage and monitor these systems [28,15].

In addition, ML-DSF with continuously learning abilities would result in an unknown and undocumented version of the medical device software. This would lead to an illegal product under the current regulations. The FDA has recognized this issue and is actively addressing it within their “Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (ML)-Based Software as a Medical Device [23]”. This framework aims to create a clear pathway for ML-DSF that are subject to continuously learning and adaptation, allowing them to improve over time while ensuring patient safety and device effectiveness.

### 3 Regulatory Context - Approval of Medical Devices

In the United States, the Food and Drug Administration (FDA) serves as the regulatory authority responsible for overseeing Food, Pharmaceuticals, and Medical Devices (MD). Its primary mission is to guarantee that these products adhere to the highest standards of safety, effectiveness, and quality before they enter the market. The FDA achieves this through rigorous and comprehensive regulations designed to safeguard consumers’ health and well-being.

---

<sup>5</sup> This is the essence of machine learning (ML)

The FDA defines a medical device as any product, including accessories, intended for diagnosing, curing, treating, or preventing diseases in people or animals [25]. It covers a vast range of items, from simple tools like bedpans to complex technologies like pacemakers. These devices work primarily through physical or mechanical means, rather than chemical action, and are classified based on their *intended use* and the *level of risk* they pose.

The FDA employs three key processes for approving Medical Devices (MDs) to meet quality and safety standards. The *510(k) Process* is the most common one and used for devices that are substantially equivalent to an approved one. For manufacturers, it suffices to give details showing that their medical device is in substantial equivalence with a previously cleared (approved) one. The *De Novo Classification Process* is for innovative MDs without an existing comparable device. Here, manufacturers request a new classification by providing unique device evidence for FDA review. The *Pre-market Approval (PMA) Process*, which is for high-risk MDs, requires extensive data and a clinical study demonstrating clinical benefits. The FDA reviews this to ensure safety and efficacy. These processes are risk-based, with regulatory scrutiny varying according to each MD's risk level.

### 3.1 Change-management for Software in Medical Devices

The FDA regulatory frameworks lay out guidelines for managing software changes in MDs to ensure the continued safety and effectiveness of these devices. The FDA emphasizes the importance of effective change control procedures e.g. for software changes [26]. The types of changes differ based on their impact on the device's safety and effectiveness.

*Minor changes* typically have a low impact on the device's safety and effectiveness. Manufacturers can implement these changes without prior FDA approval but must notify the FDA within 30 days of making the change. This type of change encompasses modifications to labeling or bugfixes in the medical device software.

*Moderate changes* have a more significant impact on the device's safety or effectiveness. Manufacturers are generally required to submit a new 510(k) submission or a PMA supplement for these changes, seeking FDA clearance before implementing them. The FDA will review the submission to ensure that the modifications do not compromise the device's safety and efficacy. This type of change encompasses modifications to design or functionality which could affect the medical device software's performance.

*Major changes* have the potential to significantly affect the safety and effectiveness of the device. Manufacturers typically need to submit a new 510(k) submission or a PMA supplement to the FDA for approval before implementing these changes. The FDA's review process for PMAs is more rigorous and involves a comprehensive assessment of the new information to ensure that the modified device continues to meet regulatory standards. In cases where a device that was originally cleared through a 510(k) process undergoes significant changes that might push it into a higher risk category or significantly alter its intended

use, the FDA may require a new PMA application instead of just a new 510(k) submission. This would typically be the case if the changes affect the fundamental technological characteristics or the safety and effectiveness of the device, thereby necessitating a more comprehensive review than what is covered under the 510(k) process. However, simply submitting a PMA supplement for a system that was approved via a 510(k) is not a typical pathway. The decision to require a PMA, instead of another 510(k) submission, is based on the nature of the changes and the potential risks associated with them.

Alterations to the fundamental design or intended use of an MD may be considered significant changes. These changes often require thorough evaluation and may necessitate updated clinical evidence. Changes to software, including updates, bug fixes, or enhancements, are relevant for MDs with software components. The regulations require careful consideration of the potential impact on safety and performance. It is crucial for manufacturers to thoroughly assess and document these changes in accordance with the requirements. Depending on the nature and impact of the change, manufacturers may need to update their technical documentation, conduct additional testing, or even perform a re-approval.

### **3.2 Predetermined Change Control Plan for Machine Learning-Enabled Device Software Function**

The change management described above must also be applied to ML-DSF. This leads to the fact that only static ML models would be accepted by the FDA for approval as continual learning would lead to a modification of the approved MD, resulting in a loss of approval. Manufacturers would be forced to re-validate their devices each time the continual learning ML-DSF would adapt the ML model.

