

Whitepaper

How to build medical AI:

Combining best practices from machine learning and medical device development

Dr. Gabriel Krummenacher, Lead Data Scientist
Dr. Stefan Weiss, MBA, Lead Business Innovation Consultant Pharma & MedTech
Dr. Patrick Steiner, Project Manager Health Tech
Uwe Szymanski, Lead Regulatory Expert and Software Architect

Table of Contents

4 Abstract

5 Introduction

7 The Past, Present and Future of Medical AI

- 9 A proven track record – image analysis and big data processing
- 11 Early adopters required – Federated learning and language processing
- 12 AI offers little benefit – Isolated applications where suitable data is lacking

13 Regulatory Considerations and Current Regulatory Environment

- 15 Current regulatory environment
- 16 General considerations
- 18 Medical AI development best practices and guidelines
- 19 FDA framework
- 20 The need for more guidelines

21 Principles of AI in Health Tech

- 23 Principle 1: High data and labelling quality
- 25 Principle 2: Good machine learning practices
- 25 Principle 3: Simple, transparent, interpretable models
- 27 Principle 4: Thorough evaluation
- 29 Principle 5: Reproducible machine learning pipelines
- 29 Principle 6: Manage data, model and code history
- 31 Principle 7: Comprehensively document code, data, models and design decisions
- 31 Principle 8: Communicate results and uncertainty carefully

33 Conclusion and Outlook

35 Glossary

36 Authors

37 References

Abstract

AI-driven applications and devices are expected to deliver significant improvements to healthcare systems. It is therefore no surprise that this fast-developing field is attracting a great deal of attention. Despite this strong market pull and proven ability to deliver game-changing insights and outcomes, medical AI applications are still rare. A key obstacle to wider commercial success is a lack of detailed regulatory guidelines for developing and maintaining medical AI solutions. This is leading to significant delays in bringing much-needed digital innovation to healthcare – with a significant impact on patient health. To facilitate a successful market entrance of AI-driven devices and software, in this white paper we draw on our experience from numerous AI development projects. The white paper is divided into three parts:

In [part one](#) we review key milestones in medical AI, evaluate existing applications and share our perspective on current trends.

In [part two](#) we discuss existing regulations relevant to medical AI and highlight areas in need of further refinement.

In [part three](#) we share our 8 best practice principles for developing market-ready medical AI applications. These combine insights from machine learning and medical device development.

Our aim is to improve development of user-centric applications including AI-driven medical devices and clinical applications.

Introduction

If we consider recent technological developments in the [Health Tech ↗](#) field, few concepts have influenced how we define, develop and imagine future healthcare systems as profoundly as [artificial intelligence \(AI\) ↗](#). Less than 6 years after the first AI-enabled [medical device ↗](#) was approved (AliveCor 2014)¹ ↗, AI has become omnipresent in the field. Almost all major Health Tech corporations are currently considering or planning to use some form of AI in future products and services. Looked at more closely, this phenomenon should not come as a surprise. AI concepts and methods have been used successfully in a wide range of domains and across many different industries for some years, and statistical approaches have also been in use in medical image analysis solutions for some time. In retrospect it was an obvious match.

In Health Tech, the use of statistical methods such as AI has been further fuelled by the success of computer vision, the digitization of medical records and diagnostic data, and a sharp increase in the amount of patient-generated data being produced by apps and wearables. Recently, the trend started to point towards using AI in a broader range of applications with examples being automated patient diagnosis and automated processes throughout hospitals and clinics. But rapid growth in AI-driven applications has its drawbacks. AI is still a relatively new technology and a lot of medical device manufacturers and regulating bodies do not yet fully understand the opportunities and challenges it brings.

While AI looks to be an excellent fit for the healthcare sector, there are some important obstacles to overcome – obstacles that are not present (or present to a far smaller degree) in non-regulated environments. In contrast to most other applications of AI, incorrect predictions made by an automated system forming part of a medical device can cause direct harm. While the cost of a mistake by an algorithm aimed at predicting whether a user will click on an online advertisement is small and purely monetary, the consequences of incorrect diagnosis and treatment can be fatal.

To manage and minimize the risk to patients, governments around the world regulate the development and approval of medical devices and software. These regulations apply equally to medical AI solutions. Unfortunately, at the present time there are no standards for developing medical AI devices and very few guidelines exist. The result is considerable uncertainty. Correct deployment and evaluation of AI systems is complex, and AI systems pose a major risk to patient safety if [models ↗](#) are biased, non-robust or not properly validated. To date, very few randomised clinical trials of deep learning systems have been conducted (Nagendran et al., 2020).

In this white paper, we will discuss challenges and opportunities for AI in Health Tech, offer insights into [regulated ↗](#) development of [machine learning ↗](#) approaches, and share the best practices we have identified in running real world projects. We have therefore structured the paper into three sections:

In [Section 2 ↗](#), we discuss past and existing Health Tech applications where AI has already proven its worth. We also highlight the areas where, based on our experience, we see the biggest opportunities for AI innovation in the near future.

In [Section 3 ↗](#), we focus on regulatory considerations for approvable AI. We discuss current regulatory requirements and highlight areas for further refinement.

In [Section 4 ↗](#), we share our Zühlke best practice principles for developing commercial medical AI solutions. Best practices are at the core of any AI project aiming for long-term impact, but need additional consideration when applied to the medical field.

In conclusion, this white paper is aimed at helping you to create medical AI solutions that are safe, reliable, satisfy regulatory requirements and can truly improve patients' lives.

The Past, Present and Future of Medical AI

Right now, the medical device/medical software field is bursting with ideas and concepts for AI solutions. Substantial sums of money are clearly being poured into research on potential applications for AI. While this is a positive development, it should be noted that AI – like any other technology – is not always the golden solution for every problem. AI is inherently better suited to repetitive tasks involving large amounts of variable data or applications requiring high accuracy and rapid analysis. Such applications also tend to fulfil the crucial requirement of having labelled data available for development and training. Highly specialized tasks performed only infrequently do not meet this key requirement. This is quickly apparent if we look at the fields in which medical AI applications are considered in the AI literature (Figure 1) ⁷.

A proven track record

Image analysis and big data processing

If we look at some of the first AI-based medical devices to be approved, one thing stands out. A substantial majority involve medical image analysis

Extract from the list of the earliest medical softwares using AI for medical image analysis

- AliveCor; Atrial fibrillation detection; September 2014
- Arterys; Cardiac MRI interpretation; January 2017
- MaxQ-AI; CT-based diagnosis of brain haemorrhages; January 2018
- Viz.ai; CT-based stroke diagnosis; February 2018

That it was in the medical imaging field that the idea of exploiting the potential of AI first gained momentum is unsurprising. Medical imaging has long been a key driver in the development and utilization of automated and feature-driven analysis of large quantities of image data (CT, MRI, OCT scans, etc.). These solutions frequently employ statistical approaches. Software designed to support medical image analysis has an impressive track record and has earned substantial trust from [health care professionals \(HCP\)](#). So the degree of resistance to taking the plunge with AI has been notably smaller than in other fields. Disciplines such as radiology, pulmonology and ophthalmology provide vast, meticulously labelled datasets on which to test new approaches. For

HCPs familiar with computer-assisted image analysis, accepting the idea of a clinical decision support system (CDS) – which is what almost all currently approved AI medical software applications claim to be – is a relatively small step. Systems describing themselves as CDSs aim to provide HCPs with fast, reliable information on which to base diagnostic and treatment decisions.

The key benefit offered by these systems relies on the fact that medical image analysis often involves examining and interpreting large sets of images, where relevant features may be hard to detect, especially if the examiner is tired. Additionally, diagnostic sensitivity and specificity is rarely dependant on single features, but is often improved by taking into account multiple features and relations between features and feature groups.

