Classification of AI technologies

Phase 1

Authors alphabetically:

Anastasia Chalkidou

Jamie Erskine

Kate Goddard

Farhad Shokraneh

Contents

[1 Abbreviations 4](#_Toc59055548)

[2 Executive summary 6](#_Toc59055549)

[3 Background 6](#_Toc59055550)

[4 Aim 7](#_Toc59055551)

[5 Methods 7](#_Toc59055552)

[6 AI classification 8](#_Toc59055553)

[6.1 Literature findings 8](#_Toc59055554)

[6.1.1 Theoretical Classifications 8](#_Toc59055555)

[6.1.2 Practical Classifications 9](#_Toc59055556)

[6.2 Expert elicitation 13](#_Toc59055557)

[6.2.1 Definition of AI 13](#_Toc59055558)

[6.2.2 Pre-existing AI classifications 13](#_Toc59055559)

[6.2.3 Thoughts on AHSN and CQC suggestions for classifying AI 14](#_Toc59055560)

[6.2.4 Examples of useful categorisations 14](#_Toc59055561)

[7 Existing regulatory frameworks 14](#_Toc59055562)

[7.1 Literature findings 14](#_Toc59055563)

[8 Reporting guidelines 15](#_Toc59055564)

[8.1 CONSORT & SPIRIT AI extensions 15](#_Toc59055565)

[8.2 Model types 16](#_Toc59055566)

[8.3 Conference Proceedings 17](#_Toc59055567)

[9 AI evaluation – key considerations 18](#_Toc59055568)

[9.1 Literature findings 18](#_Toc59055569)

[9.1.1 Systems vs. product-based approach 18](#_Toc59055570)

[9.1.2 United States Government Accountability Office 19](#_Toc59055571)

[9.1.3 Park and Han (2018) 19](#_Toc59055572)

[9.1.4 Magrabi et al. (2019) 21](#_Toc59055573)

[9.1.5 Safety Analysis 21](#_Toc59055574)

[9.1.6 Olthof et al. (2020) 21](#_Toc59055575)

[9.2 Expert elicitation 22](#_Toc59055576)

[9.2.1 Overlap with other evaluation processes 22](#_Toc59055577)

[9.2.2 Key consideration in evaluating ML 22](#_Toc59055578)

[9.2.3 Approach to evidence generation 23](#_Toc59055579)

[9.2.4 Future proofing – changes expected in the next 5 years 23](#_Toc59055580)

[10 Phase 4 AI award technologies 24](#_Toc59055581)

[11 Discussion 24](#_Toc59055582)

[12 Conclusions 26](#_Toc59055583)

[13 References 28](#_Toc59055584)

[14 Appendices 31](#_Toc59055585)

[14.1 Reporting Guidelines 31](#_Toc59055586)

[14.2 Computer Science Conference Topics 40](#_Toc59055587)

[14.3 HTAs of AI 41](#_Toc59055588)

[14.4 Regulatory frameworks 47](#_Toc59055589)

[14.5 Technographic review 73](#_Toc59055590)

[14.6 Classification example 76](#_Toc59055591)

# Abbreviations

|  |  |
| --- | --- |
| ACP | Algorithm Change Protocol |
| ACM CCS | Association for Computing Machinery Computing Classification System |
| AI | Artificial Intelligence |
| AIMD | Artificial Intelligence Medical Devices |
| AUC | Area Under the Curve |
| CADe | Computer-Assisted Detection Devices |
| CDS | Clinical Decision Support |
| CLA | Continuously Learning Systems |
| COCIR | European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry |
| CONSORT | Consolidated Standards of Reporting Trials |
| DHT | Digital Health Technology |
| ECBIOS | Eurasia Conference on Biomedical Engineering,. Healthcare and Sustainability |
| EMA | European Medicines Agency |
| FDA | Food and Drug Administration (United States) |
| FMEA | Failure Mode and Effects Analysis |
| GMLP | Good ML Practices |
| HCPs | Health Care Professionals |
| HMA | Heads of Medicines Agencies |
| HTA | Health technology assessment |
| IEEE | Institute of Electrical and Electronics Engineers |
| IMDRF | International Medical Device Regulators Forum |
| IVDR | In-Vitro Diagnostic Devices Regulation |
| MDED | Medical Device Evaluation Division (Japan) |
| MDR | Medical Device Regulation |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| NMPA | National Medical Products Administration (China) |
| PMA | Premarket Approval Application |
| PMDA | Pharmaceuticals and Medical Devices Agency (Japan) |
| PSEHB | The Pharmaceutical Safety and Environmental. Health Bureau (Japan) |
| RCT | Randomised Controlled Trial |
| ROC | Receiver Operator Curve |
| SaMD | Software as Medical Device |
| SPIRIT | Standard Protocol Items: Recommendations for Interventional Trials |
| SPS | SaMD Pre-Specifications |
| TGA | Therapeutic Goods Administration (Australia) |
| TPLC | Total Product Life Cycle |
| TRIPOD | Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis |
| WIPO | World Intellectual Property Organization |

# Executive summary

A framework to evaluate the clinical and cost-effectiveness of Artificial Intelligence (AI) in healthcare does not exist. Understanding the classification ‘building blocks’ of AI technologies that may impact on how the evidence is developed to demonstrate their effectiveness and safety is an important step to develop this framework. This work describes the systematic search and review of AI literature and AI-related reporting guidelines to understand how AI technologies can potentially be classified at the lowest level of granularity to enable an accurate initial triage for health technology assessment (HTA). The output is a preliminary group of categories based on the literature review and expert elicitation from the London Medical Imaging and AI Centre for Value-Based Healthcare.

This preliminary report, concluded that a low granularity classification, seems to be the most pragmatic and applicable approach, especially since currently ML applications are dominating the healthcare field. A classification that is valid on a theoretical level, such as the AHSN complexity scale, may be remote from the practical applications of AI in healthcare, and therefore, operationally irrelevant if the majority of existing applications are skewed towards one of the suggested categories in each class.

From a regulatory perspective, triaging risk seems to be a recurring theme. Autonomous systems and systems that are continuously evolving ML are suggested as higher risk. Harder to evaluate as autonomous systems may be more “black box” – potentially requiring a more pragmatic approach based on outcomes. This will shift the evaluation focus on carefully selecting those outcomes and providing auditing mechanisms for the long-term to mitigate the risk. A classification based on a systems-based approach will require to reconsider the evidence hierarchy as real-world evidence and pragmatic studies will be able to capture the variance of operational and clinical complexity more adequately.

# Background

The adoption of innovative healthcare technologies that improve patient outcomes and ensure cost-savings is a [priority for the NHS](https://www.england.nhs.uk/2019/06/nhs-aims-to-be-a-world-leader-in-ai-and-machine-learning-within-5-years/). The adoption of digital health technologies (DHTs) driven by intelligent computer programs and big data is such a priority area. These so called Artificial intelligence (AI) technologies are broadly defined as technologies which include algorithms that can learn from new experiences, adjust outputs and perform human-like tasks. NICE’s focus is on the health technology assessment (HTA) of these technologies. A robust framework for validating and evaluating the use of AI in digital health technologies, in terms of both their clinical effectiveness and their economic value does not exist. An important step to develop this framework understand the classification ‘building blocks’ of AI technologies that may impact on how the evidence is developed to demonstrate their effectiveness and safety.

# Aim

The aim of this report is to explore how to classify AI technologies at the lowest levels of granularity to enable an accurate initial triage for health technology assessment (HTA).

# Methods

Given the complexity of the topic a number of different resources were deemed necessary to provide a comprehensive view in a relatively short timeframe (Figure 1).

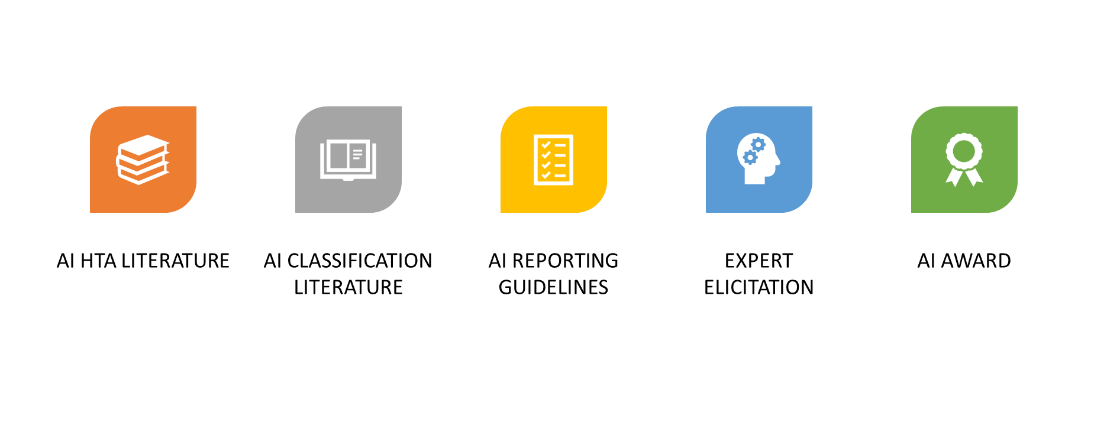


Figure 1: Information resources used during phase 1.

Since the topic was covering different aspects of AI, we used multiple searches to capture the aspects separately. The systematic search was run between 13 and 17 November 2020. The snowballing for references continues until 8 December 2020. The systematic search in MEDLINE and PubMed was done on 16th November 2020.

We followed the previous example of search methods and strategies for Precision Medicine project (Love-koh et al. 2018) and used the terms Artificial Intelligence, Machine Learning, ML, and AI combined with terms such as Classification, Technology Evaluation, Technology Assessment, Regulatory, Regulation and Economic Evaluation.

We used MEDLINE, Google, Google Scholar, and Google Images to search for the literature. We also looked at recent conference themes on AI.

We also run Website-specific searches using their Search function as well as Google Site search to capture both archived and new information for the national and international regulatory and HTA agencies websites such as the IMDRF, EUNETHTA, INAHTA, CADTH, AHRQ, HTAi, NICE, FDA, EMA, TGA.

We also browsed the reporting guidelines from the Equator Network for AI-specific guidelines. In addition, we consulted AI Experts from the KCL London Medical Imaging AI Centre for Value Based Healthcare to get their input on AI classification and the implications for HTA.

To find examples of AI technologies currently available in the UK we looked at the applications submitted as part of round 1 of the AI Awards from NIHR.

# AI classification

## Literature findings

Identifying a widely accepted and comprehensive classification of AI is a challenging task. We did not identify any published paper systematically focusing on classifying this rapidly emerging topic in a comprehensive way. Often, researchers and practitioners in the field try to adapt, change or modify the existing classification adding new classes based on their own preferences making an agreement on classification a challenging issue. This report tries to provide a comprehensive sample of objective[[1]](#footnote-1) and subjective[[2]](#footnote-2) classifications reported in the published and grey literature. In the following section, we refer to the published academic literature as well as webpages written by experts in the field to cover both types of classifications.

### Theoretical Classifications

There have been at least two proposed theoretical/philosophical classifications for AI:

#### Classification based on Functionality (EDUCBA 2020)

This classification relies on the degree to which a system can perform, imitate, replicate or adapt human functions. The classification tries to categorise the technologies into three sets based on their similarity to human mind in thinking and feeling (Joshi 2019):

* 1. *Reactive AI* with no memory that responds/reacts only to stimuli;
  2. *Limited Memory AI* mainly refers to the technologies than can use memory to learn and better their responses majority of which are Machine Learning (ML) models (Johnson 2020).
  3. *Theory of Mind AI* that refers to the system that can understand what other intelligent entities need.
  4. *Self-Aware AI* systems that possess human intelligence and an independent awareness.

While classes c and d are still pending to find a way towards practicality, classes a and b have found practical applications.

#### Classification based on Capabilities (EDUCBA 2020)

Depending on the range of tasks that the machine can perform and improve its performance, this classification suggests three categories (Joshi 2019):

1. *Artificial Narrow Intelligence* performs a task automatically following a process similar to humans. It can only perform the tasks that it has been programmed to perform and nothing beyond that. This class can include *Reactive AI* and/or *Limited Memory AI* from the Functionality classification.
2. *Artificial General Intelligence* can learn, understand and perform all tasks like a human. This class can contain *Theory of Mind AI* and *Self-Aware AI*.
3. *Artificial Superintelligence* is not only replicating the intelligence of human being but also goes beyond the capabilities of human mind through possessing more memory and processing speed that in combination with decision-making skills can potentially stand higher than human intelligence.

Class a from this classification is the only class to have existing practical applications and the other classes have been providing material for Sci-Fi movies and occasionally are referred to as singularity.

### Practical Classifications

The concept of AI emerged in the field of computer science, as a result, the existing literature classifications are generally following those suggested by computer science taxonomies with some modifications and adaptions.

#### Association for Computing Machinery Computing Classification System (ACM CCS)

The ACM CCS provides one of the most comprehensive AI-based classifications in the academic literature. This classification is hierarchically structured in four levels and each level can have more subcategories; however, its online edition represents five levels per concept. AI is a sub-level of Computing Methodologies and has two sub-levels in the hierarchy:

1. *Natural Language Processing* (8 categories as last level);
2. *Knowledge Representation and Reasoning* (12 categories);
3. *Planning and Scheduling* (5 categories);
4. *Search Methodologies* (7 categories);
5. *Control Methods* (3 categories);
6. *Philosophical/Theoretical Foundations of Artificial Intelligence* (2 categories);
7. *Distributed Artificial Intelligence* (4 categories);
8. *Computer Vision* (4 categories).