The FDA recognized this as a problem and stated the “Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions” early 2023 [27]. In that recommendation, the FDA renders the PCCP as “the documentation describing what modifications will be made to the ML-DSF and how the modifications will be assessed”. Thus, the FDA requests manufacturers to identify and assess the anticipated changes to their ML-DSF in a PCCP, which will be submitted during the approval process. If the PCCP states that the assessment of the anticipated change has no impact on the general performance or safety of the MD, then the FDA is likely to approve the continual learning-enabled MD. If the ML-DSF changes as defined in the PCCP, the device’s approval persists without the need for re-approval.

According to the FDA, the following modifications fall under the scope of the PCCP: Modifications to an ML model which is “implemented automatically (i.e., for which the modifications are implemented automatically by software)” [27], which does not explicitly involve continual learning but also would not exclude this possibility, and, modifications to an ML model which is “implemented manually (i.e., involving steps that require human input, action, review, and/or decision-making, and therefore are not implemented automatically)” [27].

It is the manufacturer’s responsibility to ensure that the changes are indeed following the PCCP. When conformance with the PCCP is erroneously assumed, the approval of the system vanishes. As such, it is in the benefit of the manufacturer to install monitoring means to ensure compliance with all previously defined requirements, especially the PCCP. To this end, manufacturers prepare Standard Operating Procedures (SOPs) that detail the ongoing monitoring and evaluation processes for PCCP-approved medical devices. These SOPs should outline how data on device performance, safety, and efficacy will be collected and managed post-market, how risk management will be conducted continuously, and how changes to the ML-DSF will be assessed and documented. The SOPs must also specify the roles and responsibilities of personnel involved in monitoring, the methods for reporting and communicating findings, and the procedures for maintaining compliance with FDA regulations. By implementing these SOPs, manufacturers can ensure that any modifications to the ML-DSF remain within the approved scope and that the device continues to meet safety and performance standards without requiring re-approval.

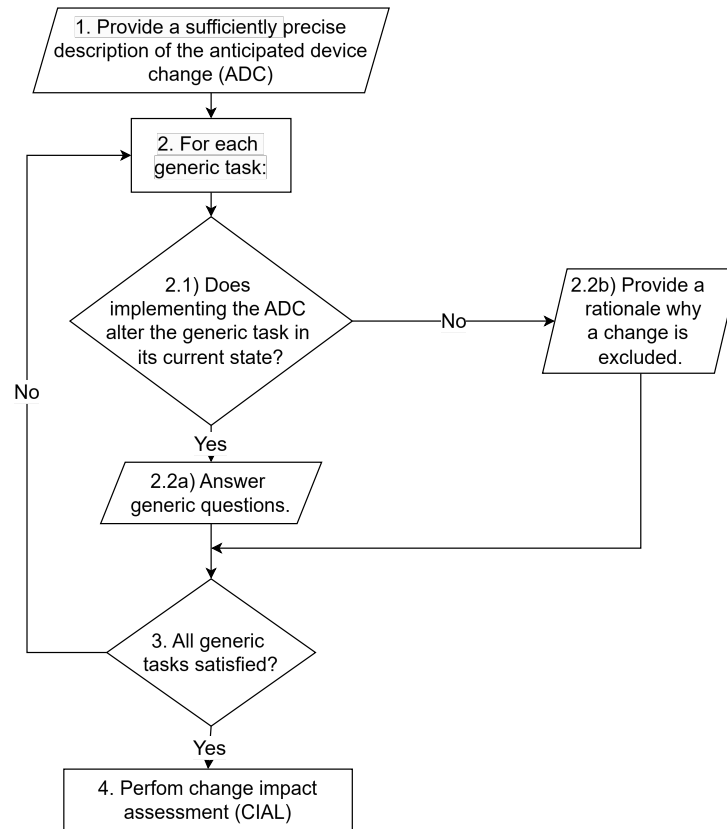
#### 4 A CRISP-PCCP

We come now to the main contribution of this paper by introducing CRISP-PCCP, the tailored process to systematically anticipate the consequences of a change of a Machine Learning-Enabled Device Software Function (ML-DSF). It aims to assist in formulating the PCCP described in FDA’s recent draft [27]. If a manufacturer follows the process, he or she will receive an estimation on how high the chance of acceptance by the FDA would be for a particular anticipated device change (ADC). Moreover, the manufacturer may use the CRISP-PCCP as a documentation input for the PCCP.

Note that CRISP-PCCP neither addresses non-ML device changes nor provides general advice on developing ML-DSFs. CRISP-PCCP focuses purely on changes to the ML component. CRISP-PCCP builds heavily on the *CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance methodology* (CRISP-ML(Q)) framework proposed by Studer et al. [19].