If your application fulfils the following conditions, there is a good chance that AI could be a valuable extension to your product:

- Large, structured, standardised datasets are available
- Image analysis is used to identify, highlight and classify features for diagnostic or other decision-making
- Analysis involves browsing through large datasets with potentially small or weak features

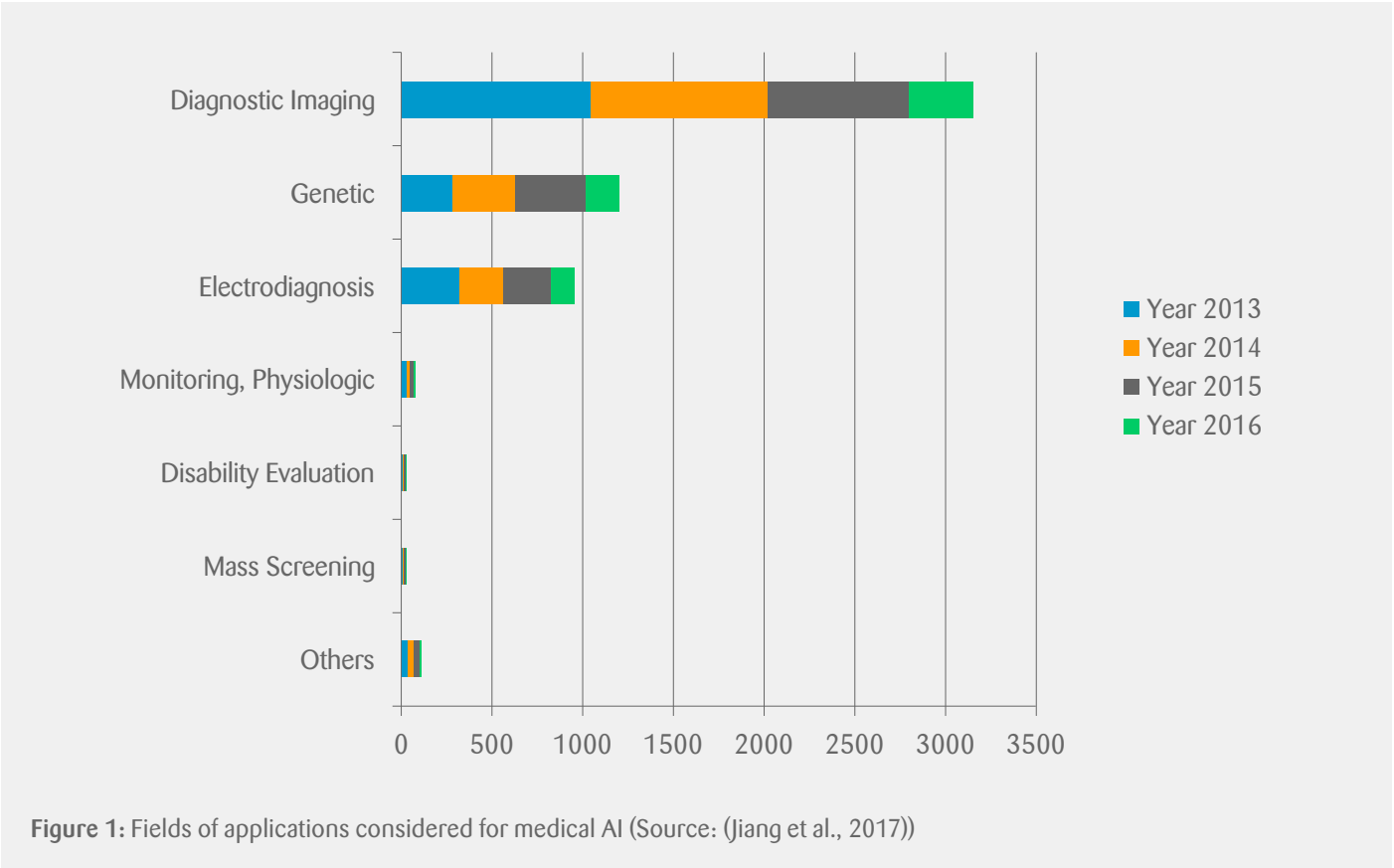


Figure 1: Fields of applications considered for medical AI (Source: (Jiang et al., 2017))

Early adopters required

Federated learning and language processing

In addition to image processing and analysis, AI has recently reached the point where it becomes feasible to consider its use for a number of other medical applications, including [natural language processing](#) and [federated learning](#). These approaches have achieved promising results in research studies and in proofs of concept for health care applications, but are not yet in widespread clinical use. They reason they lag behind image analysis is that the technology and methods are relatively new and are still being researched and because the structured data needed for these applications has only just started to be collected and provisioned.

Research and development of these applications is to some extent also hindered by privacy concerns and the value of individually identifiable health information. [Protected health information \(PHI\)](#) is subject to the highest protection and anonymisation requirements, making it difficult to create and provision large datasets while at the same time ensuring that the data ownership remains with the patient. Patient anonymisation meets some PHI privacy requirements and is widely used in clinical information systems. But such approaches are often designed to work only if data stays within the system managing the anonymisation.

For patient reports and clinical documentation – where natural language processing would offer significant benefit – anonymisation is inherently more complex, as, in addition to metadata, patient specific data is also present

in the text being analysed. Patient specific data is often the most important attribute to be extracted from the text. Storing such data in a publicly accessible database is therefore almost impossible. Federated learning is an attempt to overcome these limitations. It involves allowing models to be individually trained on isolated datasets, sharing only the trained models, which do not contain any personal information. Promising results have been achieved in a research environment (see [MELLODDY](#)), but the concept has yet to prove itself in a productive and regulated environment.

Once the technology and the approaches to properly handle personalized data are defined, by automating much everyday health care paperwork, such as patient reports, literature searches, analysis of study protocols, etc., AI could facilitate a major breakthrough. One potential application is disease subtyping for precision medicine. While various studies have shown promising results using proteomics and molecular data (Sinkala et al., 2020), other results suggest that such models also need to include electronic patient histories and longitudinal health data (Saria and Goldenberg, 2015; Zhang et al., 2019). However, this again raises issues of data ownership, responsibility for data and anonymisation.

Despite some uncertainties, early adopters could gain a highly rewarding head start if they now lay the foundations of future projects including concepts for sharing results while retaining full ownership of valuable data.

AI offers little benefit

Isolated applications where suitable data is lacking

While AI has shown itself to be one of the most promising developments in the Health Tech field over the last decade, it's important to acknowledge its limitations. As with any other technology, AI may be unsuitable for certain problem sets or simpler, more efficient methods may be available. Focusing solely on AI solutions just because AI is today's next big thing would be counterproductive.

One of the biggest obstacles to using AI is a lack of available data. Because AI is, by its nature, a statistical method, suitable, well-labelled data sets are essential for achieving the kind of model and algorithm performance required. This is even more true for medical AI, where reliable performance is critical and needs to be demonstrated to regulatory authorities. Even where data is available, selection bias or poor labelling may mean that the quality of the data set (see also [Principles of AI in Health Tech](#)) is inadequate.

Data may be lacking because the appropriate data has not been collected and structured or simply because the

number of patients or treatment events is limited. Where this is the case, a simple expert system or standard operating procedures may have a comparable or better impact on decision making. In this case, the data processing, model development and algorithm training workload is simply not justified. Questions should be asked about whether AI is the appropriate tool if the application meets any of the following criteria:

- low incidence surgery or treatment
- diagnosis and treatment of rare diseases
- situations involving extensive patient interaction
- procedures with low cases numbers and large variability

AI solutions can still be valuable for the above applications (or aspects thereof), but before investing in model development the feasibility of the proposed approach should be carefully evaluated.