Unlike other classifications, this system has decided to consider AI and ML at the same level, rather than ML being a sub-level to AI. ML also is a sub-level of Computing Methodologies and has its own sub-levels:

1. *Learning Paradigms*
   * 1. Supervised Learning (6 sub-levels) including Ranking, Supervised Learning by Classification [related to Medical Imaging] and by Regression;
     2. Unsupervised Learning (7 sub-levels) including Cluster Analysis and Anomaly Detection;
     3. Reinforcement Learning (5 sub-levels) including Sequential Decision Making;
     4. Multi-Task Learning (3 sub-levels) including Lifelong Machine Learning and Learning under Covariate Shift;
2. *Learning Settings* (7 sub-levels)
3. *Machine Learning Approaches* (14 sub-levels) including Classification and Regression Trees, Kernel Methods, Neural Networks, Rule Learning, Markov Decision Processes, and Bio-Inspired Approaches
4. *Machine Learning Algorithms* (5 sub-levels)
5. *Cross-Validation* (no sub-level)

Although this classification system covers most of the academic literature in the field, its focus is on research-based literature and not on technologies with practical applications. Last but not least, this classification has officially been updated in 2012 while the ACM Digital Library seems to use a living version of this classification. This might be the reason why it does not cover some of the most recent and important areas such as Deep Learning which is classified under Neural Networks in the ACM Digital Library. Regardless of all the shortcomings, this classification has been the basis for almost all the other suggested classifications for AI including the classifications in textbooks and encyclopaedias such as Wikipedia (Wikipedia 2020).

#### Classification of AI Techniques (WIPO 2020)

The classification suggested and used by the World Intellectual Property Organization (WIPO) is one of the most practical, realistic, and easy to understand classifications. While it does not provide much information about AI classifications, it follows and expands the ACM CCS system for ML. The WIPO classification uses the following classes of AI patents in a 2-level system:

* 1. *Fuzzy Logic* (no sub-level)
  2. *Ontology Engineering* (no sub-level)
  3. *Probabilistic Reasoning* (no sub-level)
  4. *Logic Programming* (3 sub-levels): General, Description Logistics, and Expert Systems
  5. *Machine Learning* (15 sub-levels): General ML, Supervised Learning, Unsupervised Learning, Reinforced Learning, Multi-Task Learning, Rule Learning, Instance-Based Learning, Deep Learning, Logical and Relational Learning, Classification and Regression Trees, Support Vector Machines, Neural Networks, Probabilistic Graphical Models, Latent Representation, and Bio-Inspired Approaches

Machine learning is the dominant AI technique disclosed in patents and is included in more than one-third of all identified inventions. For healthcare specifically, ML is the main application[[3]](#footnote-3). Chen et al. (2008) report 10 classes of AI techniques half of which overlap with the WIPO's system. The other five classes include Multi-Agent Systems that are using more than one Intelligent Agent, and Hybrid Systems that are using two AI techniques such as Fuzzy Logic and Neural Networks at the same time. They also consider algorithms such as Evolutionary Algorithms (i.e. Genetic Algorithms), Swarm Intelligence and Cellular Automata as other classes.

#### Classification of AI Functional Applications (WIPO 2020)

The Functional Applications of AI have also been suggested and used by WIPO for the purpose of reporting progress with patent registrations. It seems similarly practical to ACM CCS without being very granular.

#### Classification of Ways to Achieve AI

Classification of Ways to Achieve AI is what some computer science practitioners have used to classify the AI methods (Kumar GN 2020). This classification is generally a shorter version of the ACM CCS that a) uses only the terminology/classes that are familiar/interesting to lay audience; b) considers ML as a Way to Achieve AI and a sub-level of AI; and c) merges several sub-levels into one level.

* 1. *Classification of Machine Learning* (Certes 2018) as the main sub-level of *Ways to Achieve AI* have also been suggested in ACM CCS but focusing only *ML Learning Paradigms* and *ML Algorithms* (Fumo 2017) ignoring *Learning Settings*, and *ML Approaches*. Two main consideration for this classification is 1. *Deep Learning* is listed in as same level as *Supervised Learning*; 2. *Semi-Supervised Learning* has been suggested as a new *ML Paradigm* which is not a class of ACM CCS.
  2. *Grouping ML Algorithms by Similarity* (Brownlee 2019) partially follows ACM CCS however provides a more updated classes of Algorithms compared to ACM CCS.

This classification clarifies that it has excluded task-specific algorithms or the algorithms for subfields of ML.

* 1. *Classification based of Types of Classification Tasks in Machine Learning* (Brownlee 2020b) or Classification Predictive Modelling is also an important way of classifying ML classification task because many AI technologies used in Medical Imaging are using Supervised Learning by Classification. Classification can be Binary, Multi-Class, Multi-Label or Imbalanced.
  2. *Classification of ML based on Types of Learning* (Brownlee 2019) follows ACM CCS ML Paradigms adding new types of learning which are unique (such as Multi-Instance Learning) and sometimes borrowed from educational literature (Inductive and Deductive Learning).

Cognilytica (2019) tried to cover all the businesses that are using AI and Corea (2018) visualised an AI Knowledge Map that highlights the overlapping areas of AI knowledge. ML could be supervised, unsupervised, or semi-supervised. Semi-supervised systems are both supervised and unsupervised. In addition, while assigning a device/technology to a class, some of them may be using ML and NLP at the same time to deliver the output.

Based on WIPO (2019), Medical Imaging is one of 11 sub-fields in Life and Medical Sciences with registered patents and holds the second place in terms of number of patents after the field of Physiological Parameter Monitoring. AI in Diagnostic Imaging could be considered under Computer Vision Class of AI which is using ML more than any other AI techniques to achieve AI. Among ML Paradigms, Diagnostic Imaging is using Supervised Learning by Classification and among ML Approaches, it uses Deep Learning sub-class of Neural Networks class (Could be Unsupervised) more than other paradigms and approaches. Under ML class, AI systems can be classified based on Classification Task (Binary, Multi-Class, Multi-Label or Imbalanced), based on Type of Learning, or their algorithms can be grouped based on Similarity. It is possible but not necessarily useful to provide more levels and sub-classes for each of these classes for example there are several ways to classify artificial neural networks or algorithms used in AI systems.

There are several ways to visualise AI classifications: Hierarchical Diagrams, Algorithmic Process Diagrams, Venn Diagrams, and Knowledge Maps. Each of these visualisations has their own advantages and disadvantages.

## Expert elicitation

### Definition of AI

Initial discussions were held with 2 AI experts from the KCL London Medical Imaging AI Centre for Value Based Healthcare (Professor Michael Luck and Dr Jorge Cardoso). A formative open-ended questionnaire was developed to structure the discussion. The items in the questionnaire were based on NICE research questions and preliminary review of the literature.

Neither expert believed there was a standard definition of AI. At the most basic level, AI was defined as a system able to perform tasks normally requiring human intelligence[[4]](#footnote-4). Experts highlighted that the terms Machine Learning (ML)[[5]](#footnote-5) and AI often used interchangeably. ML is often used to describe AI, however ML is a subset of AI (with other subsets including Natural Language Processing (NLP) and computer vision).

One expert suggested that looking at AI conferences[[6]](#footnote-6) and topics is one way to “cut the cake”. ML was expected to be at least a component of most/all upcoming health AI products and that ML was likely to increasingly feature in new healthcare software. ML was described as the “revolutionary” area of AI. It is also the subset of AI that is often perceived as a “black box” as the system learns, and this is where a lot of potential risk will be found.

### Pre-existing AI classifications

Experts were not aware of any pre-existing AI classifications beyond the standard classes of the medical device directive. The mentioned that dichotomies could be created to generate classifications, such as supervised/unsupervised, descriptive/predictive/prescriptive models, fixed/continuously-learning models, however, experts warned that classifying ML systems using technically defined classifications[[7]](#footnote-7) (“artificial partitions”) may not be useful due to the number of “edge cases” (edge case are problem or situations that occur only at an extreme operating parameter). Risk of AI was deemed to be well defined, but technical clustering was not. Most AI (and ML in particular) software was described as between fields. Experts noted any AI classification system would benefit from being flexible or open ended to account for the heterogeneity in technologies.

### Thoughts on AHSN and CQC suggestions for classifying AI

The experts felt that organising technologies by complexity (as per [AHSN](https://ai.ahsnnetwork.com/about/complexity-scale/)) would work in theory but would be too challenging to operationalise. It was felt that organising technologies by clinical risk and autonomy (as per the [CQC suggestion](https://www.cqc.org.uk/sites/default/files/20200324%20CQC%20sandbox%20report_machine%20learning%20in%20diagnostic%20services.pdf)) was more pragmatic in terms of real differences that could be measured. Level of autonomy is important for classifying potential levels of risk/safety and could be roughly grouped into autonomous or not, noting there are grey areas, such as augmented intelligence. In terms of the clinical scope groupings, one expert noted that the narrower scope categories were currently most relevant in terms of healthcare.

### Examples of useful categorisations

Experts suggested a number of general categorisations:

1. Supervised versus unsupervised algorithms may be useful method of classifying ML algorithms, however, this is not exhaustive of all types of AI.
2. ML systems that are fixed versus continuously learning. Algorithms may learn after deployment; however, this is a challenging regulatory area.
3. Algorithms that interact versus work in parallel with clinical systems. Parallel systems are less complicated and may therefore be simpler to assess.
4. Ground truth/gold standard to refer to versus something that is medical opinion.

# Existing regulatory frameworks

## Literature findings

A number of international regulators such as the FDA are currently in the process of exploring the changes needed to evaluate AI technologies in the context of healthcare. The included regulatory frameworks represent organisations from the USA, EU, Australia, Japan, Russia, China, Canada, South Korea and Singapore. A detailed summary is provided in appendix 14.4. Leading organisations in the field seems to be the IMDRF, as most countries seem to adopt and/or adapt their proposals. The IMDRF is also hosting the Artificial Intelligence Medical Devices (AIMDs) Working Group. The AIMDs group has two initial aims, to achieve a harmonised approach to the management of AI medical devices and to establish a guidance to share the views on establishing a common terminology.

# Reporting guidelines

Several reporting guidelines for AI technology in healthcare have been published in recent years, with several more currently in production. These guidelines can provide some understanding of the various ways that AI algorithms can be classified, based on the way they are reported and evaluated in the literature. Publication of their draft recommendations is expected in July 2021.

## CONSORT & SPIRIT AI extensions

[CONSORT](https://www.nature.com/articles/s41591-020-1034-x) is a reporting checklist for Randomised Controlled Trial (RCT), while [SPIRIT](https://www.clinical-trials.ai/spirit) is used for the reporting of protocols for RCTs. AI-specific extensions to both checklists were developed in parallel and published in 2020. Several Delphi rounds were used in the development of the checklists and stakeholders included healthcare professionals, methodologists, statisticians, computer scientists, industry representatives, journal editors, policy makers, health ‘informaticists’, experts in law and ethics, regulators, patients and funders.

Important recommendations from SPIRIT-AI include the specification of the intended use for AI algorithms (CONSORT-AI 1ii and 6i), as some AI interventions may have multiple intended uses or the intended use may evolve over time and to also clarify if the AI is replacing, augmenting or adjudicating components of clinical decision-making. Item 9 about the study setting has been expanded to also include information about the operational environment, including hardware and software requirements. Extension for item 10 requests to state the inclusion and exclusion criteria not just at the level of the participants but also of the input data. A lot of emphasis is also placed on how missing data was handled and how performance errors were analysed and reported (items 11 and 22). For the purpose of this review we looked at the following reporting guidelines (Cruz Rivera, S., et al. (2020); Hernandez-Boussard, T., et al. (2020); Liu, X., et al. (2020); Luo, W., et al. (2016); Moons, K. G., et al. (2015); Norgeot, B., et al. (2020); Sengupta, P. P., et al. (2020); Sounderajah, V., et al. (2020)). A detailed summary of the reporting guidelines changes is presented in appendix 14.1.

The items on the checklists vary based on their purpose but the majority of items can be categorised into one of the following:

* 1. Acquisition and selection of data for training and validation
  2. How poor quality/missing data is handled
  3. Human-AI interaction
  4. Accessibility of code
  5. How AI output will contribute to decision making
  6. How performance errors are analysed and identified, and the practical costs related to them
  7. Algorithm ‘fine-tuning’
  8. Type of algorithm (although there are no strict definitions)

Several of these can be used to classify algorithms into different categories. Fixed and continuous algorithms can be considered separately, for example. The level of human-algorithm interaction can also be separated into discrete (but subjective) categories. This may include fully autonomous, partly autonomous or non-autonomous. For example, an algorithm which is fully autonomous may provide a diagnosis with no input from a clinician, while a partly autonomous algorithm would require a clinician to make a final decision. The output of a non-autonomous system would require a clinician to interpret and actioned. Experts suggest that these classifications are currently skewed, in terms of healthcare applications, towards fixed and partly autonomous AI algorithms (although this could change in the future).