CRISP-PCCP is implemented by performing the following steps (see also Figure 1):

1. provide a sufficiently precise description of the anticipated device change (ADC)
2. for each generic task:
  - 1) argue whether implementing the ADC alters the generic task in its current state
  - 2a) if so, answer generic questions, interpreting them appropriately (see Section 4.1)
  - 2b) if not, provide a rationale on why a change is excluded
3. revisit generic tasks as needed (e.g. if side effects and dependencies are discovered), until satisfied



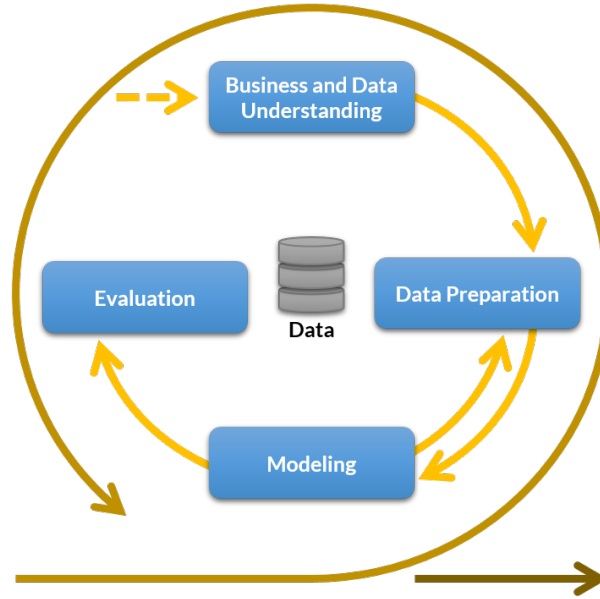
**Fig. 1.** Flowchart for Iterative Assessment and Implementation of Anticipated Device Changes (ADC) in CRISP-PCCP.

#### 4. performing a change impact assessment (see Section 4.5)

These steps must be repeated for each phase in the CRISP-PCCP. The phases of CRISP-PCCP can be seen in Figure 2. It is assumed that if a manufacturer follows this process and document the result as a PCCP request he or she will be able to demonstrate a thorough understanding and control over the anticipated changes to the device, ensuring that each modification is evaluated for its impact on safety and effectiveness.

Let us use an example to guide us through the description of the various generic tasks to come. The example is closely related to one of the prototypes we used to develop our approach. However, depending on the generic task to explain, we vary the purpose and the capabilities of the example, without restricting ourselves to the actual prototype. For more information regarding the prototype, see <https://ki-sigs.de/projekt/AP%20310> (in German).





**Fig. 2.** Overview of the phases of CRISP-PCCP and their interaction.

*Example 1.* Imagine a portable Optical Coherence Tomography (OCT) scanning device [12], which captures an image of the patient’s retina, performs a semantic segmentation to localize macular degeneration (if present) and diagnoses whether the macular degeneration is dry or wet. The computation is performed locally on the device. The image capturing and preprocessing is realized by rule-based (non-ML) software. The semantic segmentation is done by ML software based on a U-Net<sup>6</sup> model. The classification, whether macular degeneration is present and whether it is dry or wet, is performed by rule-based software. Further downstream tasks, like presenting the classification result to a user, are realized by rule-based software.

Starting at Section 4.1, we present the relevant generic tasks within our approach. We adjusted some tasks to better fit the development of ML-DSF in the medical context and assigned an impact level to each task, indicating the acceptability of the anticipated device change within a PCCP.

In the overview given above, we mentioned to answer generic questions, if a generic task is altered by an ADC. By that, we mean to answer thoroughly

- (i) *what* is going to be changed,
- (ii) *why* is the generic task going to be changed
- (iii) *where* is the change going to take place (potentially affected region, emphasis on physical location), and

<sup>6</sup> A U-Net is a special Convolutional Neural Network that was primarily developed for the segmentation of image data in medical image processing [17].

(iv) *who* is going to be responsible for the change (required qualifications still met)?

If there is no clear answer to one of those questions or the change does not affect the ML part of the ML-DSF, a rationale for that generic task has to be given to explain why the change may be excluded. Further we use a set of generic tasks as the basis, derived from CRISP-ML(Q). A manufacturer is free to add or remove generic tasks to the phases of CRISP-PCCP to fully meet the needs of their ML-DSF-enhanced product. Since CRISP-PCCP is modular, adapting the process to suit the manufacturer’s specific requirements is often beneficial. We will now walk through the relevant generic tasks of CRISP-PCCP, phase by phase.

#### 4.1 Data and Business Understanding

Quoting Studer et al. [19], this “initial phase is concerned with tasks to define the business objectives and translate it to ML objectives, to collect and verify the data quality, and to finally assess the project feasibility.” We build on this definition, but leave out the part regarding the project feasibility, as we assume in our context that the feasibility analysis has already been performed at an earlier stage of development, initiated by other requirements. Besides that, we identified several key topics to be considered when assessing the impact of an ADC, like the intended use, quality goals, and capturing processes, just to name a few.