Regulatory Considerations and Current Regulatory Environment

Any treatise on new technology in the Health Tech environment would not be complete without addressing regulatory requirements. This is certainly no different for AI. Working with emerging technologies often pushes the limits of the existing regulations, because they no longer or not completely cover new approaches and concepts. The following section analyses the current regulatory environment and discusses best practices and potential pitfalls as we see them.

Current regulatory environment

Because of their legislative nature, regulations – and medical regulations in particular – exhibit a high degree of inertia and often fail to keep pace with emerging technologies. This is certainly also the case with medical AI. Existing legislation fails to take account of specific features of this novel technology. While the FDA has risen to the challenge and published a proposed regulatory framework for AI³, at the time of writing (October 2020) the EU has been notably quiet about this issue. The European Commission appointed 52 experts to the High-Level Expert Group on AI (AI HLEG) in 2018, but still only few guidelines have been published and even the current centrepiece, the Ethics Guidelines for Trustworthy AI⁴ remains a draft guideline only (as at October 2020).

In the European Union, the end of the transition period from the MDD to the MDR in May 2021 adds an additional layer of complexity. Because the MDR will place many standalone software in a higher risk class than the old MDD, manufacturers are still having their products approved as class 1 devices under the older directive. This is a reasonable approach, since it allows manufacturers to start collecting additional data (e.g. post-market clinical follow-up data) crucial for approval under the MDR.

General considerations

As discussed above, at present there are no harmonized standards aimed at specifically addressing the issue of AI in medical software. This is not expected to change in the near future. So what should manufacturers of AI products do?

Quality assurance standards generally require the use of best practices and state-of-the-art development. In the absence of more specific guidelines, that means nothing more and nothing less than looking at existing regulations and applying them to the maximum extent possible to your new problem set. In short, it's essential that you comply with existing regulations and essential requirements applicable to all medical device development:

- There must be a well-formulated [intended use](#). The performance of the medical device in general and the AI algorithm must be verified and validated for the specific intended use.
- Key and quality-critical parameters must be defined and described and it must be demonstrated that the device performs as claimed (sensitivity, specificity, etc.).
- Software must be developed in accordance with existing harmonized standards.
- The development must be properly planned and documented. The methods used for development, verification and validation must be specified.
- There must be a quality assurance process in place and people working on development must be capable of performing the tasks assigned to them.
- Development must be risk-based. There must be proper risk analysis and mitigations must be defined and verified. The legal manufacturer is responsible for any harm caused by the medical device.

Existing, non-AI specific standards and regulations are a valuable source of information for the above points. The following standards and regulations should be taken into account whatever technology or concepts are used:

- IEC 62304 for medical software lifecycle management
- ISO 14971 for risk management
- ISO 13485 for quality systems
- 21 CFR and/or MDR/IVDR for quality management and general requirements for medical devices

While existing software development and risk management standards are a useful starting point, they do not take account of the unique nature of AI. The use of statistical effects and principles in AI calls for additional caution and additional processes to ensure that products are safe and effective. Figure 2 shows an overview of the Zühlke medical AI process and highlights the stages at which each of the above regulations apply.

Zühlke Medical AI Process



Figure 2: The stages of the medical AI development process with corresponding regulations and documentation

Medical AI development best practices and guidelines

In addition to general regulatory requirements, a number of factors highly specific to AI-based medical software also need to be considered. At the time of writing, the number of best practice guides and guidelines on AI for medical software is limited but, encouragingly, growing steadily. While none of these guidelines have an official character, all aim to anticipate the approach regulators will take. This makes them very useful.

Paper/guideline/framework	Scope	Content/focus
Johner Institute Guideline for AI ⁵ ↗	EU and FDA	Guideline listing state of the art and best practices, as well as describing management of risks to maximise patient safety.
Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) ⁶ ↗	FDA	Total product lifecycle-based regulatory framework for AI technologies. The framework allows modifications to be made from real-world learning and adaptation, while ensuring that software as a medical device remains safe and effective.
Review Key Points of Aided Decision-Making Medical Device Software Using Deep Learning Techniques (NMPA, 2018)	China	Guidelines for decision-making medical device software with a focus on deep learning. Issued by the Chinese National Medical Products Administration (NMPA).
Regulatory Science on AI-based Medical Devices and Systems (Chinzei et al., 2018)	Japan	Technical report on the clinical positioning of AI medical systems and their applications from a regulatory science viewpoint. Issued by the Japanese Science Board of the Pharmaceuticals and Medical Devices Agency (PMDA).
BSI/AAMI: Machine Learning AI in Medical Devices ⁷ ↗	EU and FDA (AAMI)	This paper examines the implications of the differences between conventional medical software and AI, and discusses the controls needed to ensure AI in healthcare is safe and effective.

AI is often developed for use as a standalone medical device, rather than forming part of a physical medical device. This adds another layer of complexity, which it shares with non-AI [Software as Medical Device](#) (SaMD) solutions. Fortunately, for SaMD solutions, existing [IMDRF](#)⁸ ↗ and [MDCG](#)⁹ ↗ guidelines represent an established and valuable source of information. For AI solutions, these guidelines provide essential information on issues such as risk categorization, risk-benefit analysis and clinical evaluation of software.

FDA framework

As one of a handful of organizations responsible for the admission of medical devices, the FDA has embraced the concept of AI and has published a set of documents detailing how the agency plans to handle AI solutions in the future. Additionally, the FDA has also set up a pre-certification program for companies planning to submit medical AI solutions for admittance¹⁰ ↗ and launched a Digital Health Centre of Excellence in September 2020¹¹ ↗. It's therefore reasonable to assume that, in the short term, the FDA will play a leading role in defining AI-specific needs and requirements. As a result, it's worth paying close attention to new FDA guidelines and recommendations.

Based on frameworks published by the FDA and discussions with notified bodies, in order to receive market approval for AI products, additional AI-specific requirements must be met:

- Quality management systems will need to include processes for AI. Where available, [good machine learning practice \(GMLP\)](#) ↗ will need to be followed.
- Transparency will be key. Algorithm outputs will be expected to be reasonably transparent.
- There will need to be proper separation of training, [tuning](#) ↗ and test datasets.
- The expected changes to the algorithm and planned procedures to change and/or update the algorithm need to be described.
- Post-market surveillance plans will be expected to consider AI-specific issues (including real-world evidence for AI models, root cause analysis for mispredictions, analysis of data shift in distributions, etc.). Manufacturers will need to be transparent about performance and about changes made to the product.

The FDA framework also discusses options for algorithms which continue to learn in the field. This is a fundamentally new concept for medical software. Such an approach would previously have been unthinkable – medical devices must be validated before they can be marketed and there are clear rules and requirements for updating software in the field. If we are to realise AI's full potential, however, continuous learning will eventually have to come into play. When it does, particular attention will need to be paid to how algorithm modification is defined.

In its framework, the FDA tries to explain under what circumstances:

- it will expect manufacturers to document a modification only, without requiring a new premarket submission
- it will wish to review modifications and perform validation before the manufacturer can market the modified product
- it will insist on a (completely) new submission or approval

Other regulatory authorities are likely to follow the FDA's lead and establish similar rules clearly setting out what sort of algorithm modifications can be made without requiring notified body involvement. While this point is definitely something to bear in mind and investment in developing such systems may prove worthwhile, at present we nonetheless continue to recommend keeping models locked and dealing with updates in the conventional way.

The need for more guidelines

As outlined before ↗, a number of institutions and interested parties are currently putting forward proposals for guidelines and best practices for medical AI. There are, however, few or no harmonized standards. There is significant demand for more detailed regulations, guidelines and standards, and calls for such guidelines appear to have been heard. A WHO and ITU initiative has been launched with the aim of establishing a rigorous standardized evaluation framework for AI in healthcare (Wiegand et al., 2019) and some independent institutes have also started to publish guidelines for AI in medical products (Johner et al., 2020).

In the meantime, as ever the best approach is to talk to notified bodies and competent authorities at an early stage.