The acquisition and selection of data for training and validation is crucial in assessing the efficacy, safety and generalisability of any algorithm. Any classification should take into consideration that some systems will require more robust evaluation of this information (see HTA Insights section).

Performance errors and failure analysis (see also HTA insights section) are particularly important as all AI systems will make some mistakes that may be difficult for humans to foresee. Therefore, where cases of performance error are identified in the development or validation of an algorithm, these should be reported such that risk-mitigation strategies can be implemented.

## Model types

The majority of reporting guidelines do not specify their applicability based on the ‘type’ or model of AI[[8]](#footnote-8). Luo et al. 2016, however, in their guideline on the use of predictive models in clinical settings name the following as types of predictive model:

1. Decision tree
2. Random Forest
3. Lasso Regression
4. Gradient boosting machines
5. Support vector machines

They also suggest that the ‘form’ of the model should be reported. This can be **classification** if the target variable is categorical, **regression** if the target variable is continuous, **survival prediction** if the target variable is the time to an event. These categories may be useful for these types of prediction model, which make up the majority of AI technologies in healthcare currently. It is anticipated that TRIPOD-ML, the ML-specific extension of the TRIPOD checklist for reporting the development and validation of prediction models will also provide further information on this.

## Conference Proceedings

Several conference proceedings were searched to try to understand how AI is being split into topics for discussion. Two computer-science conferences were suggested (see section 6.2.1). Appendix 14.2 provides a detailed summary of how the topics were categorised. There is a wide range of topics and little consistency between the conferences. Machine learning is the main topic (in terms of the number of presentations and panels that fall into that category). Although it was suggested as a possible starting point from the experts, it appears to be counterproductive to attempt to break AI applications down in this way for the purposes of developing a classification for use in healthcare.

A search for “Artificial Intelligence” on the IEEE conference website was performed, returning 131 conferences for the year 2020. It was therefore considered unfeasible to systematically search all AI conference proceeding. Instead, conferences on AI in medicine specifically were also searched on google. The [IntelligentHealth AI 2020](https://intelligenthealth.ai/programme-download/) programme had the following themes:

1. The future of healthcare
2. Data everywhere
3. Covid-19
4. Regulators Perspective AI Health transformers
5. Healthcare without borders
6. Creating an AI enabling landscape
7. The role of AI in genomics.

This list is significantly shorter than those found in the computer-science conferences. However, these themes are broad and similarly unhelpful. A further general healthcare conference was searched for AI-related topics. The [ECBIOS 2021](http://2020.ecbios.asia/) topic list included the following:

1. Smart healthcare system analysis and design
2. Computer and human-machine interaction of healthcare systems
3. Application of IoT (Internet of Things) on healthcare systems
4. Big data and artificial intelligence enabled healthcare systems.

Again, these are very broad but could potentially be useful starting point for a high-level classification.

# AI evaluation – key considerations

## Literature findings

### Systems vs. product-based approach

Most international regulators and HTA organisations are product-based - they review and ultimately approve or reject healthcare products such as pharmaceuticals and devices. For these healthcare technologies the regulatory and best-practice frameworks are well-established and allow all stakeholders to confidently develop, assess, procure and maintain these technologies. AI-based technologies present unique clinical and operational complexities and risks that challenge the evaluation under existing regulatory and HTA frameworks. It is likely that AI technologies will present more variance between their performance in an artificial testing environment and in actual practice settings, and thus potentially more risks and less certainty over their benefits (Gerke et al., 2020). Variance can increase due the complexity of these systems and how they interact with their environment. For example, [FDA approved in 2018 the IDx-DR](https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye), the first ever autonomous AI-based system cleared to provide a diagnostic decision, however, recently the ACR and RSNA (two of the largest radiological professional associations) wrote an [open letter to the FDA](https://www.acr.org/-/media/ACR/NOINDEX/Advocacy/acr_rsna_comments_fda-ai-evolvingrole-ws_6-30-2020.pdf) asking the agency to wait for more rigorous testing and surveillance of the modality before authorizing its autonomous implementation in medical imaging. According to their statement they believe it is unlikely that the FDA could provide reasonable assurance of the safety and effectiveness of autonomous AI in radiology patient care without more rigorous testing, surveillance, and other oversight mechanisms throughout the total product life cycle.

The impact of AI-based technologies depends on many factors of a broader system including human and organisational factors. This indicates that the regulatory focus should be on designing an appropriate process for managing this new ecosystem taking a more system-based approach than a product-based one Gerke et al. (2020). A systems approach does not divide the development stages chronologically and it does not divide the process in discreet steps.

### United States Government Accountability Office

The AI in Healthcare report states that for high-risk AI tools, evidence on clinical safety and effectiveness is of paramount importance even as other metrics are considered. High-risk tools will require evidence from rigorous studies for regulatory purposes and will require substantial monitoring at the time of and following implementation. For low-risk clinical AI tools such as those used at point-of-care, or those that focus on administrative tasks, evaluation may focus on process of care measures and metrics related to the AI’s usage in practice to define its positive and negative effects. The report accepts a wide range of evaluation study design, ranging from RCTs to quasi-experimental approaches such as stepped-wedge designs or even carefully matched retrospective cohort studies.

### Park and Han (2018)

Park and Han (2018) published a methodological guide for the evaluation of imaging-based AI technologies in healthcare (diagnosis and prediction). This in-depth report provides useful understanding of the main requirements for evaluating AI models. The methods highlighted can provide a starting point towards a classification, once the evidence requirements for different models are apparent.

The authors also provide an explanation why ML is particularly useful in diagnosis is its ability to detect patterns corroborating the experts view that ML dominates most healthcare applications. They also note that DL makes use of artificial neural networks, which consist of a stack of multiple layers of artificial neuronal links that loosely simulate the brain’s neuronal connections. A convolutional neural network conceptually mimics the visual pathway, making it particularly good for analysing images. These points reinforce the expert’s view that machine learning is used in almost all healthcare applications of AI and the finding that machine learning is the largest AI conference topic.

Park and Han posit that good understanding and application of relevant epidemiologic and statistical principles is necessary to evaluate AI technologies. They define ‘validation’ as a specific step in model development, in a sequence of steps consisting of training, validation, and testing. Validation is the step that fine-tunes the model or selects the most optimised model. Testing, then, is the process of evaluating model performance, and so is more like how the term validation is used in the field of medicine more generally.

#### Statistical Methods

**Discrimination Performance**

When evaluating a diagnostic tool, with a binary outcome (disease vs no disease) sensitivity and specificity are generally used to assess performance. AI algorithms work slightly differently, however. Even if an AI algorithm output is binary, to generate this output it will first calculate a probability-like continuous output and then use a threshold value to convert this into a dichotomous result. So, a Receiver Operator Curve (ROC) should be generated and the Area Under the Curve (AUC) can be used as a metric for performance. It should be noted, however, that like sensitivity and specificity, AUC is independent of disease prevalence.

**Calibration Performance**

If an algorithm produces a continuous output, it should be evaluated differently. In this case, how close the predicted probability value is to the real probability value is what is being tested. This is known as calibration. Here, a goodness-of-fit test, such as the Homer-Lemeshow test, should be used to evaluate the algorithm.

**Internal vs. External Datasets**

Evaluating model performance using the same data that were used to develop the model is known as **internal validation**, while the use of other data is known as **external validation. I**nternal validation can substantially overestimate the algorithms performance and so external validation should be used.

**Overfitting** is where a learning model ‘customises’ itself to the training data to such an extent that it cannot be generalised to be used on new populations. This has been shown to be a particular problem in DL algorithms, and so external validation is an essential requirement for these types of algorithm. External validation can be **temporal** or **geographic.** Temporal external validation uses a validation dataset made up of newly recruited patients from the same setting as the test data set, while geographic external validation uses data collected by independent investigators at another site. Geographic external validation would be considered ideal, as it avoids overfitting by using a completely new population.

**Effect of the Spectrum on diagnostic or predictive performance**

A very large dataset is generally required for the development of deep neural networks, sometimes called “big data”. This can make standard prospective clinical trials very difficult and usually data has to be collected from multiple heterogeneous sources in various ways. Sometimes, developers may intentionally collect data from more positive or negative subjects, rather than in an equal ratio, to morph the data into the mathematical model. This is known as **convenience sampling** (as opposed to consecutive or random sampling) and could obscure severity, stage or duration of disease, the presence of comorbidities and other important factors. It also carries a risk of **spectrum bias**, where the test population does not align with the target population giving poor results. Retrospective collection of data is prone to this type of bias and so in ML, the spectrum of the population should be clearly reported and evaluated.

### Magrabi et al. (2019)

Magrabi et al. (2019) published a position paper on AI in clinical decision making. The paper proposes potential risk areas related to these models. Similar to Park and Han (2018), this paper also highlighted the problem of overfitting, even in restricted tasks like image interpretation. It was also noted that differences in image capture workflow can lead to erroneous results.

Ethical considerations are also discussed at length and it is noted that they can be completely separate from clinical benefit, as algorithms trained on a non-representative population may discriminate against certain demographics. For example, an algorithm used for prioritising patients for an organ donated for transplant might include the expected lifespan of the patient as a predictor variable. This ignores the correlation with socio-economic factors and their effect on health and lifespan. These potential ethical or equality considerations may be less obvious than in other areas of health and particular attention should be paid to them when assessing AI systems.

The authors of the paper suggest that auditing may be a pragmatic approach to “black-box” applications but also argue that the onus should be on developers to make their applications more explainable and accessible. Continuous algorithms in particular would require continued surveillance to ensure that performance does not suffer.

### Safety Analysis

Xie et al. (2020) in their paper on AI in Diabetic Retinopathy Screening, describe another important form of AI-model evaluation. FMEA (Failure Mode and Effects Analysis) involves creating a detailed map of processes for a service or activity to identify all the ways that a process may fail. This can be used in AI HTA but may be very resource intensive. This analysis can be considered to be an evaluation of the performance errors that should be reported as per the guidelines mentioned in section 8 and can help to mitigate the risks associated with them.

### Olthof et al. (2020)

Olthof et al. (2010) suggested categorising Ai by the type of impact on the work of neuroradiologists. According to the technographic review, most functionalities of 37 applications (39; 54%) are designed to ‘support’ radiologists in performing their current tasks more efficiently. Example: visualization the images and information. Some other functionalities of applications ‘extend’ the work of radiologists by providing quantitative information that would not be possible before the introduction of these applications (23; 32%). This adds a new task that may require improving human capabilities and does require human intervention but solves the problem via an algorithm. Example: provide diagnostic information that was not available before, such as a heatmap of suspicious areas. Only a few functionalities of applications (10; 14%) offer functionalities that take over certain tasks. This does not require human involvement and fundamentally changes a task. Example: autonomous reading and reporting radiology cases. The authors note that a number of AI applications for neuroradiology include a number of different functions, for example, the AI reader (Qmenta) supports the radiologist in making a diagnosis, extends radiology work by providing quantitative information and can replace the radiologist in drafting an automated report for the referring clinician.

## Expert elicitation

### Overlap with other evaluation processes

There are generic elements of the process in terms of evidence generation and NICE level evaluation that will be the same as for Medical Technologies and for Digital Health Technologies (DHTs), including AI. In terms of evaluation, one expert did not believe there was currently a large difference between AI and other technologies, with the main difference being that an AI technology’s risk profile may need more stringent assessment.

### Key consideration in evaluating ML

Two key considerations were noted by the experts, first the variability in data types and patterns is key to testing and evaluation. Training algorithms on highly curated data[[9]](#footnote-9) results in mistakes in real-world applications. Evaluation should usually involve providing algorithms with edge cases, for example atypical images acquired in unusual ways or images acquired with different hardware. The aim would be to change the distribution of data. The population distribution that an algorithm is trained on is very relevant (e.g. socio-economic, ethnic, diet etc.). Processing unusual data is where an otherwise good algorithm can fail. A strong algorithm can extrapolate to unseen populations. Example: Mia AI algorithm for mammography - training algorithm did not translate well to British population. Algorithm was good, but the data it was trained on did not generalise to British population.

Secondly, a major difference between ML and medical devices is the un-explainability of ML. One expert highlighted that uncertainties include how to carry out failure analysis in “black box” situations. The expert suggested that previously we knew more about internal processes of ML, but performance tended to be not as good, and that increasingly ML performs better now despite humans having less knowledge of these processes. The expert posed the question of whether knowledge of the internal processes was important as long as the outcomes were improving.

### Approach to evidence generation

Locally, evidence generation may vary significantly as many hospitals are not AI-ready. Clinical pathways are also different across trusts, meaning that certain AI models would only work under certain pathway assumptions. Clinical pathways also change in time (new guidelines), so fitness-for-purpose may change. One expert noted, however, that the same applies to other types of devices/drugs/procedures, so AI shouldn’t be an exception.