*Intended Use* The FDA refers to *intended use* as the general purpose of a product, which is the objective intent of the legally responsible representative (e.g. the manufacturer or a reseller) who labels the product [25]. This intent may be claimed over the package or in the instruction for use, the design, or the composite of the product. The FDA offers a tool that shall help to determine whether a product’s software functions may fall potentially under the scope of the FDA’s oversight [24].

*Scope* By scope, we mean the specific functionality of the ML-DSF that is achieved through an ML model. It is important to distinguish between the intended use and the scope, as the scope may change without affecting the intended use. The following argument explains this distinction: The scope of Example 1 covers performing semantic segmentation to localize macular degeneration, given an RGB image. Stating whether macular degeneration is present (e.g. if the segmentation exceeds some threshold), or even deciding whether the degeneration is wet or dry, is not determined by ML software and therefore not part of our scope. An ADC is to replace the non-ML classification software part with an ML model, leaving everything else as is. The scope is affected but not the intended use.

Therefore, in our Example 1, the generic questions must be answered, which we do in Table 1. For brevity reasons, we omit the questions in the remaining examples.

**Table 1.** Answers to generic questions for an ADC affecting the scope of Example 1.

What?	The rule-based component which classifies the segmentation is replaced by an ML model.
Why?	Experiments indicate a higher accuracy of the ML model in comparison to the rule-based approach.
Where?	All devices that have the ML-DSF or for which it is subsequently made available by means of an update.
Who?	ML engineer in the development unit, test engineer during integration test

*Intended Patient Population* Citing the patient-focused drug development glossary<sup>7</sup>, the intended patient population is defined as “the group of individuals (patients) about whom one wishes to make an inference.”

To give an example how the intended patient population is affected by an ADC, assume that in the case of Example 1, the initial intended patient population is defined as persons of age 50 to 70 years. However, anticipating that one year after market submission, enough training data will be available of persons of age 40 to 80, the intended patient population is extended to that range.

*Quality Goals* We highlight several quality goals, relevant to ML in the medical field (derived from the success criteria mentioned in CRISP-ML(Q) [19]). Namely, the *diagnosing performance goal*, *runtime performance goal*, *robustness goal* and the *human understandability goal*. We emphasize the term *goal* in these cases, as we are interested in the intended consequences of an ADC. When applying our method to the example, we obtained several useful insights: Most ADCs can impact certain qualities in some way, even if that was not the original intention. To avoid meaningless answers to our questions, we adopt a specific approach. Often, the side effect of an ADC on a specific quality (not the goal) cannot be ruled out. Therefore, we focus only on intended consequences, which we refer to as quality goals. This means we aim to maintain the effects on diagnosis performance, runtime performance, robustness, and human understandability within reasonable limits. It is worth noting that existing quality assurance processes already cover these aspects.

Ideally, every goal is defined by a measurable metric or verifiable [27,1,16]

By diagnosing performance we mean properties like accuracy, specificity, sensitivity, and so on Runtime performance includes computation time, memory and storage consumption, and energy requirements. Robustness denotes “the degree to which a component can function correctly in the presence of invalid inputs or stressful environmental conditions,” see [7]. By human understandability, we refer to “the ability to explain or to present in understandable terms to a human”, see [3].

<sup>7</sup> Refer to <https://www.fda.gov/drugs/development-approval-process-drugs/patient-focused-drug-development-glossary>.

*Capturing Process and Digital Representation* By capturing process we mean the transition of turning a physical object or an action into a digital representation. The digital representation is the result of the capturing process, usually the output of the capturing device. For example, recording an image is a capturing process and the image file is the digital representation.

Considering again the above mentioned portable OCT device. Assume that the training set of the original device (OD) consists of images captured by a stationary OCT scanner in a lab environment, creating images of high quality. Expanding the training set by retina images collected during production use of the portable OCT device, which by construction generates images of lesser quality and operates in diverse environments, is an example of a change in the capturing process and its digital representation.

## 4.2 Data Preparation

Data preparation covers all tasks involved in the transformation of data in its digital representation (see above) to a form which is accessible by ML models. This includes tasks operating directly on the digital representation, like selection, cleaning, and imputation, but also conversion steps like construction, integration and formatting.

In contrast to CRISP-ML(Q), we do not divide this phase into smaller parts (generic tasks). One reason is that in practice, we experienced that data preparation (preprocessing) closely intertwines the mentioned generic tasks. For example, many deep learning frameworks provide functions which convert JPEG images to a multidimensional array of floating point values in the range of  $[0, 1]$ . These functions combine construction and formatting into one step. Often, they allow to integrate cropping, linearly transforming, and normalizing functionality, which then also covers selection and cleaning. Another reason is that for our approach, it is not necessary to distinguish between the generic tasks, as they are treated equally regarding their Change Impact Assessment Level (refer to Section 4.5 for further details).