Despite the lack of official documents, notified bodies in both the US and Europe have reached out to AI experts and have recently formulated their idea about what they will expect from manufacturers to start the approval process for AI products (see the guidelines and frameworks on page 18 ↗). With AI, getting official agencies involved at an early stage is particularly important. That's because a lot of AI-specific requirements relate to collection, labelling and processing of the data used for developing and training algorithms, which takes place early in the project. Getting a notified body involved or considering documentation requirements at too late a stage could end up being a costly mistake.

Principles of AI in Health Tech

When it comes to turning a machine learning research project into a marketable product, safety and quality need to be top priority throughout. Output from machine learning models needs to be systematically verified and validated, and all software for training and deploying models must be developed using a quality management system. The development should be patient-centred to ensure that the technology is clinically useful, effective, and tested for unintended consequences.

We have brought together machine learning expertise from more than 100 AI projects and expertise from over 50 years of medical device development to produce internal guidelines and processes for building health care products which incorporate machine learning. Figure 2 [↗](#) shows our medical machine learning process. It outlines key steps and considerations at each stage of the product development process. It brings together our data science process and requirements from our medical device development process.

Below, we discuss 8 key principles for developing medical AI systems. These principles complement existing medical device development guidelines and legislation. (See also Figure 4 [↗](#).)

Principle 1

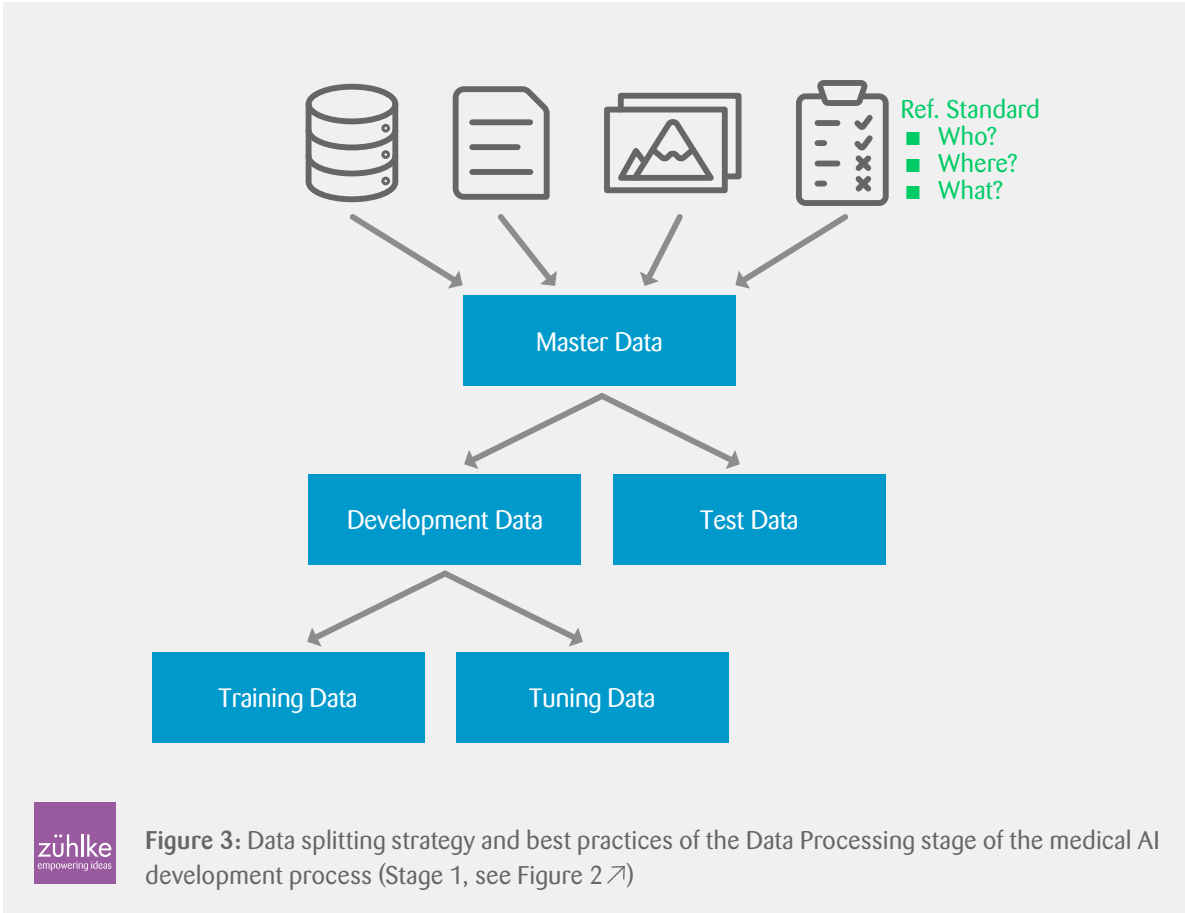
High data and labelling quality

Supervised machine learning – the de-facto standard AI methodology – is heavily dependent on large datasets with high-quality labels. Detecting skin cancer in images, for example, requires the availability of large numbers of images of both cancerous and non-cancerous tissue. Similarly, pre-clinical validation of machine learning models always involves comparing model outputs to known reference labels. The quality of datasets, especially reference labels used for training and evaluating machine learning models, is therefore critical.

So it's not surprising that regulators have high expectations for how data, labelling and reference standard (the ground truth data used for training and evaluation) are documented (US Food and Drug Administration, 2012). Key questions documentation needs to answer are:

- Does the distribution of data match the population defined in the intended use?
Any machine learning model will only be valid for the data it was trained and evaluated on. It is therefore important that the patient population and method used to obtain the data are similar to the patient populations and methods to which it will be applied. This includes, for example, demographic data, medical history, disease state, and test methods and equipment.
- How were the reference labels obtained?
If this involved human annotators, what qualifications and expertise did the experts possess and what instructions did they receive?
- Was data removed and if yes, according to what criteria?
- What confounders and effect modifiers are there?
What is the distribution of the data, conditioned on confounders or effect modifiers?
These could be variables such as the type of measuring equipment, the clinic in which the data was recorded or patient medical history.

Stage 1 – Data Processing



Best Practices

- Validated code for data processing and splitting
- Controlled access to master data
- Analysis of correlation, confounders, and groups
- Necessary size of test data estimated (power analysis)
- Data representative of intended use population
- Documented process for obtaining reference standard

Principle 2

Good machine learning practices

There is no official good machine learning practice (GMLP) standard like there is for other fields (GxP). Nonetheless, there is some consensus among machine learning researchers about what constitute essential practices for ensuring high quality machine learning models.

An important task for a company aiming to build medical AI solutions is to have a senior data scientist write down these practices in the form of guidelines, and to ensure

that every data scientist or machine learning engineer working on building the solution adheres to them.

GMLP should describe things like how to split data into training, tuning and testing data, how to select a model, perform hyperparameter optimization and features selection, best practices for data collection, pre-processing and outlier removal, and best practices for model validation and explainability.

Principle 3

Simple, transparent, interpretable models

While the main goal of training a machine learning model is maximizing its performance, it is important to remember that the only performance that matters is performance on new, unseen data. It is not enough to look at just a single performance metric computed on a test data set. Simpler models may not fit the training data as well as more complex ones, but the improvement in performance often fails to translate into actual performance in the field. So if multiple models achieve similar performance on test data, choose the simplest.