One expert supported the idea that clinical effectiveness can be demonstrated with either retrospective or prospective data (depending on the model), and be measured against current clinical performance targets, often defined by commissioning groups. One expert suggested that interactive systems such as prescribing advice systems (that need to interact with the rest of treatment) need prospective analysis as every new patient is different and the important metric is improved clinical outcomes (rather than “is this the right drug”). On the contrary, diagnostic systems may use retrospective data. The important factor is how data is curated and fed into an algorithm and not necessarily the whether the data is retrospective or prospective[[10]](#footnote-10). Aversion to retrospective data is primarily due to poor design of retrospective data collection.

### Future proofing – changes expected in the next 5 years

The experts suggested that in the future AI systems will be more complex (with complexity pertaining to the different sources of data used to feed the system). Experts felt that understanding the AI decisions in complex scenarios will be more challenging, if not impossible. Currently systems are usually applied to single “ologies” that use data from one speciality. Clinical benefits may be more difficult to define and measure once technologies start interacting. There will also be an increase in “black box” products. There may be risk involved with not understating internal processes. Finally, performance of the algorithms will likely start to surpass that of humans, and consequently will require new ways to assess system performance that is not comparing to human performance, raising questions about reference standards applicability.

# Phase 4 AI award technologies

Using the parameters suggested in the briefing note and the experts views we attempted to classify[[11]](#footnote-11) the [10 AI technologies](https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/ai-health-and-care-award-winners/) successfully awarded a phase 4 AI award. The categories used were complexity (suggested by the AHSN), autonomy (supervised, semi-supervised, un-supervised) as suggested by the CQC, algorithm learning status (fixed vs. continuous), explainability (no black box, some-information provided, black box), interaction (system interaction, for example interaction with existing NHS IT software like PACS), clinical pathway (as suggested by the CQC) and functionality (Olthof et al. 2020). For the categories of complexity, autonomy and learning status the majority of the technologies are clustering in the same category. Eight out of 10 AI systems encompassed between 4 and 7 of the functions listed in Olthof et al. (2020). For a detail summary of the classification please look at appendix 14.6.

# Discussion

This work describes the systematic search and review of AI literature and AI-related reporting guidelines to understand how AI technologies can potentially be classified at the lowest level of granularity to enable an accurate initial triage for HTA. The output is a preliminary group of categories based on the review of literature and discussion with experts at the London Medical Imaging and AI Centre for Value-Based Healthcare.

This is not the first effort to classify AI as a field of research, policy, and practice, however, the substantial overlap between areas makes it difficult to rely on binary classifications, such as the ACM CCS, where a class is fuzzy between two or more other classes. The lack of consensus in the literature around a classification was corroborated by the experts who were not aware of any pre-existing AI classifications beyond the standard classes of the medical device directive. The increasing number and heterogeneity of AI applications and the speed at which the field is evolving also inherently challenges the development of meaningful classification frameworks for AI technologies. A single AI technology may increasingly encompass multiple functionalities. In a technographic review, Olthof et al. (2020) noted that a number of AI applications for neuroradiology include different functions, for example, the AI reader (Qmenta) supports the radiologist in making a diagnosis, extends radiology work by providing quantitative information and can replace the radiologist in drafting an automated report for the referring clinician. If we would attempt to use the Digital Health Technologies evidence standard framework, to categorise Qmenta it can be widely classified between Tier 2 and 3b depending on how it is going to be used in practice. This can also be seen in the technologies selected for the NIHR Accelerated Access Collaborative’s AI in Health and Care Phase 4 Awards; 8 out of 10 AI systems encompassed between 4 and 7 of the functions listed in Olthof et al. (2020). Discussions with experts elicited potential ways to categorise apps for evaluation, but also drew caveats, such as that classifying ML systems in particular using technically defined classifications (“artificial partitions”), may not be useful due to the high number of “edge cases” ML could be supervised, unsupervised, or semi-supervised. In addition, while assigning a device/technology to a class, some of them may be using ML and NLP at the same time to deliver the output. It may, therefore, be more practical to consider a classification based on more general criteria such the ones propose by the CQC. To this end, the experts recommended that classifications at the early stages of development remain simple, flexible, open-ended, and potentially formative to account for the heterogeneity of the technology and the evolution of the field in the following years.

The risk associated with introducing AI to healthcare is widely understood to be a key consideration in evaluation. Most critical issues with risk and safety would be addressed by regulators before a product can be marketed. Risk itself is, however, a broad category and though theoretically it is well documented, the nature of risk will potentially be different for each software e.g. risk of incorrect advice, risk of omitting important information, physical risk, more broadly risk of deskilling. Defining risk (both anticipated and unanticipated) and how this relates to different categories of AI requires further research.

In terms of assessing effectiveness, experts felt that AI technologies would require similar metrics to non-AI technologies, but that the scope would potentially need to change. Similarly, Gerke et al. (2020) argues that as AI becomes increasingly complex and interactive that a system- rather than product-based approach will be required. New comparators or reference standards may need to be included as AI performance begins to reliably outperform human. In addition to comparator, the scope of evaluation may need extra thought (system or product as the ML gets increasingly complex and interactive). A more holistic assessment that places more emphasis on human factors will also be important as there may be variation in how the product is used on an individual level. Likewise, the setting of evaluation and generalisability will be important considerations as factors such as organisational and operational factors are more intertwined with product performance.

For this preliminary phase I we used a combination of the proposed classifications in the NICE briefing note, the experts’ suggestions and the functionality dimension from Olthof et al. (2020) to classify the winner technologies of the Phase 4 AI award. The results demonstrate why a classification that is valid on a theoretical level (for example the AHSN complexity scale) may be remote from the practical applications of AI in healthcare and therefore, operationally irrelevant if the majority of existing applications are skewed towards one of the suggested categories in each class. To address this latter point, further iterative development of the proposed classification for AI technologies is required. The classification will need to be validated by expert review and applied to a larger sample of AI technologies and conclude with a checklist that can be used to classify AI technologies.

# Conclusions

* The findings from the literature review and patent applications information from the WIPO report corroborate the experts view that ML applications are currently dominating the healthcare field.
* A low granularity classification (closer to the CQC rather than the AHSN approach) seems to be the most pragmatic and applicable approach. This finding is corroborated by the plurality of AI classifications, the dominance of ML applications in healthcare and the experts view. Some countries such as Australia and South Korea have already adopted this high-level approach in their proposed regulatory frameworks.
* From a regulatory perspective, triaging risk seems to be a recurring theme. Autonomous systems and systems that are continuously evolving ML are suggested as higher risk. Harder to evaluate as autonomous systems may be more “black box” – potentially requiring a more pragmatic approach based on outcomes. This will shift the evaluation focus on carefully selecting those outcomes and providing auditing mechanisms for the long-term to mitigate the risk.
* The complexity of evaluating AI requires a system-approach; the current product-based approach of the regulatory system is not fit for purpose.
* A classification that is valid on a theoretical level, such as the AHSN complexity scale, may be remote from the practical applications of AI in healthcare, and therefore, operationally irrelevant if the majority of existing applications are skewed towards one of the suggested categories in each class.
* A classification based on a systems-based approach will require to reconsider the evidence hierarchy as real-world evidence and pragmatic studies will be able to capture the variance of operational and clinical complexity more adequately.

# References

ACM. ACM Computing Classification System. 2012 [Update?]

BENJAMENS, S., DHUNNOO, P. & MESKO, B. 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med, 3, 118.

Brownlee, Jason. 14 Different Types of Learning in Machine Learning. Machine Learning Mastery November 11, 2019. Available from: <https://machinelearningmastery.com/types-of-learning-in-machine-learning/>

Brownlee, Jason. 4 Types of Classification Tasks in Machine Learning. Machine Learning Mastery April 8, 2020. Available from: <https://machinelearningmastery.com/types-of-classification-in-machine-learning/>

Brownlee, Jason. A Tour of Machine Learning Algorithms. Machine Learning Mastery August 12, 2019. Available from: <https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>

Chen SH, Jakeman AJ, Norton JP. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. Mathematics and Computers in Simulation 2008; 78:379-400

Cruz Rivera, S., et al. (2020). "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension." Nat Med 26(9): 1351-1363.

Certes Computing Ltd. Types of Artificial Intelligence: A Detailed Guide. Certes December 20, 2018. Available from: <https://certes.co.uk/types-of-artificial-intelligence-a-detailed-guide/> [Accessed: November 13, 2020]

Cognilytica. Cognilytica’s Classification of the AI Vendor Ecosystem: Part I. January 16, 2019. Available from: <https://www.cognilytica.com/2019/01/16/cognilyticas-classification-of-the-ai-vendor-ecosystem-overview-and-bottom-3-layers/>

Cognilytica. Cognilytica’s Classification of the AI Vendor Ecosystem: Part II. January 21, 2019. Available from: <https://www.cognilytica.com/2019/01/21/cognilyticas-classification-of-the-ai-vendor-ecosystem-part-ii/>

Corea, Francesco. AI Knowledge Map: How To Classify AI Technologies. KDnuggets™ News September 5, 2018; 18:n33. Available from: <https://www.kdnuggets.com/2018/08/ai-knowledge-map-classify-ai-technologies.html>

EDUCBA. Introduction to Types of Artificial Intelligence. <https://www.educba.com/types-of-artificial-intelligence/>

Fumo, David. Types of Machine Learning Algorithms You Should Know. Towards Data Science June 15, 2017. Available from: <https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861>

GERKE, S., BABIC, B., EVGENIOU, T. & COHEN, I. G. 2020. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. npj Digital Medicine, 3, 53.

Johnson, Jonathan. 4 Types of Artificial Intelligence. BMC June 8, 2020. Available from: <https://www.bmc.com/blogs/artificial-intelligence-types/> [Accessed: November 13, 2020]

Gómez-González, E., et al. (2020). "Artificial intelligence in medicine and healthcare: a review and classification of current and near-future applications and their ethical and social Impact." arXiv:2001.09778.

Gottinger, H. W. (2018). "Technology Assessment and Forecasting of Medical Expert Systems (MEST)." Methods of Information in Medicine **27**(02): 58-66.

HARVEY, H. B. & GOWDA, V. 2020. How the FDA Regulates AI. Acad Radiol, 27, 58-61.

Haverinen, J., et al. (2019). "Digi-HTA: Health technology assessment framework for digital healthcare services." Finnish Journal of eHealth and eWelfare **11**(4).

Hernandez-Boussard, T., et al. (2020). "MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care." J Am Med Inform Assoc.

Joshi, Naveen. 7 Types Of Artificial Intelligence. Forbes June 19, 2019 [10:54pm EDT]. Available from:<https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/> [Accessed: November 13, 2020]

Kumar GN, Chethan. Artificial Intelligence: Definition, Types, Examples, Technologies. Medium 31 August, 2018. Available from: <https://medium.com/@chethankumargn/artificial-intelligence-definition-types-examples-technologies-962ea75c7b9b> [Accessed: November 13, 2020]

Liu, X., et al. (2020). "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension." Nat Med 26(9): 1364-1374.

Love-Koh, J., A. Peel, J. C. Rejon-Parrilla, K. Ennis, R. Lovett, A. Manca, A. Chalkidou, H. Wood and M. Taylor (2018). "The Future of Precision Medicine: Potential Impacts for Health Technology Assessment." Pharmacoeconomics 36(12): 1439-1451.

Luo, W., et al. (2016). "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View." J Med Internet Res 18(12): e323.

Moons, K. G., et al. (2015). "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration." Ann Intern Med 162(1): W1-73.

Magrabi, F., et al. (2019). "Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications." Yearb Med Inform **28**(1): 128-134.

Park, S. H. and K. Han (2018). "Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction." Radiology **286**(3): 800-809.

Norgeot, B., et al. (2020). "Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist." Nat Med 26(9): 1320-1324.

Park, S. H. and K. Han (2018). "Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction." Radiology 286(3): 800-809.

Sengupta, P. P., et al. (2020). "Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): A Checklist: Reviewed by the American College of Cardiology Healthcare Innovation Council." JACC Cardiovasc Imaging 13(9): 2017-2035.

Sounderajah, V., et al. (2020). "Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group." Nat Med 26(6): 807-808.

Xie, Y., et al. (2020). "Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening." Transl Vis Sci Technol **9**(2): 22.