## 4.3 Modeling

By modeling we mean the declaration (implicitly and explicitly) of a space of learnable functions which is later systematically searched for a good (in terms of some metric) solution. To find, at least in principle, a good solution within said space, it is necessary that the space of learnable functions contains a good solution in the first place (the space should not be too small). To find the solution in a reasonable amount of time, it also should not be too large. Thus, it is important to carefully select the space of learnable functions.

The following paragraphs describe the principles that affect its size and its content.

*Domain Knowledge & Data Assumptions* Domain knowledge and data assumptions incorporated into the OD may not be valid with regards to an ADC. For example, assume that the image capturing device of Example 1 can record more image modalities than just those by cameras using visible light. The other modalities were recorded during production use and integrated into the training set, which did not contain such images during the development of the OD. The segmentation component of Example 1 will no longer be a two-dimensional U-Net, but multidimensional. Domain knowledge and data assumptions that expected two-dimensionality may no longer be valid for higher dimensions.

*Modeling Technique* By modeling technique, we mean the selection of the class of models to choose from. For example, modeling the scope via decision trees is a modeling technique. Modeling it instead via fully connected deep neural networks is another. Pre-training, transfer learning, and assembling also fits into this generic task.

The modeling technique is affected by an ADC, if for example a novel feed forward network module is introduced in the literature and the manufacturer decides to incorporate it into the ResNet<sup>8</sup> of the OD.

*Tuning Procedure* Identifying the space of potential learnable functions constitutes an initial step within the model development framework, whereas the precise selection of a singular function from this space encompasses a distinct and complex challenge. By tuning we mean the guided selection of a learnable function, typically facilitated by a designated dataset known as tuning data<sup>9</sup>. This selection process commonly adopts the formulation of an optimization problem aimed at evaluating and ranking the candidate functions within the specified space. Usually, the candidates of a given space are ranked along a specifically formulated optimization problem. By defining loss functions and regularization terms, candidates are valued, preferring the ones with higher value (or equivalently, lower loss). To systematically (and hopefully efficiently) search only for promising candidates, optimizer are applied. Depending on the way the optimizer operates (initial solution, local optimization, global optimization, . . .), some candidates are effectively excluded. Considering again the example we gave for the scope, replacing a non-ML unit with an ML unit will affect the tuning procedure of the OD, or introduce a second one.

*Reproducibility* The modeling is reproducible if, based only on the modeling documentation, the previously learned function can be identically recreated. Often this fails due to implicit data assumptions, non-written domain knowledge, or (hidden) randomness.

---

<sup>8</sup> A ResNet (Residual Network) is a type of Convolutional Neural Network specifically designed to train deeper networks by addressing the vanishing gradient problem [6].

<sup>9</sup> This is often called validation data, when splitting the available data into the training, validation, and test set. We prefer, and so does the FDA, the term tuning, as it avoids confusion with the meaning of validation in the medical context.

Looking again at our running example, an ADC might add more stochasticity to the ML model (e.g. by introducing variational parts). If the manufacturer misses to keep track of the randomness passed through the training, a later reproduction fails.

#### 4.4 Evaluation

In the evaluation phase, we check whether the performance, robustness, and human understandability goals, defined in phase 1, are met by the learned function obtained from phase 3. In its draft of the PCCP [27], the FDA states that the modification protocol describes “[...] the methods that will be followed when developing, validating, and implementing modifications [...].” Thus, it is mandatory to examine whether the evaluation procedures for the OD are still suitable for evaluating an ADC.

For example, an ADC of the portable OCT device mentioned above is to diversify the intended patient population, by adding corresponding training data which will be acquired after the first market admission. Additional *diagnosis performance* evaluation explicitly targeting the newly affected patient population is necessary.

#### 4.5 Change Impact Assessment and Summary

In the previous sections, we presented those generic tasks of the CRISP-PCCP phases, which we deemed appropriate for our approach. Treating each of these generic tasks and eventually answering the related questions gives a rather detailed view of the impact an ADC may have. However, a broad picture or a conclusion might be unclear. For this reason, we propose a systematic method to condense the change impacts into a single number: the Change Impact Assessment Level (CIAL).

The FDA identified in [23] three broad categories of changes of an ML-DSF: performance changes, input changes and intended use changes. In our view, changes related to the ML-DSF’s performance generally have the highest chance of being compatible with a PCCP (meaning that the FDA will probably accept anticipated changes of this kind in most cases). For example, the usage of additional training data from the intended patient population, gathered from field usage to increase the accuracy of the ML-DSF, is a change that is likely to be compatible with a PCCP. To such a change we assign the Change Impact Assessment Level (CIAL) **3**.