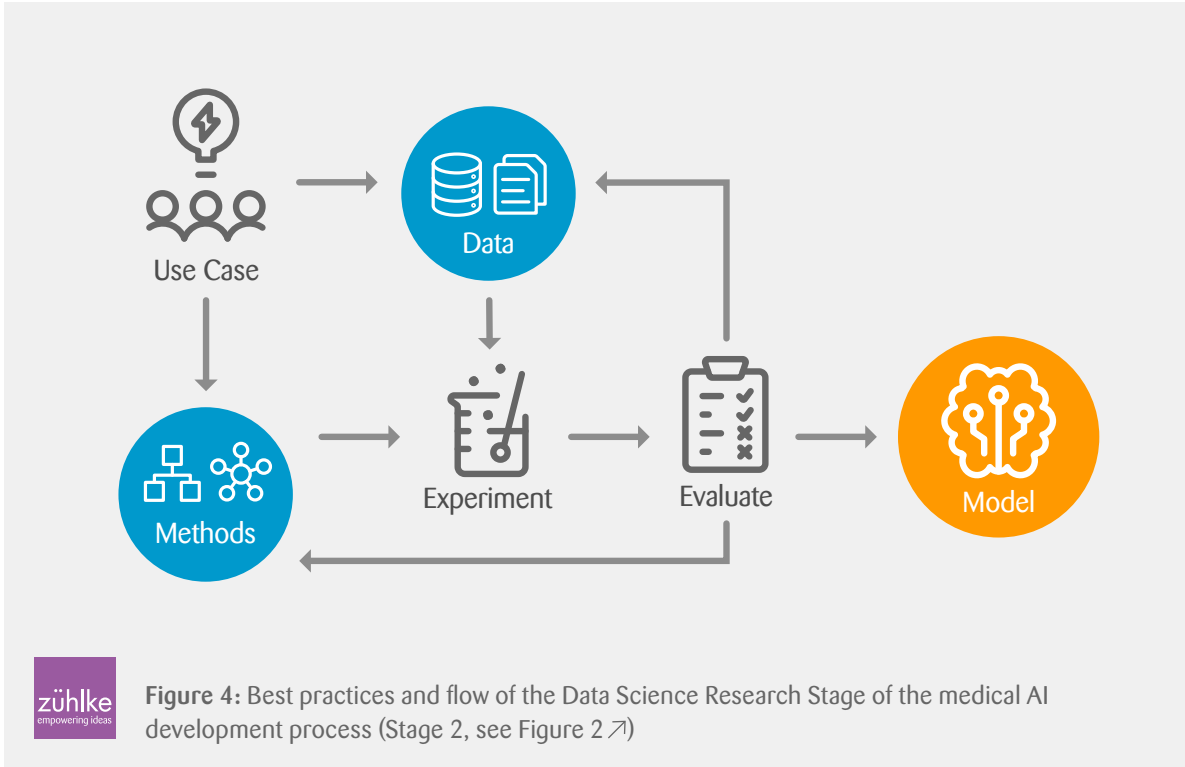
The robustness of the model in the face of slight perturbations in the data also needs to be investigated. Rather than computing point estimates of model performance, multiple evaluations under different experimental conditions (random seeds, sub-samples of the data) should be performed. This permits an evaluation of performance differences between different models.

Machine learning models should not be black boxes. The output from a model needs to be explainable both to

medical staff and to patients, who have a right to know on what symptoms and patient information a specific diagnosis is based. The data scientists who develop and test a model should also be able to understand why it comes up with a specific output. This information can be used to debug the model and check that its predictions make sense. Some models are interpretable by design, for example decision trees and linear regression. For more complex models like gradient boosting and neural networks, a number of approaches to interpreting outputs are available. The book “Interpretable Machine Learning” (Molnar, 2020) gives a good overview on the topic.

Choosing simple, transparent, and interpretable models is therefore at the core of managing the patient risks that come with applying machine learning techniques in healthcare and increases the success of regulatory admission.

Stage 2 – Data Science Research



Best Practices

- Documentation of all experiments and results
- Version Control System for code, models and data
- Re-sampling and cross-validation to compare models
- Clinically motivated metric for evaluation
- Computer System Validation
- Algorithm Research & Development Report
- Favor simpler, more robust models with better explainability, less and simpler features

Principle 4

Thorough evaluation

One of the most important aspects of building a machine learning model is understanding how well it will work outside the lab. One of the big challenges for medical AI is transferring research results into clinical applications. If performance is not properly evaluated, there is a significant risk that machine learning performance will be overestimated.

Identifying trustworthy performance indicators early in the product development pipeline allows costly clinical trials for underperforming models to be avoided and investment to be directed towards more promising models. As discussed above, this can only be achieved by subjecting models to extensive and rigorous testing. Otherwise resourcing decisions will be based on untrustworthy data and are likely to prove to be misguided.

It is important to remember that the key metric for evaluating machine learning models is not their performance on the data used for developing the model, but their performance on new, previously unseen data. This is obviously not computable, and proper evaluation should strive to obtain the best possible estimate of future performance.

There have been a number of initiatives aimed at standardizing reporting of model performance:

- The TRIPOD Statement (Moons et al., 2015) features a 22-item checklist for use when reporting on the development and performance of predictive models for individual prognosis or diagnosis.
- A recent initiative has adapted the TRIPOD guidelines specifically for machine learning models. (Collins and Moons, 2020).
- The final step for obtaining market approval is usually a clinical study. Existing guidelines on clinical trial protocols and reporting have recently been adapted for medical interventions involving AI, giving rise to the SPIRIT-AI (Cruz Rivera et al., 2020) and CONSORT-AI extensions (Liu et al., 2020).

Estimating future performance of a predictive model – model validation – is generally a hard problem. To obtain a good estimate, a number of factors need to be taken into account. The process starts with an idea of the distribution of the data to which the model will be applied. As with intended use for medical devices, the manufacturer needs to define the intended patient group. Once the intended use population has been identified, you need to collect data with the same distribution of features (age, sex, disease, etc.) and target variable. This is used to evaluate and train the machine learning model.

When estimating performance, some of the main challenges relate to the process of correctly splitting data into training and testing sets. It is important to ensure that the split retains stratification, but does not distribute correlated groups of data points across different data sets. If, for example, a model is to be applied to patients from clinics that are not represented in the training data, it is essential that data from any one clinic is not split between the training and testing data sets.