Wikipedia contributors. (2020, November 23). Artificial intelligence. In Wikipedia, The Free Encyclopedia. Retrieved 13:53, November 25, 2020, from <https://en.wikipedia.org/w/index.php?title=Artificial_intelligence&oldid=990175533>

WIPO. WIPO Technology Trends: Artificial Intelligence. ISBN: 978-92-805-3007-0 Available from: <https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf>

# Appendices

## Reporting Guidelines

| **Checklist** | **Purpose** | **Methods** | **Definitions** | **Items** | **Notes** |
| --- | --- | --- | --- | --- | --- |
| CONSORT-AI 2020 | Reporting of RCTs | Iterative Delphi method with 169 experts. Rated items 1-9 in importance and suggested other items. 2 day consensus meeting in Jan 2020, with 31 stakeholders. | **- Artificial Intelligence** The science of developing computer systems which can perform tasks normally requiring human intelligence.  - **Fine-tuning** Modifications or additional training performed on the AI intervention model, done with the intention of improving its performance.  - **Performance error** Instances in which the AI intervention fails to perform as expected. This term can describe different types of failures, and it is up to the investigator to specify what should be considered a performance error, preferably based on prior evidence. This can range from small decreases in accuracy (compared to expected accuracy) to erroneous predictions or the inability to produce an output, in certain cases. | * Onsite and offsite requirements of AI (Data collected onsite vs where data is stored offsite for example) * How input data was acquired and selected * How poor quality or missing data is handled * Human-AI interaction + expertise of clinicians * How performance errors are analysed and identified * Can AI code be accessed? |  |
| SPIRIT-AI 2020 | Reporting of Protocols for RCTs | As above | As Above | * Specify type of model * Specify intended use of model * Pre-existing evidence of use of model * Inclusion and exclusion criteria * Version of algorithm * Onsite and offsite requirements of AI * How input data was acquired and selected * How poor quality or missing data is handled * Human-AI interaction + expertise * Explain how the AI output will contribute to decision making or other elements of clinical practice. * How performance errors are analysed and identified * Can AI code be accessed? | * More items than CONSORT |
| Luo 2016 | Guideline on use of predictive models in clinical settings | Delphi method, 4 iterations, 11 members of panel. | Gives the following as types of predictive model:   * Decision tree * Random Forest * Lasso Regression * Gradient boosting machines * Support vector machines   Descriptions in an appendix I of the paper.  Quality metrics:   * Internal validation * External validation * Response variable * Independent variable * TP, TN, FP, FN, Sensitivity, specificity, PPV, NPV * ROC | * Data sources (for training, validation and evaluation) * Performance metrics in both point estimates and Cis * Practical value of model * Review of prediction accuracy of existing models * Clinical setting including facility type, size, volume and duration of available data * Defined measurement of prediction goal (per patient or per hospitalisation etc) * Determine the form of the prediction model: (1) classification if the target variable is categorical, (2) regression if the target variable is continuous, (3) survival prediction if the target variable is the time to an event. * Translate survival prediction into a regression problem, with the target measured over a temporal window following the time of prediction. * Explain practical costs of prediction errors * Quality metrics (see prior column) * Define the observational units on which the response variable and predictor variables are defined. * Define the predictor variables. Extra caution is needed to prevent information leakage from the response variable to predictor variables. * Describe the data pre-processing performed, including data cleaning and transformation. Remove outliers with impossible or extreme responses; state any criteria used for outlier removal. * State how missing values were handled. * Specify the internal validation strategy. Common methods include random split, time-based split, and patient-based split. * For retrospective studies, split the data into a derivation set and a validation set. For prospective studies, define the starting time for validation data collection. * If possible, report the parameter estimates in the model and their confidence intervals. When the direct calculation of confidence intervals is not possible, report nonparametric estimates from bootstrap samples. * Interpretation of the final model. If possible, report what variables were shown to be predictive of the response variable. State which subpopulation has the best prediction and which subpopulation is most difficult to predict. * Report the clinical implications derived from the obtained predictive performance. For example, report the dollar amount that could be saved with better prediction. How many patients could benefit from a care model leveraging the model prediction? And to what extent? * Report unexpected signs of coefficients, indicating collinearity or complex interaction between predictor variables | * Some of these fairly standard for diagnostics, some more specific. |
| MINIMAR | Minimum reporting for medical AI reporting | No information |  | * Intended user of output * How data were split for training, testing and validation * Labelled data used to train and test * Algorithm type (e.g. Machine learning, deep learning etc) * Model task (classification or prediction) * List of variables used in model and how they were used in terms of categories or transformation * How missing data was addressed * Model or parameter tuning * Internal and external validation |  |
| TRIPOD - AI | Multivariable prediction models for individual prognosis or diagnosis | Not yet published. |  |  |  |
| PRIME | Cardiovascular imaging-related machine learning | Independent multidisciplinary panel |  | - Describe how the data were processed in order to make it clean, uniform, and consistent  - Describe whether variables were normalized and if so, how this was done  - Provide details on thefraction of missing values (if any) and imputation methods  - Describe any feature selection processes applied  - Identify and describe the process to handle outliers if any  - Describe whether class imbalance existed, and which method was applied to deal with it  - Provide a clear description of data used for training, validation, and testing  - If well performing models were tested on a hold-out validation dataset, detail the data of that validation set with the same rigor as that of training dataset | - Gives a detailed description of the steps in developing an algorithm for this application.  - Gives various ways for validation (i.e. k-fold, monte-carlo etc) |
| STARD - AI | Diagnostic accuracy studies assessing AI interventions | Steering Group. Not yet published. |  |  |  |
| MI - CLAIM | Clinical artificial intelligence modelling | Not reported |  | - Separation of data into partitions for model training and model testing.  - Optimization and final model selection  - Performance evaluation  - Model Examination  - Reproduceable pipeline | * Defining principle is: sharing of raw data is neither ethical or commercially prudent. So we have to assess clinical impact and allow rapid replication of the technical design process of any clinical AI study. |

## Computer Science Conference Topics

|  |  |
| --- | --- |
| [IJCAI 2020](https://www.ijcai.org/) | [NeurIPS 2019](https://nips.cc/) |
| * Agent-based and Multi-agent Systems * AI Ethics * Computer Vision * Constraints and SAT * Data Mining * Heuristic Search and Game Playing * Humans and AI * Knowledge Representation and Reasoning * Machine Learning * Machine Learning Applications * Multidisciplinary Topics and Applications * Natural Language Processing * Planning and Scheduling * Robotics * Uncertainty in AI * Special track on AI for CompSust and Human well-being * Special Track on AI in FinTech | * Active Learning * Adaptive Data Analysis * Adversarial Learning * Activity and Event Recognition * Audio and Speech Processing * Body Pose, Face and Gesture Analysis * Communication-or-Memory Bounded Learning * Computational Biology and Bioinformatics * Computational Photography * Computational Social Science * Computer Vision * Bandit Algorithms * Classification * Clustering * Collaborative Filtering * Density Estimation * Large Margin Methods * Large Scale Learning * Meta-Learning * Metric Learning * Missing Data * Model Selection and Structure Learning * Nonlinear Dimensionality Reduction and Manifold Learning * Online Learning * Ranking and Preference Learning * Regression Applications |

## HTAs of AI

| **Report** | **Purpose** | **Methods** | **Definitions** | **Additional information** |
| --- | --- | --- | --- | --- |
| United States Government Accountability Office - Artificial Intelligence in Health Care | To discuss 1) current and emerging AI technologies available for drug development (2) challenges to development and adoption (3) policy options to address challenges to the use of machine learning in drug development. | Discussions between expert working group comprised of leaders from various disciplines—public health, informatics,  biomedical ethics, and implementation science | **AI** – Oxford dictionary definition: “The capacity of computers or other machines to  exhibit or simulate intelligent behavior; the field of study concerned with this”  Summary of AI domains:    **Machine learning** is a family of statistical and mathematical modelling techniques that uses a variety of approaches to automatically learn and improve the prediction of a target state, without explicit programming (Witten et al., 2016).  **Natural language processing (NLP)** enables computers to understand and organize human languages (Manning and Schütze, 1999).  **Expert systems** are a set of computer algorithms that seek to emulate the decision-making capacity of human experts (Feigenbaum, 1992; Jackson, 1998; Leondes, 2002; Shortliffe and Buchanan, 1975). | High impact AI tools with immediate clinical implications require more stringent explanation requirements than low risk tools with proven accuracy that clearly conveys recommendations to the end user.  **Best Practices for Machine-Learning Model Development and Validation:**  Datasets used to train AI are heterogeneous, complex, and nuanced in ways that are often subtle and institution specific.  AI systems should be rigorously evaluated before deployment to ensure their competency and safety, in a similar process to drugs and medical devices.  Establishing Utility: With AI in health care, it is necessary to know how a member of the care team would act, given AI output. While AI model evaluation typically focuses on metrics, such as positive predictive value, sensitivity (or recall), specificity, and calibration, constraints on the action triggered by the model’s output (e.g. continuous rhythm monitoring might be constrained by availability of Holter monitors) often can have a much larger influence in determining model utility (Moons et al., 2012).  External validity depends on the aim of the AI model, the degree of agency ascribed to the model, and the nature of the action triggered by the model.  Biased data will result in biased models.  To the clinical user, interpretability could mean one of two things: a sufficient enough understanding of what is going on, so that they can trust the output and/or be able to get liability insurance for its recommendations; or enough causality in the model structure to provide hints as to what mitigating action to take.  One systematic review suggested that the AI algorithms could significantly improve with the use of: multicentre datasets, incorporation of time varying data, assessment of missing data as well as informative censoring, and development of metrics of clinical utility (Goldstein et al., 2017). As a reasonable starting point for minimizing data quality issues, the authors of the NAM Special Publication recommend that data should adhere to the FAIR (findability, accessibility, interoperability, and reusability) principles in order to maximize the value of data (Wilkinson et al., 2016).  **Deploying AI in clinical settings**  For higher risk AI tools, a focus on clinical safety and effectiveness—from either a noninferiority or superiority perspective—is of paramount importance even as other metrics are considered. High-risk tools will likely require evidence from rigorous studies for regulatory purposes and will certainly require substantial monitoring at the time of and following implementation. For low-risk clinical AI tools used at point of care, or those that focus on administrative tasks, evaluation may rightly focus on process of care measures and metrics related to the AI’s usage in practice to define its positive and negative effects.  In some instances, due to feasibility, costs, time constraints or other limitations, a randomized trial may not be practical or feasible. In these circumstances quasi-experimental approaches such as stepped-wedge designs or even carefully adjusted retrospective cohort studies, may provide valuable insights. |
| Gerke et al. 2020 | Discussion paper | The authors argue that regulators like the FDA need to widen their scope  from evaluating medical AI/ML-based products to assessing systems. This shifts the perspective from a product view to a system view. | “Software as a Medical Device” (SaMD) is used to refer to software that is on its own a medical device, “without being part of a hardware medical device. | Many patients and physicians are particularly  concerned about such “autonomous” devices. The current AI/ ML-based software as a medical device that received marketing authorization by the FDA have what the FDA has called “locked” algorithms—they do not evolve over time and do not use new data to alter their performance. If the algorithm changes through usage, such SaMD will, at present, likely require another FDA round of review.  For these reasons, the most valuable asset of AI/ML, its ability to improve by learning from data, may not be fully harnessed.  Due to their systemic aspects, AI/ML-based SaMD will present more variance between performance in the artificial testing environment and in actual practice settings, and thus potentially more risks and less certainty over their benefits.  AI/ML-based SaMD also differ from other medical technologies, such as the da Vinci surgical system, because (1) they have the capacity to continuously learn, (2) they have the potential to become ubiquitous in medical interactions and make recommendations (unlike robotic-assisted surgical systems), and (3) the way they reach their recommendations is often opaque to physicians |
| HTA AI AETSA (Spanish language publication. English abstract only. | Literature review to identify AI applications in medicine. | Search of MEDLINE database as an approach to the scope of AI in medical care. | None | Conclusion of the review: No systematic implementation of AI in medical care was identified. The literature on AI in clinical care and healthcare management shows different frequencies in the distribution of the different forms of AI considered. In clinical care, diagnosis is the most represented area. The specialties most covered are oncology, neurology and cardiovascular diseases. |
| Canadian Agency for Drugs and Technologies in Health (CADTH) RAPID RESPONSE REPORT (2019): SUMMARY OF ABSTRACTS  Artificial Intelligence for  Classification of Lung Nodules: Clinical Utility,  Diagnostic Accuracy, Cost-effectiveness, and  Guidelines | Research Questions – investigating the clinical utility, diagnostic accuracy cost-effectiveness and evidence-based guidelines of AI for nodule classification in lung cancer. | A limited literature search was conducted by an information specialist on key resources including MEDLINE, the Cochrane Library, the University of York Centre for Reviews and  Dissemination (CRD) databases | None | Two RCTs and twelve non-randomised studies were identified regarding the clinical utility and diagnostic accuracy of artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer. |
| CADTH RAPID RESPONSE REPORT (2019): SUMMARY WITH CRITICAL APPRAISAL Artificial Intelligence for Classification of Lung Nodules: A Review of Clinical Utility, Diagnostic Accuracy, Cost-Effectiveness, and Guidelines | To summarize  and critically appraise the evidence initially identified in the Summary of Abstracts, based  on additional screening and review of the full text (above) | Same as above | None | Seven diagnostic case-control studies were identified regarding the diagnostic accuracy of artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer. No evidence regarding the cost-effectiveness,  clinical utility or evidence-based guidelines regarding artificial intelligence for nodule classification in screening, incidental identification, or known or suspected malignancies for lung cancer were identified. |
| Wolff et al. 2020 | Systematically review and summarize the cost-effectiveness studies dedicated to AI in health  care and to assess whether they meet the established quality criteria | Systematic literature review | None | Very few publications have thoroughly addressed the economic impact assessment, and the economic assessment quality of the reviewed publications on AI shows severe methodological deficits. No study comprised a methodologically complete cost impact analysis  There are two areas for improvement in future studies. First, the initial investment and operational costs for the AI infrastructure and service need to be included. Second, alternatives to achieve similar impact must be evaluated to provide a comprehensive comparison.  The literature review revealed significant success factors for AI, for example, regarding the legal framework, such as compliance with data security, protection, and privacy policies, and also universally accepted technological requirements to enable comprehensive data collection and to analyze content while complying with data privacy requirements. Despite the benefits in assisting diagnostic and therapeutic decisions, so far, no standards for these legal and technological issues have been defined, and these aspects should be analysed in future research with a broader focus. |