For changes related to the ML-DSF’s input, e.g. the dimensions or resolution of input images and inclusion of additional features, we assume the chance of being PCCP compatible to be on par with the chance of being PCCP incompatible. The broad range of ways to change the input with (more or less) wide-ranging effects justify our view. Therefore, we assign to each input change the CIAL **2**. For each CIAL 2 change individually, we suggest consulting experts and/or reaching out to the FDA as early as possible, to increase PCCP compatibility chance of that change.

**Table 2.** The CIAL compatibility classes per generic task (*or phase*). The higher the class, the more likely we deem a corresponding change to be PCCP compatible (3 being the highest, 1 being the lowest).

Generic Task ( <i>phase</i> )	CIAL
<i>Data and Business Understanding</i>	
Intended Use	1
Scope	2
Intended Patient Population	2
Quality Goals	3
Capturing Process & Digital Representation	2
<i>Data Preparation</i>	
<i>Modeling</i>	
Domain Knowledge & Data Assumptions	2
Modeling Technique	1
Tuning Procedure	2
<i>Evaluation</i>	
	1

Changes to the ML-DSF’s intended use have, according to the FDA ([27, p. 17]), a low (but non-zero nevertheless) chance of being PCCP compatible. We denote the CIAL of such changes by **1**.

Guided by the principles above, we assign to each generic task (or phase) of the CRISP-PCCP process model the most fitting change category (performance changing, input changing, intended use changing), yielding the corresponding CIAL (see Table 2).

By definition, the *intended use* will receive a CIAL of 1.

To the *evaluation phase*, i. e. the evaluation of the four quality goals diagnosis performance, runtime performance, robustness, and human understandability, we also assign a CIAL of 1. The reason is that the approval of a medical device strongly correlates with the degree to which the ML-DSF fulfills the quality goals and the validity of the degree itself depends on the comprehensiveness and thoroughness of the evaluation method [18,22].

Since increasing the *scope* of an ML-DSF probably lessens the PCCP compatibility, but decreasing the scope of an ML-DSF probably raises the PCCP compatibility, we assign the CIAL 2, to take the unclear situation into account. Considering the case individually may allow for a change of the CIAL to 1 or to 3.

If changing the *intended patient population* resembles rather an extension of the existing population (e. g. enlarging the age interval to both sides) and supporting arguments to do so exist, we suspect a high probability of PCCP compatibility. On the other hand, if the intended patient population is expanded by a rather “orthogonal” group, that only has a minor overlap with the existing population, many other generic tasks would presumably also be affected by this expansion, which in turn make it difficult to argue for PCCP compatibility.

Another generic task with CIAL 3 is *quality goals*. We assume that changing the quality goals will only result in more sophisticated goals (never less sophisticated ones), for economic reasons, and accordingly expect a high PCCP compatibility. The *tuning procedure* is a generic task that would fall under CIAL 2 with the potential to tend to a CIAL 1. On the one hand, most of the time the tuning procedure only intends to increase the performance or to find even better solutions to the optimization problem. On the other hand, the solution is based on actions performed by the manufacturer and falls more in the category of a software change.

The generic tasks *capturing process & digital representation*, *data preparation*, *domain knowledge & data assumption* and *tuning procedure* each encompass a diverse range of possible changes, for which we do not see ourselves in a position to proclaim either a high or a low probability of approval. That is why we assign the CIAL to all of them, resorting to the consideration of individual cases.

The last step of the change impact assessment is to check for the lowest CIAL from the generic tasks. To do so, first select only those generic tasks, which are affected by the ADC, i. e. which have answers to the generic questions and no rationales supporting their exclusion. Second, order them ascending by their associated CIAL. If there is a CIAL of 1 the chance is low that the FDA would accept the ADC for approval. If there is a CIAL of 2 the chance of FDA approval increases but it is recommended to get in touch with the FDA to discuss further conditions. And finally, if there is a CIAL of 3 the chance that the ADC will be approved by the FDA is high.

One should carefully consider whether it is worth the effort to continue the approval procedure of an ADC with overall CIAL 1. An ADC of overall CIAL 2 is more promising in that regard, but we strongly suggest involving experts and/or the FDA early in the further development process. An ADC of overall CIAL 3 is likely PCCP compatible.

## 5 Discussion

The CRISP-PCCP process was initially developed in response to the FDA’s Proposed “Regulatory Framework for Modifications to ML Software as a Medical Device” [23], as part of the BMWK-funded KI-SIGS project [10]. It was evaluated and refined in three different projects, including “PASBADIA” [21] and two others from the KI-SIGS initiative, all focusing on ML-supported medical devices.