Stage 3 – Model Performance Evaluation

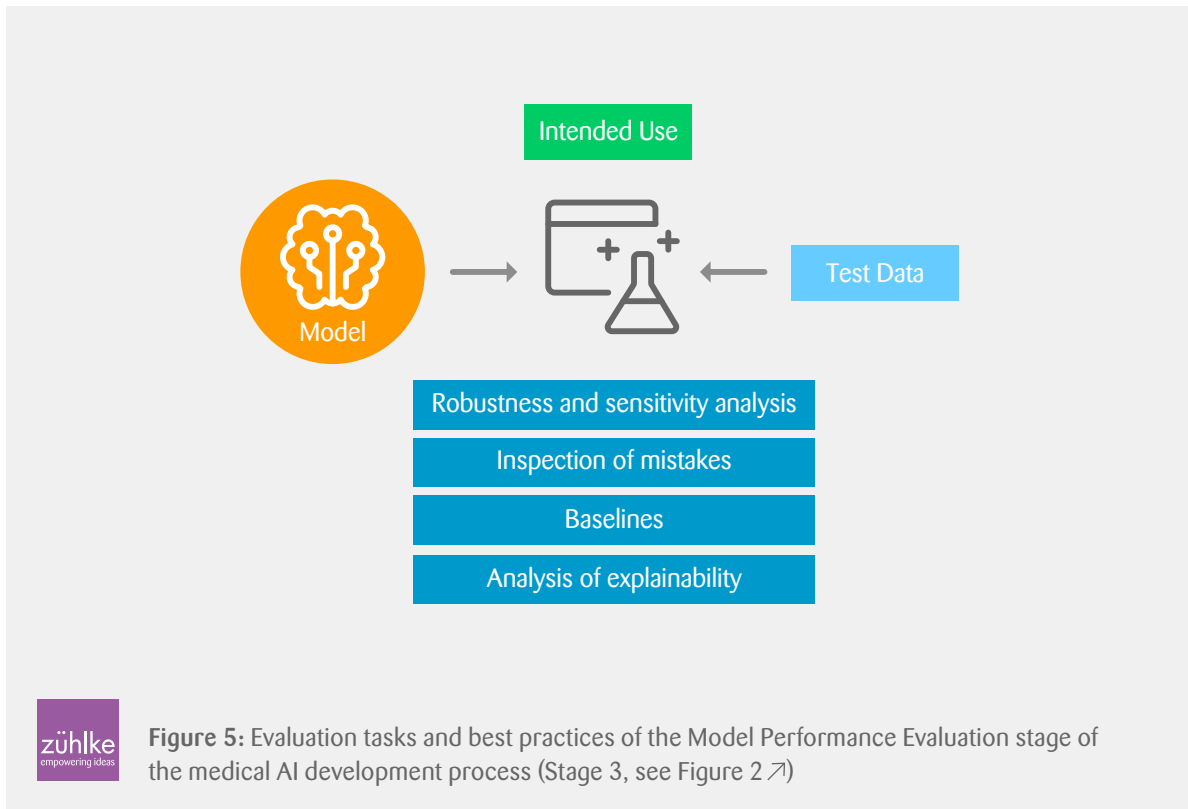


Figure 5: Evaluation tasks and best practices of the Model Performance Evaluation stage of the medical AI development process (Stage 3, see Figure 2 ↗)

Best Practices

- Point of no return: Validation data can not be used repeatedly
- Validated code for evaluation and metric computation
- Algorithm Research & Development Report
- Evaluate metric, data and performance requirements against intended use
- Comparison to simple baseline, human performance and state of the art
- Analysis of explainability
- Performance on out-of-distribution data

Principle 5

Reproducible machine learning pipelines

It is vital that data science experiments generate consistent results on repeated runs. To ensure reproducibility, it is necessary to automate all aspects of training and testing of machine learning models. That means that all

steps in the machine learning pipeline – data pre-processing, feature selection, hyperparameter optimization, cross-validation, etc. – need to be executable as software and not be dependent on manual processes.

Principle 6

Manage data, model and code history

In a conventional software engineering project, source code is generally subject to version control. In a machine learning project it is also necessary to track the history of and changes to data and trained models. This triplet of data, models and code forms the core of a machine learning application. They are strongly interlinked.

It is important that data, model and code history are managed in tandem, as any modifications to the data or

code also modify the trained model. It therefore follows that when a model is evaluated or used in a finished product, the code and data need to be in the same state they were in when that version of the model was trained.

Having a managed data, model and code history also means you can run large numbers of different experiments, thereby accelerating the data science process.

Stage 4 – Medical Device Development

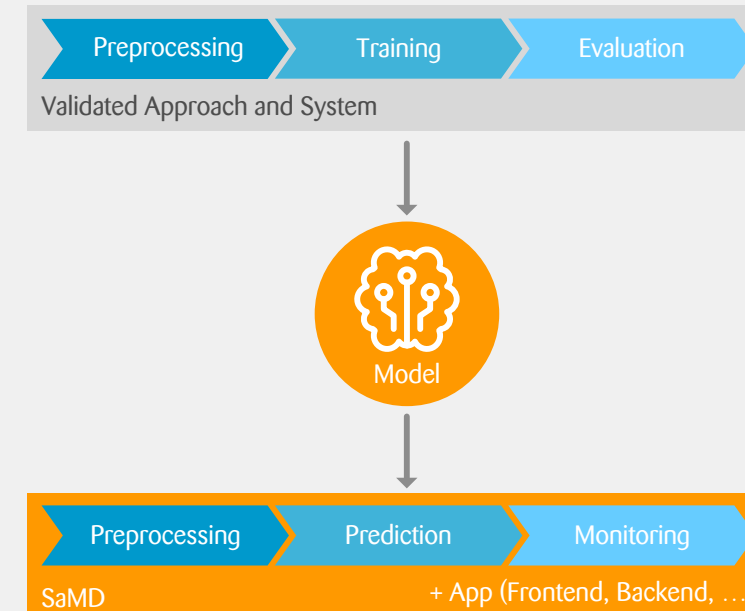


Figure 6: Best practices and flow of the Medical Device Development Stage of the medical AI development process (Stage 4, see Figure 2 ↗)

Quality Control depending on Classification

- Regulated Development (MDR/IEC 82304/IEC 62304)
- Model Specification
- Risk Management (ISO 14971)
- Verification & Validation (MDR/IEC 82304/IEC 62304)
- Quality Management (ISO 13485)
- Computer System and Process Validation for Training
- SOUP analysis for SaMD
- Usability (IEC 62366)

Principle 7

Comprehensively document code, data, models and design decisions

The development process during the research phase of a medical machine learning project does not need to be IEC 62304 compliant, but it is nonetheless vital that every decision and experiment is documented.

This documentation should cover the following points:

- ground truth labels
- data processing
- data distribution and intended use population – see “Datasheets for datasets” (Gebru et al., 2020) and “Dataset Nutrition Label” (Holland et al., 2018) for examples
- model validation
- software bug risk mitigation
 - code reviews
 - unit tests
 - code documentation
- sampling and data splitting methods
- model selection
- hyperparameter optimization
- feature selection
- explainability and feature importance

Principle 8

Communicate results and uncertainty carefully

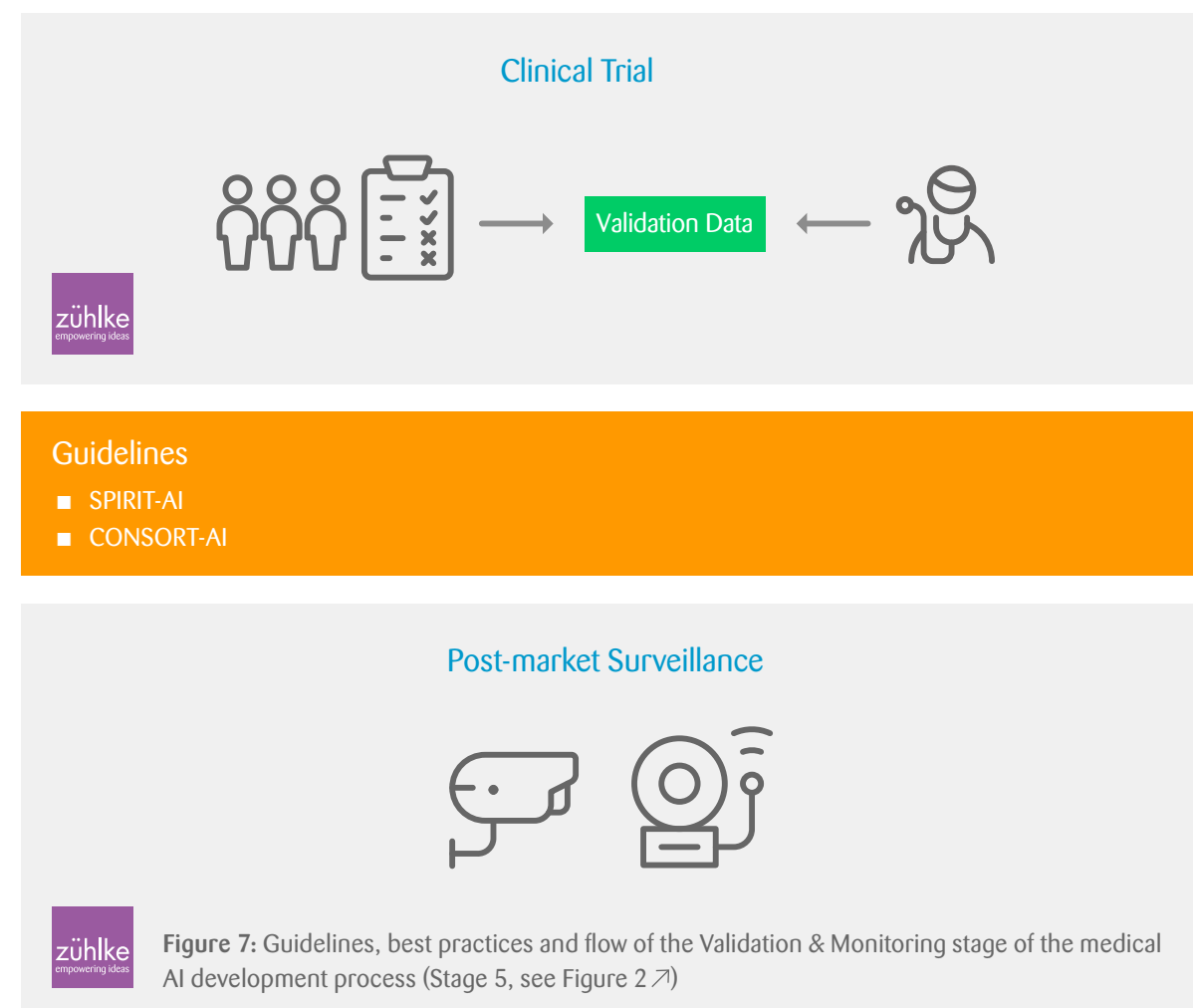
Machine learning models can be used by doctors and other clinical staff in a variety of ways. A diagnostic model, for example, could simply output the most likely disease, it could state the estimated likelihood of various diseases, it could show a list of the 5 most likely diseases, it could just show predicted risk factors that can lead to one of the diseases and so on.