## Regulatory frameworks

|  |  |  |
| --- | --- | --- |
| **Body** | **Document Information** | **Relevant Content** |
| USA | Food and Drug Administration (FDA)  Discussion Paper and Request for Feedback  Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback  April 2, 2019  **[Adapting IMDRF]** | AI/ML-based SaMD exists on a spectrum from locked to continuously adaptive algorithms, a common set of considerations for data management, re-training, and performance evaluation can be applied to the spectrum of SaMD. FDA has approved some 'Locked Algorithm' SaMD.  The current proposal is based on:   * Relies on IMDRF’s risk categorization principles * The FDA’s benefit-risk framework * Risk management principles described in the software modifications guidance * Organization-based TPLC (also envisioned in Digital Health Software Precertification Program) * Practices from FDA's current premarket programs, including the 510(k), De Novo, and PMA pathways   This proposed TPLC approach is to assure that ongoing algorithm changes are:   * implemented according to pre-specified performance objectives * follow defined algorithm change protocols * utilize a validation process committed to improving performance, safety, and effectiveness and * include real-world monitoring of performance   Based on the 510(k) software modifications guidance, categories of software modifications may be considered as part of a software change to an existing device guidance. The changes that may require a premarket submission include:   * A change that introduces a new risk or modifies an existing risk that could result in significant harm * A change to risk controls to prevent significant harm and * A change that significantly affects clinical functionality or performance specifications of the device   For AI/ML-based SaMD, the above approach would require a premarket submission to FDA when the modification:   * significantly affects the device performance, safety and/or effectiveness * is to the device’s intended use or * introduces a major change to the SaMD algorithm   The types of modifications generally fall into three broad categories:   * Clinical and analytical **performance** * **Inputs** used by the algorithm and their clinical association to the SaMD output * The **intended use** of the SaMD, as outlined above and in the IMDRF risk categorization framework, described through the significance of information provided by the SaMD for the state of the healthcare situation or condition.   **Total Product Lifecycle Regulatory Approach for AI/ML-Based SaMD**  TPLC approach is based on following general principles to balance benefits and risks, and to provide access to safe and effective AI/ML-based SaMD:   1. Establish clear expectations on quality systems and Good ML Practices (GMLP); 2. Conduct premarket review for those SaMD that require premarket submission to demonstrate reasonable assurance of safety and effectiveness and establish clear expectations for manufacturers of AI/ML-based SaMD to continually manage patient risks throughout the lifecycle;    1. Predetermined Change Control Plan would include types of anticipated modifications:       1. **SaMD Pre-Specifications (SPS)** – based on the retraining and model update strategy, and the associated methodology. SaMD manufacturer’s anticipated modifications to “performance” or “inputs,” or changes related to the “intended use” of AI/ML-based SaMD. These are the types of changes the manufacturer plans to achieve when the SaMD is in use. The SPS draws a “region of potential changes” around the initial specifications and labelling of the original device. This is "what" the manufacturer intends the algorithm to become as it learns.       2. **Algorithm Change Protocol (ACP)** – being used to implement those changes in a controlled manner that manages risks to patients. Specific methods that a manufacturer has in place to achieve and appropriately control the risks of the anticipated types of modifications delineated in the SPS. The ACP is a step-by-step delineation of the data and procedures to be followed so that the modification achieves its goals and the device remains safe and effective after the modification:          1. Data Management;          2. Re-training;          3. Performance Evaluation; and          4. Update Procedure  * FDA also lists Scope and limitations for establishing SPS and ACP.  1. Expect manufacturers to monitor the AI/ML device and incorporate a risk management approach and other approaches outlined in “Deciding When to Submit a 510(k) for a Software Change to an Existing Device” Guidance18 in development, validation, and execution of the algorithm changes (SaMD Pre-Specifications and Algorithm Change Protocol); and 2. Enable increased transparency to users and FDA using postmarket real-world performance reporting for maintaining continued assurance of safety and effectiveness.   An independent report (Benjamens et al. 2020) shows that FDA has approved 64 AI-ML based medical devices and algorithms among which 85.9% was approved with a 510(k), 12.5% through de novo pathway and one (1.6%) with premarket approval clearance.  A recent report from Journal of American College of Radiology on 20 October 2020 lists critical points and suggestions regarding international frameworks including FDA's proposed framework. (See Larson et al. 2020 in the Table).  Another report from an official journal of The Association of University Radiologists (Harvey et al. 2010) highlights that   * The FDA has made its intention clear to regulate as devices software products used in the interpretation of diagnostic imaging…However, this should not be taken to exclude AI products employed in workload reduction or logistical capacities within a radiology context. These would presumably bypass device designation [per 3060(a)](https://www.fda.gov/media/109622/download) and find their way into the market without requiring FDA approval. * FDA has additionally sought to draw a distinction between Computer-Assisted Detection Devices (CADe) as opposed to the more stringently regulated computer-assisted diagnosis devices; the former are intended to “identify, mark, highlight, or in any other manner direct attention” to imaging features, rather than autonomously diagnose, stage, or triage pathology (23). Indeed, a number of imaging AI products offering CADe functions have received approval as CDS…In a move to further ease the regulatory onus on SaMD developers, the FDA is considering reclassifying CADe used in the visualization of breast lesions, lung nodules, and dental caries to a lower-risk category, to require 510(k) submission rather than premarket approval. |
| USA | Food and Drug Administration (FDA)  Draft Guidance  Clinical Decision Support Software: Draft Guidance for Industry and Food and Drug Administration Staff  September 27, 2019  **[Adapting IMDRF]** | Targets Clinical Decision Support (CDS) Software for Industry and FDA Staff  **Examples of Device CDS intended for Health Care Professionals (HCPs) for which, based on our current understanding of the risks of these devices:**  **FDA does not intend at this time to enforce compliance with applicable device requirements**:   * ML algorithm, for which the logic and inputs are not explained, that trends and classifies patient-specific data to alert HCPs to potential triggers that may be indicative of cholesterol management issues. It is an aggregation of data intended to provide clinical information for a non-serious situation or condition.   **Device CDS on which FDA intends to focus its regulatory oversight [they intend to inform clinical management for a serious or critical situation or condition]**  Note: If the HCP could evaluate the basis for the software’s recommendations, because the logic and inputs for the ML algorithm and data inputs/criteria for risk of events used for the algorithm were explained and available to the HCP, then this software would be considered Non-Device CDS.   * ML algorithm, for which the logic and inputs are not explained, that categorizes likely symptoms of seasonal influenza for each flu season based on location and current electronic health records of patients diagnosed or suspected to have influenza to assist HCPs in differentiating between common flu symptoms and other illnesses (e.g., common cold) in a particular season. It is intended to inform clinical management for a serious situation or condition. * ML algorithm, for which the logic and inputs are not explained, that identifies hospitalized, type 1 diabetic patients at increased risk of postoperative cardiovascular events. It is intended to inform clinical management for a critical situation or condition. |
| International | International Medical Device Regulators Forum (IMDRF)  [Presentation at IMDRF]  Artificial Intelligence Medical Devices (AIMD) Working Group Update  September 21-25, 2020 | From Artificial Intelligence Medical Devices (AIMD) Working Group intends to   * Achieving a harmonised approach to the management of AI medical devices * Establishing a guidance (draft release in July-October 2021) to share the views of member jurisdictions on terminology; main contents of the guidance   + - * Definitions and scope of medical devices       * Standardisation of terminology |
| EU | European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR)  Legislation Review  Artificial Intelligence in EU Medical Device Legislation  September, 2020  **[Adapting IMDRF]** | Targets AI in EU Medical Device Legislation  In their white paper published in 2019 (COCIR 2019), they highlight that to a large extent AI may be simply considered as a specific type of software, and in that respect current classification models established by the Medical Device Regulation (MDR) and IVD Regulation (IVDR) are applicable and suitable. The general safety and performance requirements are generic principles which do not necessarily require adaptation to a new technology. Hence existing regulations should be applied. The also recommend: There is a need to clarify how existing legislative frameworks can be made more inclusive so that all forms and applications of Artificial Intelligence in Healthcare benefit from the same legal clarity and certainty. A more thorough evaluation and an evolution of the understanding of industry’s responsibility should be required prior to assessing any new or additional policy options. This next step in AI will not be possible without appropriately addressing the following non-exhaustive list of questions:   * How would one determine and manage a change, either significant or non-significant to an AI system? * Can we distinguish (non-)significance of changes to the system in terms of the performance of the AI system, particularly if this would result in different decisions made by the AI? * How would changes to AI systems affect regulatory obligations related to for instance labelling and registration? * Who would be considered the manufacturer and to what extent can the manufacturer exercise control?   In September 2020, COCIR released Legislation Review which is an analysis of the legal requirements that concludes AI can be deployed in a way that is consistent with EU MDR and EU IVDR.  To facilitate the practical implementation of the legal requirements COCIR recommends the adoption of practical guidance, supported by the development of international standards and makes concrete proposals for such guidance throughout this analysis, and recommendations. Here are only the COCIR recommendations related to regulating AI:  COCIR recommends that manufacturers of an AI-based system that changes through learning during runtime add a pre-determined Algorithm Change Protocol to the technical documentation of their device for evaluation during conformity assessment. Manufacturers that make significant changes to the pre-determined Algorithm Change Protocol need to update the technical documentation, including the clinical evaluation. They need to perform a new conformity assessment. Changes that are in scope of the Algorithm Change Protocol are not significant changes. Changes that go beyond the change of the Algorithm Change Protocol can be significant.  COCIR recommends that manufacturers of an AI-based system that changes through learning to consider a design capable of storing discrete states of a learned model and capable of returning to a previously stored state in order to reproduce results.  COCIR recommends citing in the Official Journal of the European Union relevant AI standards for EU MDR and EU IVDR as they become available. This will further the common understanding, the transparency, and the trust between stakeholders. As a start, COCIR recommends harmonizing:   * [ISO 14971:2019 Medical devices](https://www.iso.org/standard/72704.html): application of risk management to medical devices to give manufacturers a clear signal that also mental risks [to people] must be controlled * IEC 62366-1:2015 Medical devices – Part 1: Application of usability engineering to medical devices to give manufacturers a clear signal to identify and control inappropriate levels of user trust   COCIR recommends updating two standards commonly used in healthcare in view of artificial intelligence:   * IEC 62304:2006/Amdt 1:2015 Medical device software—Software lifecycle processes   + Requiring among others that manufacturers define an Algorithm Change Protocol (ACP) for AI that changes through learning during runtime * IEC 82304-1:2016 Health Software – Part 1: General requirements for product safety   + by requiring manufacturers to describe AI attributes in user-facing documentation:   1. If, for what aspects and how the AI provides human oversight or control  2. If, for what aspects and how the AI changes   * Define if AI is locked or changes through learning during runtime * Provide a description of change dynamics and change boundaries * Define if and how humans can control change   + by requiring manufacturers to inform user on how to adjust the “knob” when fairness/accuracy trade-offs exist and can be adjusted by the user   + by requiring manufacturers to add to the limitations or contraindications to the use any minority group that can potentially be subject to a risk of significant bias   + by requiring manufacturers to inform the user of the performance characteristics of their AI-based device needed to discern the degree to which the software output and performance under the intended use conditions produces identical results, for example by communicating its precision, sensitivity, specificity, confidence, operating parameters, … |
| EU | Heads of Medicines Agencies (HMA) and European Medicines Agency (EMA)  Report  HMA-EMA Joint Big Data Taskforce Phase II report: ‘Evolving Data-Driven Regulation  January, 2020 | The report presents 10 priority recommendations which are fully compatible with the current EU legal framework for the regulation of medicinal products:   1. Deliver a sustainable platform to access and analyse healthcare data from across the EU (Data Analysis and Real-World Interrogation Network -DARWIN). 2. Establish an EU framework for data quality and representativeness 3. Enable data discoverability 4. Develop EU network skills in Big Data 5. Strengthen EU network processes for Big Data submissions 6. Build EU network capability to analyse Big Data 7. Modernise the delivery of expert advice 8. Ensure data are managed and analysed within a secure and ethical governance framework 9. Collaborate with international initiatives on Big Data 10. Create an EU Big Data ‘stakeholder implementation forum’   Under Regulatory considerations on Bioinformatics, Algorithms, Machine Learning and Artificial intelligence (AI), four immediate areas of AI use are important to address in a regulatory context:   * Ensure sufficient expertise and capacities are available within the European network * Enable regulatory evaluation of clinical data submitted by drug manufactures for approval where the data has been processed by AI algorithms or part of the analysis, such as patient selection, involved AI methods. * Explore regulatory use of AI in internal processes * Approval of AI-based Health Apps in devices intended for clinical decision-making |
| Australia | Therapeutic Goods Administration (TGA)  Consultation: Regulation of software, including SaMD  February 13, 2019  Comments about the requirements for AI technologies and software services  November 8, 2019  Consultation: Scope of regulated software based medical devices  March 25, 2020  Comments that this have not included AI  August 17, 2020  **[Adapting IMDRF]** | The consultation proposes to use the factors identified by IMDRF for risk categorisation of software to classify SaMD products in Australia. This means that rules would be written that consider both the seriousness of the situation or condition where the SaMD is used, and the significance of the information being provided by the SaMD.  The authors also note that although the new EU MDR 2017/745 rule is in accordance with the IMDRF recommendations, it does not provide enough detail to capture the different risk categories of SaMD identified by the IMDRF. More detail is required in the Australian classification rule to close any gaps, which will promote clarity and consistency, and a more detail approach is proposed:  Software that processes data (e.g. - images, sensor data, big data) to provide information for diagnosing a disease or condition and that is intended to:   * Make a direct diagnosis (e.g. – self testing, emergency situation, rural or remote medicine) for:   + a critical situation where the disease or condition is fatal or debilitating in a short timeframe, or poses a risk to public health, or a serious situation where the disease or condition is not life threatening but may cause a serious deterioration in a person’s state of health if not identified. The device is Class III.   + any other situation. The device is Class IIa. * Screen patients to determine the need for further assessment for:   + a disease or condition that is fatal or debilitating in a short timeframe, or that poses a risk to public health. The device is Class III.   + a disease or condition that is not life threatening but may cause a serious deterioration in a person’s state of health if not identified. The device is Class IIb.   + any other situation. The device is Class IIa. * Aid a clinician in making a diagnosis. The device is Class IIa.   Software that processes data to provide information for treatment or intervention in a disease or condition and that is intended to:   * Specify a treatment or intervention that will be administered without further consideration (e.g. – the patient will inject the amount of insulin calculated) where:   + the treatment or intervention, or its absence, could result in death or debilitation. The device is Class III.   + the treatment or intervention, or its absence, could be harmful. The device is Class IIb.   + the treatment or intervention, or its absence, is unlikely to cause harm. The device is Class IIa. * Recommend a treatment or intervention for a clinician to decide and administer. The device is Class IIa.   Software that provides therapy through direct interaction with a patient where:   * The software directs patient activity based on input from the patient and could result is patient harm (e.g. – directing a recovering heart patient to undertake activity that is too vigorous). The device is Class IIb. * The software directs patient activity based on input from the patient and the activity is unlikely to cause harm. The device is Class IIa. * The software directs patient activity based on a non-interactive intervention. The device is Class I. |
| Japan | The Pharmaceutical Safety and Environmental Health Bureau (PSEHB) and Medical Device Evaluation Division (MDED), Pharmaceuticals and Medical Devices Agency (PMDA)  Guidance [English and Japanese]  Guidance for evaluation of artificial intelligence–assisted medical imaging systems for clinical diagnosis  May 23, 2019 | **Issues and Points to Consider**  **Black Box**: Algorithm for calculating output is “black box” nature in AI based on deep learning.   * Approval review process should focus on the performance evaluation by confirming if the input yields the required output. * Manufacturers should guarantee the performance by indicating that the systems always meet the specifications on performance. * Functions that inform any unexpected outputs of the system to the users should be also required.   **Changes in Performance**: Its performance, especially after post-market training, can only be evaluated by verification of the output.   * Continuous verification of performance * Quality assurance associated with performance changes * Principles on post post-market approval process   **Assigning Responsibility**: How to consider the source or type of the data, authenticity and bias in the learning data?  Post-Approval Change Management Protocol will be introduced for medical devices including with AI to make continuous improvement possible (September 16, 2019, Mr. Fumihito Takanashi's IMDRF Presentation). |