Comprising ML and regulatory experts from the KI-SIGS project, the working groups applied CRISP-PCCP to address specific ML challenges in their respective projects. They aimed to identify and assess potential changes within the ML processing pipeline. The methodology included introducing the CRISP-PCCP’s background and goals, followed by its application. Using the current development state of the corresponding project as a baseline, upcoming development goals were identified and examined using CRISP-PCCP. This process helped in easily identifying changes outside the ML development scope,



encouraging the teams to define a rationale at the end of each investigation, which could be incorporated into technical documentation as proof of proper PCCP application.

The CRISP-PCCP process proved to be a relevant procedure for the documentation of ML development results. The underlying CRISP-ML(Q) process enabled detailed examination across various ML development phases. The CRISP-PCCP template aided in posing important questions and assessing potential changes. The ML developers from different research and development projects, which were involved in the development process of the CRISP-PCCP, independently confirmed the process’s utility in evaluating changes in ML. Regulatory experts viewed CRISP-PCCP as a useful method for capturing and assessing changes and their impact on the ML model. CRISP-PCCP has potential beyond continual learning ML applications, such as being part of the development planning for ML models while identifying associated risks. Changes to the ML model can be assessed, and implementation planning can be based on the CIAI, prioritizing changes with a higher chance of acceptance.

CRISP-PCCP can also play a role in the European jurisdiction. The support for change management processes and ML development planning is also required by the Medical Device Regulation (MDR) [20], which CRISP-PCCP can assist with. Additionally, there is a questionnaire titled “Artificial Intelligence (AI) in medical devices” published by the Interest Group of Notified Bodies [8], intended for auditors to ask questions during audits about the ML product development life cycle. Manufacturers can also use this document to identify gaps in their documentation. CRISP-PCCP, due to its structure, can then be expanded to cover identified gaps and serve as appropriate development documentation.

## 6 Conclusion

CRISP-PCCP is a process adapted from CRISP-ML(Q) [19], selectively incorporating its first four phases reflecting the developmental stage of projects. It assumed that the Deployment and Quality Assurance phases from CRISP-ML(Q) are integrated in the existing development process. Despite that, CRISP-PCCP demonstrated its potential in meeting FDA’s PCCP [27] requirements in selected projects. However, its application to projects involving Continual Learning ML models and validation by the FDA itself remains pending, mainly due to the limited number of such projects in medicine.

The integration of CRISP-PCCP in the European jurisdiction, considering the evolving AI-Act [4], poses future research questions. The AI-Act, recently passed in the European Union, seeks to regulate artificial intelligence. It aims to establish a legal framework ensuring AI systems’ safety, compliance with privacy and data protection laws, and the upholding of fundamental rights. The co-existence of the AI Act and the Medical Device Regulation imposes many questions on how to address similar overlapping concerns and requires future research.