How the model is going to be used and in what context needs to be clearly specified in the product’s intended use. The output from the model needs to be tailored for this specific use. If the model is to present clinicians with probabilities, they need to be properly calibrated. In the case of a disease vs. healthy classification model, this means that x% of patients where a model outputs a disease probability of x% indeed have the disease. Many commonly used machine learning models by default output uncalibrated probabilities. Even after calibrating a

model using training data, it can still output incorrect predictions with high certainty when used on data which does not have the same distribution as the training data (Ovadia et al., 2019; Zech et al., 2018).

To mitigate these issues it is necessary either to use models that can detect data is outside of the training data distribution (Lee et al., 2018b, 2018a) or to add additional algorithms that can identify and reject such data before it is given to the model. It is essential that the capabilities and limitations of a model are clearly communicated to the clinician. One option for achieving this is to use the “Model facts” label. The “Model facts” label is an attempt to standardize how the capabilities of and risks associated with machine learning models are communicated to clinicians (Sendak et al., 2020).

Stage 5 – Validation & Monitoring



Guidelines

- SPIRIT-AI
- CONSORT-AI

Best practices

- Generate real world clinical evidence for AI throughout its life cycle
- Continuously monitor the performance of the AI system
- Continuously monitor the data for a shift in distribution

Figure 7: Guidelines, best practices and flow of the Validation & Monitoring stage of the medical AI development process (Stage 5, see Figure 2 ↗)

Conclusion and Outlook

Fuelled by technological advances in computing and growing awareness of the value of data, AI has the potential to fundamentally change Health Tech for the better. While not (yet) suitable for use in all areas, AI has already proven its value and is now trusted for a variety of applications. We are confident that AI will continue its ascent and investment in AI will continue to grow. However, a lack of regulatory clarity and a lack of certainty regarding future changes to requirements and best practices are making the approval process for AI more costly and hindering more widespread adoption. For AI to build on its existing successes, regulatory authorities need to release guidelines on and set boundaries for using AI in regulated products without delay.

AI is already in use in applications such as clinical decision support systems and image analysis, and will almost certainly start to make inroads into language processing, disease subtyping and many other health care fields. In such fast-moving fields, waiting for a clearer regulatory environment before setting out on your medical AI journey risks missing the boat.

Until more specific guidelines are available, we recommend employing good machine learning practices and adapting existing medical device regulations. To achieve this, it is important to get machine learning and quality assurance experts involved at an early stage. Key principles when developing medical AI solutions include thoroughly evaluating model performance and the suitability of data with respect to the intended use, building interpretable models and reproducible machine learning pipelines, and comprehensively documenting all design decisions, data, models and code.

Looking beyond this paper's focus on commercial, user facing medical AI products, AI also has tremendous potential in the pharmaceutical industry. Here it is set to accelerate drug and vaccine development – a competitive advantage that is more important than ever in the COVID-19 era. In many pharmaceutical fields, any technology which is able to reduce the set of drug candidates, thereby reducing the length and duration of clinical studies, is likely to be a game changer, and AI is no exception¹² [↗](#). In combination with human-on-a-chip systems, for example, AI could – once proven – drastically reduce animal testing, and thereby significantly reduce drug development costs. Although regulatory requirements may differ for R&D-oriented applications, we are confident that that our insights on best practices as presented in this whitepaper may be easily adapted other regulated AI fields.

With a lot of attention on AI and high hopes for what it may achieve, it can be difficult to assess its real potential for a specific application. It's important to remember that AI is not a panacea. Nevertheless, it will be important to keep the current momentum of AI in order to not fall behind this highly competitive field. Developing stable, safe and secure AI algorithms takes time. It requires specialist knowledge and skills which are in great demand on the labour market and therefore may be difficult to obtain. While it certainly is a risk to invest in building up the resources needed to create AI concepts, develop use cases and to realize them, in our opinion it is a risk well worth taking.

Glossary

Artificial Intelligence (AI)	Academic discipline concerned with the study of intelligent agents, independent of how they act intelligently. As used in this paper, AI includes specific solutions such as machine learning.
Feature	An input variable to the model, for example information about a patient or a test result.
Federated learning	Decentralized machine learning technique for training a model on multiple servers without sharing data.
GMLP	Good machine learning practice
GxP	Good “x” practice. Framework of guidelines including good manufacturing practice, good clinical practice, good distribution practice, etc.
Health Care Professionals (HCP)	People professionally involved in the treatment, diagnosis, and prevention of diseases such as doctors, nurses, care personal etc.
Health Tech	The entire health care ecosystem involved in treating, diagnosing and preventing disease, and improving patient quality of life. It encompasses both devices and drugs.
Label	Annotation of the output in training and testing data. Labels are usually classes (e.g. different diseases) or actual values.
Intended use	The Intended Use defines and delimits the purpose and the function of medical products or services. The intended use is defined by the legal manufacturer of the product or service.
Machine learning (ML)	Field of computer science that gives computers the ability to learn from data without being explicitly programmed. A sub-field of AI.
Medical device	Any device (including hardware or software only) which fulfils the definition of a medical device, is subject to medical device regulations and needs marketing approval.
Medical software	Any software running on a physical medical device or any standalone software that has a medical purpose as defined in medical product regulations.
Model	A mathematical model that describes the relationship between input and output (feature → label) for a dataset. Models are trained on data using machine learning algorithms.
Natural language processing	A sub-field of machine learning that encompasses models, techniques, algorithms, services and tools for working with natural language.
Protected health information (PHI)	Any information about health status, provision of health care, or payment for health care that can be linked to a specific individual.
Regulated	Development under strict compliance with harmonized standards and subject to approval by competent authorities and/or notified bodies before market access is granted.
SaMD	Software as medical device – a medical device consisting purely of a software solution and which does not include any hardware (other than accessories).
Testing data	Data used for evaluating the output of a model.
Training data	Data used for training a model.
Tuning data	Data used for selecting a model type and the hyperparameters of a model.