| Russia | Research and Practical Clinical Center For Diagnostics and Telemedicine Technologies, Moscow Health Care Department, Government of Moscow  Clinical Acceptance of Software Based on Artificial Intelligence Technologies (Radiology)  Best Practice and Methodological Guidelines  2019 | These methodological guidelines were publicly discussed in a preprint format (arXiv:1908.00381) and recommended by the European Society of Medical Imaging Informatics (EuSoMII).  Provide a methodological framework for the process of clinical tests, clinical acceptance, and scientific assessment of algorithms and software based on the artificial intelligence (AI) technologies. Clinical tests are considered as a preparation stage for the software registration as a medical product. The authors propose approaches to evaluate accuracy and efficiency of the AI algorithms for radiology.    When identifying the software class, please specify the categories applicable to the service  When classifying the software that is a medical device, only one class may be assigned to each software (Table 1):  Class 2a: Medium-low risk software  Class 2b: Medium-high risk software  Class 3: High-risk software    **INFORMATION VALUE**  ***I – Crucial information***  Information that is (a) crucial to make an informed clinical decision when making a diagnosis and/or providing treatment to a patient, and (b) used to take immediate and timely action:  – When treating, preventing, or alleviating disease manifestations through the use of medicines, medical devices, or other treatment methods;  – To detect diseases (i.e., for diagnosis or screening).  ***II – Information that requires clarification***  Information that requires clarification and/or details due to its insufficiency to make an informed clinical (medical) decision  – information on the safe and effective use of medicines and medical devices that is used in the treatment of disease  – Information used to predict the risk of disease development, as well as supporting information used to identify the signs and symptoms of the disease or make a preliminary or final diagnosis  – Classification or identification of early symptoms of the disease  ***III – Information intended to provide the long-term treatment***  Information that (a) is intended to provide the long-term treatment, (b) does not require immediate action, and (c) is meant to inform about diseases:  – Information on the options available to diagnose, treat, prevent, or alleviate disease manifestations;  – Information obtained by the software by collecting relevant data (e.g., data on patient diseases, used medicines or medical devices, etc.).  **CLINICAL SITUATION CATEGORIES**  **Category A**  The clinical situation is classified into Category A if software is intended for use:  – In case of emergency medical care;  – In severe, extremely severe, and terminal general condition of the patient;  – When determining the need for major therapeutic or surgical intervention;  – In the diagnosis or treatment of diseases that pose a high risk to public health and/or for high-risk patients (including for vulnerable population group).  In this situation, the software can only be used by specially trained healthcare professionals.  **Category B**  The clinical situation is classified into Category B if software is intended for use:  – In case of urgent medical care;  – In moderate general condition of the patient;  – If the disease or condition does not require major therapeutic intervention;  – In the diagnosis or treatment of diseases that pose a moderate risk to public health.  The software can be used both by specially trained healthcare professionals and by patient or other individual supervised by specially trained healthcare professionals. If the software is used by patient or other individual without supervision of specially trained healthcare professionals, this clinical situation is Classified into Category A.  **Category C**  The clinical situation is classified into Category B if software is intended for use:  – In case of routine medical care;  – In satisfactory general condition of the patient;  – If the disease requires minor therapeutic intervention (usually, non-invasive) or long-term medical supervision;  – In the diagnosis or treatment of diseases that pose a low risk to public health.  The software can be used both by specially trained healthcare professionals and by patient or other individual without supervision of specially trained healthcare professionals.  N.B.: If more than one provision may apply to the software, the software class is determined according to the highest potential risk. |
| Russia | Vladimir Kutichev, Head of Medical Device Software Department, Roszdravnadzor [Federal Service for Surveillance in Healthcare]  Regulatory Framework [Presentation at IMDRF]  Regulatory Framework on Medical Devices Using AI Technology in Russia  2019 | **AI Registration Challenges for Case**: Medical decision support system based on artificial intelligence   * No specific regulatory requirements for AI * The absence of national or international databases containing validated clinical information * Inaccurate medical data recorded in the medical old records on which the AI learns can produce incorrect results * Black box testing   **Prospects activities in the field of AI**   * Technical Committee (TC) 164 Artificial intelligence " was approved by Federal Agency for technical regulation and Metrology at the end of July 2019 * The technical Committee was established to improve the efficiency of the development of the national regulatory and technical base in the field of artificial intelligence * The first meeting of the Technical Committee was held on August 6, 2019 * Roszdravnadzor became as an official member of TC 164 * Working Groups of the Technical Committee 164:   + WG 01 Terms & Definitions   + WG 02 Big Data   + WG 03 Quality of Artificial Intelligence Systems   + WG 04 Applied Artificial Intelligence Technologies   + WG 05 Artificial Intelligence Technologies in Education   **National AI Standardization Program 2020**  **Objectives**  1) Development of classification criteria (types, classes of potential risk) of software using artificial intelligence/machine learning technologies  2) To formulate a clear terminology: what is artificial intelligence/machine learning, etc.  3) Development of proposals to national standards and other regulatory documents for software using artificial intelligence/machine learning technologies  4) Development of the criteria of responsibility -in which cases the doctor can rely on the data obtained from the software based on artificial intelligence/machine learning, whether it is entitled to use them for diagnosis, and whose opinion is more important  5) Development of proposals to the regulatory framework for the organization of the collection of unified verified clinical data to configure and verify the effectiveness of artificial intelligence systems  6) Development of approaches to regulation of software using artificial intelligence and machine learning technologies, including a transparent approach to regulation to confirm the quality, effectiveness and safety |
| China | National Medical Products Administration (NMPA)  Technical Guidelines (Consultation Draft)  Real-world data used in medical device clinical evaluation technical guidelines [Chinese Language]  December 13, 2019 | [Content is not directly relevant to AI] |
| China | National Medical Products Administration (NMPA)  Review  Key points of deep learning-assisted decision-making medical device software review  No Date | In Chinese Language [Content here is from Webinar Presentation: Regulating AI-Based Medical Devices; June 25, 2020; Organised by APACMed]  Approach for Modifications to AI Software   * Major Software Updates: Affects safety/efficacy of the software   + Algorithm-Driven: Changes in the algorithm, algorithm structure, algorithm flow   + Data-Driven: Statistically significant changes in algorithm performance * Minor Software Updates: Does not affect safety/efficacy of the software   + Data Driven: No statistically significant changes in algorithm performance   The authority is also building up its staff and has established an “AI Medical Device Standardization Unit”. This unit is responsible for the standardization of terminology, technology and processes for development and quality assurance ([Quoted from Johner Institute](https://johner-institute.com/articles/regulatory-affairs/and-more/regulatory-requirements-for-medical-devices-with-machine-learning/)). |
| USA | Xavier Health in Xavier University in partnership with FDA Officials and Industry professionals  Good Practice Report [Planned Output from 2017 AI Summit]  Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare  August, 2018 | Topics that should be considered when developing CLS for healthcare   * Human Factors * Cyber-Security and Privacy * Legal liability * Regulatory considerations   Continuously Learning Systems is sometimes referred to as incremental learning. A description of the features of incremental learning is provided in an article by Karanam Supraja8, and include:   * Accommodate new information as and when available * Ability to work with unlabelled data * Ability to handle multidimensional data * Bounded complexity (e.g. amount of complexity in a problem is limited) * Learn incrementally from empirical data, and * Handle changes in concepts etc.   **Considerations for Continuously Learning Systems (CLS)**  **Characteristics of a system that contribute to CLS efficacy and robustness**  Source of data – quality and quantity of data, including expected minimums, structure, and an understanding of the context of the data sets  Fairness - should treat everyone in a fair and balanced manner and not affect similarly situated groups of people in different ways  Number of variables, features, and layers being utilized in the model (e.g. Advanced Broad-Based Analytics)  Frequency of training or retraining  Known limitations and exceptions  Established parameters of operations  **After the system has been designed, implemented, and trained, additional steps are needed to ensure its quality (both actual quality and perceived quality):**   * Performance Evaluation – Overall system test methods and tools, typically involves a comparison of what the AI is supposed to do for a certain input versus what it does * Building Confidence in AI – Includes processes that are used to provide confidence in AI operation in addition to or complementary to performance evaluation. It should be noted that a correct and robust application is useless if no one actually uses it. Building confidence in the AI system is an important success factor, and it includes:   + Inclusiveness   + Reliability   + Usability   + Clinical significance   + Ease of integration with existing systems   + Precision and accuracy   It is worth noting that any effort to standardize on a set of specific technologies, techniques, algorithms, models or toolsets is likely going to be obsolete in a short period of time given the rapid evolution within AI. This is particularly relevant because we are at a very nascent stage of the evolution of this technology and its applications.  **Data Sources, Feedback Loop, Quality and Confidence Assessment, and Limitations**  **General Software Design**  **Confidence and Explainability**  **Integrating Values & Ethics**  **Patient Consent**  **Retraining**  **Risk Management: Development Risks and Failure Risks**  Both risk areas are tied to two critical elements to determine Risk Level:   1. Risk to the patient and/or critical quality attributes 2. Level of autonomy of the system   **Cyber-Security, Authentication, Privacy, and Anonymization**  **Requirements of CLS Systems**   * Quality Assurance * Quality Improvement * Confidence, Explainability, and Trust   **Verification and Validation Challenges**   * Data Governance * Transparency/Explainability * Data Paradigms |
| EU | Joint Research Centre (JRC), the European Commission’s Science and Knowledge Service  Conference and Workshop Report  Legal and regulatory implications of AI: The case of autonomous vehicles, m-health and data mining  2019 | The followings are specific for regulatory implications of m-health:   * Refers to Regulation (EU) 2017/745 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. * This definition might be applicable to most of the AI applications used in m-health, mainly when dealing with AI incorporated in software, which is often the case46. However, the producer may have difficulties in knowing if the application is a medical device or not (as opposed to an accessory to a medical device). If the AI application is included in the definition of medical devices in the 2017/745 Regulation, companies have to comply with CE marking, information duties, etc. However, this seems to be largely unknown by producers. Besides this lack of knowledge, the cost for the administrative steps in obtaining the CE marking may in some instances be an issue. * Some actors on the field raise the issue related to the fact that Health related mobile apps are available on app stores without control on the quality. This might lead to insecurity (e.g. break of confidentiality, safety obligations, etc.) and lack of trust in product. To avoid this, an app store dedicated to health related apps might be accredited so to create a kind of label. |
| Canada | Standing Senate Committee on Social Affairs, Science and Technology  Report  Challenge ahead: integrating robotics, artificial intelligence and 3D printing technologies into Canada’s healthcare systems  October, 2017  **[Adapting IMDRF and FDA]** | **Health Canada**  Health Canada has indicated that Canada’s Food and Drug Act and Medical Devices Regulations have already informed issuing of licenses to technologies that use AI.  In April 2018, [Health Canada announced](https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/activities/announcements/notice-digital-health-technologies.html) the establishment of the Digital Health Review Division within the Therapeutic Products Directorate’s Medical Devices Bureau. The division is intended to allow for a more targeted review of rapidly changing and innovating digital health technologies, including AI technologies.  **Standing Senate Committee on Social Affairs, Science and Technology**  Health Canada indicated that under the Medical Devices Regulations (the regulations) the department has approved and issued licences for innovations in all three categories; robotics, AI and 3D printing. It suggested that the current regulatory framework is appropriate for responding to evolving technologies. However, some witnesses suggested that some changes might help to make the regulatory framework more responsive.  The Committee makes 14 Recommendations to Government and Government Responses including under Regulatory oversight that Canada works with US FDA and IMDRF.  The Committee mentions that some witnesses suggested that some changes might help to make the regulatory framework more responsive. For example, under the regulations, in order to receive marketing approval from the department, the device sponsor must provide sufficient evidence from investigational studies of the device’s safety and efficacy. Members were told that in the case of some of the innovative technologies, traditional randomized, controlled trials may not be the most appropriate approach and that the regulator should allow alternative approaches to determine safety and efficacy. Members were told that the European Parliament is currently working on a regulatory approach specifically for robotics. |
| South Korea | Ministry of Food and Drug Safety  [Presentation at IMDRF]  Perspectives and Regulatory Considerations for AI based Medical Devices  September, 2019 | Deregulation of AI-based medical device for rapid market entry   * Approval/review guidelines with the application of big data and AI technology (published on 22 Nov. 2017) * Guidelines for evaluation of the clinical efficacy of the AI-based medical device (published on 20 Dec. 2017) * Enactment of Medical device industry development and innovation medical device support act (Enactment on 30 Apr. 2019, Effective from 1 May, 2020)   **Guideline on Approval Review of AI and Big Data based Medical Devices (Nov 2017)**  **Classification**  According to the Regulations on Medical Device Items and Classification by Item  - The main items are classified into medical imaging analysis device software [2], Computer aided detection software [2], Computer aided diagnosis software [3], and Radiation treatment planning software [3];  - This pertains to four items, including in vitro diagnostic software items such as disease prognosis and predictive screening software [2], and cancer prognosis and predictive screening software [3].  **Approval Review Measures**   1. ‘Performance’ recording method 2. Performance and clinical efficacy validation items 3. Clinical efficacy validation 4. Scope of data submission 5. Subject for amendment approval and certification: If the accuracy is improved by adding learning data without changing the designs, procedures shall be exempted. However, it must be managed under the GMP system. 6. Version management 7. Management of learning data 8. Approval scope for applying cloud computing technology   **Challenge and Future Directions**  **Revision of AI-based Medical Device Related Regulations and Guidelines**  When conducting the approval process of an AI-based medical device (independent software type), challenges arise during the review due to a gap between the company and the reviewer as there is no standard for a comparison method.  Flexible regulations are required to respond reasonably to the development of innovative medical technology which is rapidly growing.  **Equivalence Review Status**: Comparison of equivalence with the predicted product  *For general medical device*: When comparing with the product, it is considered to be equivalent if it falls within the category described in the purpose of use, performance, test standard, and method of use.  *For independent software (AI medical device)*: When comparing with the product, it is difficult to determine what level would be considered as equivalent in algorithm, coding, etc.  **Medical Devices Industry Development and Innovative Medical Device Support Act (Effective from May 1, 2020)**  Considerations for Certification of Innovative Device Companies   * Securing the research personnel for the medical device, and excellence of human and material input resources for research and development * Excellence in medical device research and development activities * Corporate social responsibility and ethics * Contribution to the technical and economic excellence of medical device research and development performance and improvement of public health   Re-evaluation every 3 years after the certification |
| South Korea | Ministry of Food and Drug Safety, Medical Device Evaluation Department, (Digital Health Devices Task Force) High-tech Medical Devices Division  Guideline  Guideline on Review and Approval of Artificial Intelligence (AI) and Big Data Based Medical Devices (For Industry)  November, 2020 | This guideline describes the position of the Ministry of Food and Drug Safety regarding classification criteria, classification and review and approval process for big data and AI-based medical devices.  Medical devices are classified based on their intended use, characteristics and potential risk to human body upon use in accordance with Article 3, Medical Devices Act (Classification and Designation of Classes).  **AI classification**   * The guideline classifies AI technologies based on the type of decision-making power they have. Automatic diagnosis, prediction, monitoring, or treatment is considered a “Sole Determinant” and requires more clinical evidence than “Informing” or “Driving” technologies. The latter technologies are only designed to assist healthcare practitioners and their data analysis, insight and recommendations, need to be confirmed through additional tests and/or professional input.   **Classifications of Change in Version**   * Major change: Change in operating principles, intended use, performance (applicable to In case where the pre-approved performance (accuracy) changes beyond the previous range.) * Simple change: Graphic user interface (GUI) design change * Minor change: Bug correction, colour and menu location change of GUI, etc. * Training data change: Training data change within the range of performance (accuracy) written during approval.   The following two tables are from a [secondary source](https://asiaactual.com/blog/south-korea-software-guidance-big-data-ai-machine-learning/) not directly from the Guidance but restructures the guidance with examples:  Types of Software Outputs   |  |  |  | | --- | --- | --- | | **Sole determinant** | **Informing** | **Driving** | | Analyse vital signs measured and integrated in the emergency room to predict emergency situations such as breathing difficulties that gives warnings such as alarms. | Software that calculates the probability of a specific cancer based on medical information such as a biopsy and electronic medical record (EMR).  Screening software that detects and displays abnormal areas through upper CT image analysis.  Analyses medical images to determine quantitative values for specific areas of blood vessels such as blood flow rate and vessel diameter. | Software that predicts hypoglycaemia in advance by analysing information such as blood sugar data, food intake, and insulin injection.  Software that diagnoses or predicts arrhythmia using ECG measurement results. |   The MFDS has initially identified the following 7 categories of software used for Image analysis that uses Big Data, AI and/or Machine Learning:   |  |  | | --- | --- | | **Item Name** | **Notes** | | Medical Image Analysis Software | Software that can be used for simulation treatment, simulation procedure, and diagnosis by acquiring and analysing medical images. | | Radiation Treatment Software | Used to determine radiation simulation treatment and simulation using acquired medical images. | | Medical Image Detection Aid Software | Assist medical personnel in making better diagnoses by marking them with colour or leader lines | | Medical Image Analysis Device | Device with software that acquires and analyses medical images and can be used for simulation treatment, simulation procedures, and diagnosis. | | Medical Image Detection Aid Device | After detecting abnormal areas in a medical image. The device assists medical personnel in making diagnoses by marking them with colour or leader lines | | Medical Imaging Assistance Software | Software used to aid in diagnostic decisions. Uses medical images to determine the presence or absence of a  disease, the severity of the disease, or the degree of likelihood of the condition, etc. | | Medical Imaging Assistance Deice | Medical imaging is used to determine the presence or absence of a disease, the severity of the disease, or the degree of likelihood of the condition, etc. is automatically displayed to a device used to aid in making diagnoses. | |
| Singapore | Health Sciences Authority  Regulatory Guidance  GN-21: Guidance on Change Notification for Registered Medical Devices (Revision 4.7)  February, 2020 | Regulatory Guidelines for Software Medical Devices – A Life Cycle Approach (April 2020)  Both guidances provide comprehensive details and decision making algorithms to account for version changes. The diagram below shows what happens in cases of Continuous Learning Algorithms: |