## References

1. Altman, D.G., Bland, J.M.: Diagnostic tests. 1: Sensitivity and specificity. *BMJ (Clinical research ed.)* **308**(6943), 1552 (Jun 1994). <https://doi.org/10.1136/bmj.308.6943.1552>
2. Bende, P., Vovk1, O., Caraveo, D., Pechmann, L., Leucker, M.: A Case Study on Data Protection for a Cloud- and AI-based Homecare Medical Device. In: Lamo, Y., Rutle, A. (eds.) *Proceedings of The International Health Data Workshop. CEUR Workshop Proceedings*, vol. 3264. CEUR, Bergen, Norway (Jun 2022), [https://ceur-ws.org/Vol-3264/#HEDA22\\_paper\\_3](https://ceur-ws.org/Vol-3264/#HEDA22_paper_3), iSSN: 1613-0073
3. Doshi-Velez, F., Kim, B.: *Towards a rigorous science of interpretable machine learning* (2017)
4. European Union: EU AI Act: first regulation on artificial intelligence | News | European Parliament (Jun 2023), <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>, accessed: 08/12/2023
5. Hatcher, W.G., Yu, W.: A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends. *IEEE Access* **6**, 24411–24432 (2018), <https://ieeexplore.ieee.org/document/8351898>
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). <https://doi.org/10.48550/arXiv.1512.03385>, <http://arxiv.org/abs/1512.03385>, arXiv:1512.03385 [cs]
7. IEEE: IEEE Standard Glossary of Software Engineering Terminology. *IEEE Std 610.12-1990* pp. 1–84 (1990). <https://doi.org/10.1109/IEEESTD.1990.101064>
8. IG-NB: Questionnaire „Artificial Intelligence (AI) in medical devices“, [https://www.ig-nb.de/?tx\\_epxelo\\_file\[id\]=884878&cHash=53e7128f5a6d5760e2e6fe8e3d4bb02a](https://www.ig-nb.de/?tx_epxelo_file[id]=884878&cHash=53e7128f5a6d5760e2e6fe8e3d4bb02a), accessed: 12/12/2023
9. International Electrotechnical Commission: IEC62304:2006/AMD1:2015 Amendment 1—Medical Device Software—Software Life Cycle Processes. <https://webstore.iec.ch/publication/22790> (2015), accessed: 03/30/2021
10. KI-SIGS: AI Space for Intelligent Healthcare Systems KI-SIGS, <https://ki-sigs.de/>, accessed: 02/01/2022
11. Obermeyer, Z., Emanuel, E.J.: Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine* **375**(13), 1216–1219 (Sep 2016), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5070532/>
12. Ong, J., Zarnegar, A., Corradetti, G., Singh, S.R., Chhablani, J.: Advances in Optical Coherence Tomography Imaging Technology and Techniques for Choroidal and Retinal Disorders. *Journal of Clinical Medicine* **11**(17), 5139 (Jan 2022). <https://doi.org/10.3390/jcm11175139>, <https://www.mdpi.com/2077-0383/11/17/5139>, number: 17 Publisher: Multidisciplinary Digital Publishing Institute
13. Pechmann, L., Mildner, M., Suthau, T., Leucker, M.: Regulatorische Anforderungen an Lösungen der künstlichen Intelligenz im Gesundheitswesen. In: Pfannstiel, M.A. (ed.) *Künstliche Intelligenz im Gesundheitswesen: Entwicklungen, Beispiele und Perspektiven*, pp. 175–198. Springer Fachmedien Wiesbaden, Wiesbaden (2022), [https://doi.org/10.1007/978-3-658-33597-7\\_8](https://doi.org/10.1007/978-3-658-33597-7_8)
14. Petersen, E., Potdevin, Y., Mohammadi, E., Zidowitz, S., Breyer, S., Nowotka, D., Henn, S., Pechmann, L., Leucker, M., Rostalski, P., Herzog, C.: Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions. *IEEE Access* **10**, 58375–58418 (2022), <https://doi.org/10.1109/ACCESS.2022.3178382>, conference Name: IEEE Access

15. Pianykh, O.S., Langs, G., Dewey, M., Enzmann, D.R., Herold, C.J., Schoenberg, S.O., Brink, J.A.: Continuous Learning AI in Radiology: Implementation Principles and Early Applications. *Radiology* **297**(1), 6–14 (2020), <https://doi.org/10.1148/radiol.2020200038>, pMID: 32840473
16. Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation (Oct 2020), <http://arxiv.org/abs/2010.16061>, arXiv:2010.16061 [cs, stat]
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (May 2015), <http://arxiv.org/abs/1505.04597>, arXiv:1505.04597 [cs]
18. Stewart, J.P.: Software as a Medical Device (SaMD): Clinical Evaluation. International Medical Device Regulators Forum (2017)
19. Studer, S., Bui, T.B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., Müller, K.: Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology. *Mach. Learn. Knowl. Extr.* **3**(2), 392–413 (2021), <https://doi.org/10.3390/make3020020>
20. Union, E.: Regulation (eu) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, amending directive 2001/83/ec, regulation (ec) no 178/2002 and regulation (ec) no 1223/2009 and repealing council directives 90/385/eec and 93/42/eec (text with eea relevance. ) (April 2017), <https://lexparency.de/eu/32017R0745/>, accessed: 11/17/2022
21. University of Lübeck: PASBADIA: COPICOH, <https://www.copicoh.uni-luebeck.de/forschung/projekte/aktuelle-projekte/pasbadia>, accessed: 12/08/2022
22. U.S. Food and Drug Administration: Software as a Medical Device (SaMD) (Dec 2018), <https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd>
23. U.S. Food and Drug Administration: Proposed Regulatory Framework for Modifications to AI/ML Software as a Medical Device (Apr 2019), <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
24. U.S. Food and Drug Administration: Digital Health Policy Navigator (Dec 2022), <https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-policy-navigator>
25. U.S. Food and Drug Administration: Cfr - Code of Federal Regulations Title 21 Part 801.4 (Jun 2023), <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcr/CFRSearch.cfm?fr=801.4>
26. U.S. Food and Drug Administration: Deciding when to submit a 510(k) for a change to an existing device. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/deciding-when-submit-510k-change-existing-device> (2023), accessed: 11/29/2023
27. U.S. Food and Drug Administration: Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions (Apr 2023), <https://www.fda.gov/media/166704/download>
28. Vokinger, K.N., Feuerriegel, S., Kesselheim, A.S.: Continual learning in medical devices: FDA’s action plan and beyond. *The Lancet Digital Health* **3**(6), e337–e338 (Jun 2021), <https://linkinghub.elsevier.com/retrieve/pii/S2589750021000765>