Authors



Dr. Gabriel Krummenacher
Lead Data Scientist

Gabriel Krummenacher leads the Data Science Team at Zühlke and has ten years of experience in conducting data analytics and machine learning projects. He holds a PhD from the Institute for Machine Learning at ETH, where he worked on scalable methods for large-scale and robust learning, wheel defect detec-

tion and sleep stage prediction with deep learning. He took an MSc in computer science at ETH Zurich in 2011. He was a visiting researcher at the NICTA Bioinformatics group in Melbourne and interned in the Trading Technology Team at Axa Rosenberg in San Francisco.



Dr. Stefan Weiss, MBA
Lead Business Innovation Consultant Pharma & MedTech

Dr. Stefan Weiss is Lead Business Innovation Consultant at Zühlke. He has a broad background in Neuroscience (PhD, Helmholtz Center Munich) combined with extensive economics and (digital) innovation management expertise (MBA, Collège des Ingénieurs). Before joining Zühlke, Stefan helped shape the future of Healthcare and Life Sciences at the Merck KGaA Innovation Center. He is passionate

about using his scientific and economic expertise to realise digitalization of the pharmaceutical and MedTech sectors using innovative, AI-driven solutions and business models. At Zühlke, he extends technical excellence with domain-specific insights, helping to strengthen partnerships with clients in the pharmaceutical and MedTech sectors.



Dr. Patrick Steiner
Project Manager Health Tech

Patrick Steiner joined Zühlke in 2016. He has worked on a range of SaMD projects over the last few years, including conventional software and AI solutions in regulated environments. His focus is on pragmatic,

lean implementation of compliant development processes for regulated medical products. He has an MSc ETH in Biomedical Engineering and a PhD in Biomedical Imaging.



Uwe Szymanski
Lead Regulatory Expert and Software Architect

Uwe Szymanski joined Zühlke in October 2010 and is Embedded Software Lead. His main areas of expertise include systems & software engineering, project management, and software development in reg-

ulated markets. He has primarily worked in the healthcare, life science and mechanical engineering. His passion is the development of safety critical systems.

Impressum:

Zühlke Engineering AG
Zürcherstrasse 39J
8952 Schlieren (Zürich)
Switzerland

info@zuehlke.com
Managing Director:
Nicolas Durville

Bilder:
Getty Images Deutschland GmbH
© Zühlke 2021 all rights reserved

References

1 <https://cdn.medicalfuturist.com/wp-content/uploads/2019/06/The-Medical-Futurist-FDA-approved-AI-algorithms-in-medicine-2019-09.png> 

2 <https://www.melloddy.eu/> 

3 <https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf> 

4 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 

5 <https://www.johner-institut.de/ai-guideline/> 

6 <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> 

7 https://pages.bsigroup.com/l/35972/2020-05-06/2dkr8q4?_ga=2.46422688.647334402.1603358242-680835268.1588162584 

8 <http://www.imdrf.org/documents/documents.asp> 

9 https://ec.europa.eu/health/md_sector/new_regulations/guidance_en 

10 <https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program> 

11 <https://www.fda.gov/news-events/press-announcements/fda-launches-digital-health-center-excellence> 

12 <https://www.nature.com/articles/s42256-020-00236-4.epdf> 

Bell, D., Gachuhi, N., Assefi, N., 2018. Dynamic Clinical Algorithms: Digital Technology Can Transform Health Care Decision-Making. *Am J Trop Med Hyg* 98, 9 – 14. <https://doi.org/10.4269/ajtmh.17-0477>

Chen, P.-H.C., Liu, Y., Peng, L., 2019. How to develop machine learning models for healthcare. *Nat Mater* 18, 410 – 414. <https://doi.org/10.1038/s41563-019-0345-0>

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L.K., Segler, M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., Greene, C.S., 2018. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 15. <https://doi.org/10.1098/rsif.2017.0387>

Chinzei, K., Shimizu, A., Mori, K., Harada, K., Takeda, H., Hashizume, M., Ishizuka, M., Kato, N., Kawamori, R., Kyo, S., Nagata, K., Yamane, T., Sakuma, I., Ohe, K., Mitsuishi, M., 2018. Regulatory Science on AI-based Medical Devices and Systems. *ABE* 7, 118 – 123. <https://doi.org/10.14326/abe.7.118>

Collins, G.S., Moons, K.G.M., 2019. Reporting of artificial intelligence prediction models. *The Lancet* 393, 1577–1579. [https://doi.org/10.1016/S0140-6736\(19\)30037-6](https://doi.org/10.1016/S0140-6736(19)30037-6)

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A.K., Calvert, M.J., 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine* 26, 1351 – 1363. <https://doi.org/10.1038/s41591-020-1037-7>

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. *Nature Medicine* 25, 24 – 29. <https://doi.org/10.1038/s41591-018-0316-z>

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K., 2020. Datasheets for Datasets. *arXiv:1803.09010* [cs].

Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K., 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv:1805.03677* [cs].

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Yilong, Dong, Q., Shen, H., Wang, Yongjun, 2017. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2. <https://doi.org/10.1136/svn-2017-000101>

Johner, P.D.C., Christoph Molnar, Andreas Purde, Abtin Rad, Christian Dierks, Stefan Bunk, Sven Piechottka, 2020. Guideline for AI for medical products.

Lee, Kimin, Lee, H., Lee, Kibok, Shin, J., 2018a. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. *arXiv:1711.09325* [cs, stat].

Lee, Kimin, Lee, Kibok, Lee, H., Shin, J., 2018b. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), *Advances in Neural Information Processing Systems* 31. Curran Associates, Inc., pp. 7167 – 7177.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis* 42, 60 – 88. <https://doi.org/10.1016/j.media.2017.07.005>

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M.J., Denniston, A.K., 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine* 26, 1364 – 1374. <https://doi.org/10.1038/s41591-020-1034-x>

Molnar, C., 2020. Interpretable Machine Learning. Lulu.com.

Moons, K.G.M., Altman, D.G., Reitsma, J.B., Ioannidis, J.P.A., Macaskill, P., Steyerberg, E.W., Vickers, A.J., Ransohoff, D.F., Collins, G.S., 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 162, W1. <https://doi.org/10.7326/M14-0698>

Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H., Topol, E.J., Ioannidis, J.P.A., Collins, G.S., Maruthappu, M., 2020. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368. <https://doi.org/10.1136/bmj.m689>

NMPA, 2018. Review Key Points of Aided Decision-Making Medical Device Software Using Deep Learning Techniques.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J., 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv:1906.02530* [cs, stat].

Saria, S., Goldenberg, A., 2015. Subtyping: What It is and Its Role in Precision Medicine. *IEEE Intelligent Systems* 30, 70 – 75. <https://doi.org/10.1109/MIS.2015.60>

Sendak, M.P., Gao, M., Brajer, N., Balu, S., 2020. Presenting machine learning model information to clinical end users with model facts labels. *npj Digital Medicine* 3, 1 – 4. <https://doi.org/10.1038/s41746-020-0253-3>

Sinkala, M., Mulder, N., Martin, D., 2020. Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics. *Scientific Reports* 10, 1212. <https://doi.org/10.1038/s41598-020-58290-2>

US Food and Drug Administration, 2019. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)-discussion paper.

US Food and Drug Administration, 2012. Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data – Premarket Notification [510(k)] Submissions.

Wiegand, T., Krishnamurthy, R., Kuglitsch, M., Lee, N., Pujari, S., Salathé, M., Wenzel, M., Xu, S., 2019. WHO and ITU establish benchmarking process for artificial intelligence in health. *The Lancet* 394, 9 – 11. [https://doi.org/10.1016/S0140-6736\(19\)30762-7](https://doi.org/10.1016/S0140-6736(19)30762-7)

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine* 15, e1002683. <https://doi.org/10.1371/journal.pmed.1002683>

Zhang, X., Chou, J., Liang, J., Xiao, C., Zhao, Y., Sarva, H., Henchcliffe, C., Wang, F., 2019. Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. *Scientific Reports* 9, 797. <https://doi.org/10.1038/s41598-018-37545-z>