## Technographic review

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [Olthof et al. 2020](https://pubmed.ncbi.nlm.nih.gov/32318774/) | 1.Obtain a comprehensive, systematic overview of AI functionalities for neuroradiology and 2. Analyse the potential impacts of these applications on the work of neuroradiologists. | SR | Type of impact on clinical work | Identified 37 applications of 27 companies. All applications analyse images of one or more of the following imaging modalities: MRI (19; 51%), CT (19; 51%), MR perfusion (2; 5%), CT perfusion (3; 8%), CT angiography (5; 14%) and MR angiography (1; 3%).  Most applications are designed to be used for one pathology.  Most functionalities of applications (39; 54%) are designed to ‘support’ radiologists in performing their current tasks. Some other functionalities of applications ‘extend’ the work of radiologists by providing quantitative information that would not be possible before the introduction of these applications (23; 32%). Only a few functionalities of applications (10; 14%) offer functionalities that take over certain tasks.  An application sometimes offers functionalities related to all three categories. For instance, the AI reader (Qmenta) supports the radiologist in making a diagnosis, extends radiology work by providing quantitative information and can replace the radiologist in drafting an automated report for the referring physician. |

## Classification example

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Technology | JW classification of complexity | | | Autonomy | | | Fixed/Continuous | | Explainability/Supervision | | | Interaction | | Clinical Pathway | | | Functionality (Olthof 2020) |
|  | High | Middle | Low | Supervised | Semi-supervised | Un-supervised | Fixed | Continuous | No black-box | Some info | Complete black-box | NO | YES | Narrow Low | Narrow High | Broad High | 7 categories\* |
| Zio | High |  |  |  | Semi-supervised |  | Fixed |  |  | Some info |  | NO |  |  |  | Broad High | 1, 2, 3, 4, 6 |
| Mia | High |  |  |  | Semi-supervised |  | Fixed |  |  | Some info |  |  | YES |  | Narrow High |  | 1, 2, 3, 4, 6 |
| Veye | High |  |  |  | Semi-supervised |  | Fixed |  |  | Some info |  |  | YES |  |  | Broad High | 1, 2, 3, 4, 6 |
| DrDoctor | High |  |  |  | Semi-supervised |  | Fixed |  |  |  | Complete black-box |  | YES |  |  | Broad High | 3, 6 |
| DLCExpert | High |  |  |  | Semi-supervised |  | Fixed |  |  | Some info |  |  | YES |  |  | Broad High | 1, 2, 6, 7 |
| e-Stroke Suite | High |  |  |  | Semi-supervised |  | Fixed |  |  | Some info |  | NO |  |  | Narrow High |  | 1, 2, 3, 4, 5, 6, 7 |
| RITA | High |  |  |  | Semi-supervised |  | Fixed |  |  |  | Complete black-box |  | YES |  | Narrow High |  | 3, 6 |
| Healthy.io |  | Middle |  |  | Semi-supervised |  | Fixed |  |  | Some info |  |  | YES | Narrow  Low |  |  | 1, 3, 4, 6 |
| OptosAI | High |  |  |  | Semi-supervised |  | Fixed |  |  | Some info |  | NO |  |  | Narrow High |  | 1, 2, 3, 4 |
| Ultromics | High |  |  |  | Semi-supervised |  | Fixed |  |  | Some info |  | NO |  |  | Narrow High |  | 1, 2, 3, 4 |
| \*1: Provides quantitative information about pathology, 2: Marks regions of interest or detects change, 3: Provides classification, diagnosis or outcome probabilities, 4: Prepares report, 5: Automated derivation of brain biomarkers, 6: Workflow optimization and triaging, 7: Anatomical segmentation | | | | | | | | | | | | | | | | | |

1. Those classifications following a systematic and widely agreeable approach. [↑](#footnote-ref-1)
2. Those classifications based on personal preferences. [↑](#footnote-ref-2)
3. This finding from patent filling corroborates expert opinions on the dominance of ML in the field of healthcare. [↑](#footnote-ref-3)
4. This definition has also been adopted by the CONSORT-AI extension. [↑](#footnote-ref-4)
5. Algorithms that adapt to new circumstances and can detect and extrapolate patterns. [↑](#footnote-ref-5)
6. Such as IJCAI and NeurIPS [↑](#footnote-ref-6)
7. Such as those reported in section 6.1.2. [↑](#footnote-ref-7)
8. This is likely for two reasons; firstly, algorithm types are poorly defined and numerous and secondly because new algorithm types are always in development, such that an open-ended definition is more “future-proof”. This approach should be taken into consideration when developing a classification. [↑](#footnote-ref-8)
9. An example of a highly curated dataset would be to exclude low quality images. [↑](#footnote-ref-9)
10. It is important to note that the literature, such as the Park and Han (2018) paper, highlights the advantages of prospectively collected datasets. [↑](#footnote-ref-10)
11. We only had access to limited publicly available information for retrieving related information for each category for each technology. [↑](#footnote-ref-11)