Imperial College London

Regulating AI in Healthcare products

1. Key organisations in the UK

In the UK there are a range of public sector bodies providing guidance or exercising regulatory roles for AI in healthcare, focussed on different stages from initial research to adoption in the NHS.

The **Health Research Authority (HRA)** protects the interests of patients and the public in research in England [i]. It oversees the Research Ethics Committee system, advises on using confidential patient information[ii], and approves any research done in the NHS.

The **Medicines and Healthcare Products Regulatory Agency (MHRA)** oversees the release of any new healthcare products onto the market (whether for sale or as free software). The equivalents in Europe are the EMA (with delegations to national authorities) and the FDA in the USA.

The **National Institute for Health and Care Excellence (NICE)** guides the use of products or methods in the NHS in England[iii] based on performance and cost-effectiveness.

The **Care Quality Commission (CQC)**[iv] regulates health and social care services, registering providers, and monitoring services.

Bodies such as the General Medical Council and Nursing and Midwifery Council, and Royal Colleges can also have influential roles in setting standards within professions.

NHS Digital provides guidelines and ensures that any IT systems not already regulated as healthcare products are evaluated before and during use.

NHS[X] connects the Department of Health and Social Care, NHS England, and NHS Improvement, and supports digital advances in health and care services, setting national policies and strategy.

Most AI-enabled healthcare products and services will also be regulated under wider legislation.

The General Data Protection Regulation (UK GDPR) defines how personal information can be kept and used and covers automated decisions based on it. The Information Commissioner's Office works to uphold UK GDPR and other information rights.

The Equality Act addresses discrimination and bias.

The Health and Safety Executive is responsible for general workplace safety and accident investigation.

The UK regulatory bodies influence, and are influenced by, wider international views on medical products and AI, which ensures some consistency between national approaches.

- The international state of the art in science, medicine, and technology is a significant factor in individual decisions and in changing regulatory expectations over time.
- International standards such as those published by the ISO (International Organization for Standardisation) or IEC (International Electrotechnical Commission) can be influential, and the British Standards Institute (BSI) is a significant contributor.
- Information sharing and networking between regulators supports harmonisation and cooperation, especially on emerging technologies such as AI – the International Medical Device Regulators Forum (IMDRF) is particularly influential. Higher-level bodies such as the WHO can also influence general strategy, policies, and ethics on Digital Health and AI.

2. Product Regulation for Medical Devices and Software

In the UK, the Medicines and Healthcare Products Regulatory Authority, an executive agency of the Department of Health and Social Care, is responsible for ensuring products meet applicable safety, quality, and efficacy standards.

At time of writing, the Medical Devices Regulations 2002[1] [1] remain in force, as the EU's new Medical Devices Regulations [2] did not come into effect[v] before the UK left the EU. However, the Medicines and Medical Devices Act 2021 [3]empowers the Government to change the 2002 regulations to reflect international or national best practice, starting this year.

Scope of regulations

A medical device is any product that diagnoses, monitors, or treats an illness, alleviates a disability or modifies normal function, but which does not act as a pharmaceutical or biological agent. AI and software can be regulated as components of a physical device, as accessories, or as stand-alone products.

The medical device legislation does not apply to products that:

- are for all-round wellness, fitness, healthy eating, etc without a link to specific diseases
- are not intended to benefit individual patients (e.g. professional medical education simulators, or lab workflow control systems)

[1] These align with EU Directives from the 1990s

- are produced in a health centre for local use only, without any commercialisation intent

The contribution software makes within a medical decision is an important factor. Software which merely organises and displays diagnostic information, or shows relevant published guidelines for treatment, might not be classed as a medical device[vi].

Principles

The core principle is that before reaching the market, there must be an assessment to ensure: either minimal risk to patients or a risk proportionate to benefits; that stated functions and outcomes are achieved in practice; that products conform to current standards expected in the area; and that user instructions are suitable.

Products that meet these standards can use the UKCA mark (replacing the EU CE mark) and can be marketed, though the producer must monitor ongoing safety.

The level of evidence needed for approval can be lower than for new medicines. Specially designed clinical studies are not always required for medium- and low-risk products unless they are very novel: the scientific literature, known features of equivalent devices, and lab tests may be sufficient. Assessments can focus on the intended product functions rather than medical benefit endpoints.

The assessment needed depends on the level of risk. New products are placed in one of four broad classes, ranging from Class I to Class III (see Table 1). The evaluation must be done by an approved, expert, independent body (a "UK Approved Body[vii]") – except in the lowest risk classes where the producer can make the assessment.

Software in the 2002 regulations

As well as the usual expectations of performance and safety in medical technology, current UK regulations specify some requirements particularly applicable to software[viii]:

- Demonstration that its use in combination with other systems is safe, and users have clear instructions on use limits.
- User instructions on how to verify installation, monitoring, recalibration, updates etc.
- Repeatability, reliability and performance, and minimal risks from a single fault condition
- Software validation to the current state of the art in development lifecycle, risk management, validation, and verification
- The information display, alarm/alert systems etc., must be clear, ergonomic, and practical.
- Measurement or diagnosis software must have sufficient accuracy and stability for the task, and limits on accuracy must be set out by the manufacturer.

Table 1: UK product classes and assessments (excluding IVD[ix]s)

Device examples	*Typical assessments (current legislation)*
Class III	
Examples : active implanted devices such as pacemakers or cochlear implants Artificial Intelligence is likely to feature in these products in future.	Specifically designed clinical investigation almost always needed Full quality assurance OR type testing combined with other verification and QA checks. Occasional audits of production and QC
Class II(b)	
Example : radiotherapy devices or ventilators AI-enabled products which support or control active therapeutic devices (e.g. drug administration, surgery, radiotherapy) , and AI for critical diagnostic / monitoring tasks where error could lead to immediate danger.	Full quality assurance OR type testing combined with other verification or QA checks. Specifically designed clinical investigations *often* needed Occasional audits of production and QC
Class II(a)	
Example: surgical instruments; general diagnostic, assay, monitoring instruments. AI enabled products supporting diagnosis – for example many radiology applications - and disease / treatment monitoring would be in Class II(a).	Review of manufacturer's assessments of conformity, combined with other verification or QA check; OR full QA. Clinical evidence from equivalent products with of the same design can be taken into account. Occasional audits of production and QC
Class I	
Example: syringes, wheelchairs, software providing basic information, prioritising or tracking activity. Chatbots enabled with basic AI to aid decisions on whether to visit a GP; or professional systems for prioritising cases for attention could be in Class I.	Usually based on manufacturer's assessments of conformity, without independent review.

European Union

The EU and UK regulations evolved together until the UK left the EU. The EU's new Medical Devices Regulations continue many of the features of UK law, but with separate legislation for IVDs[x] and medical devices, and with new features that include:

- Post-marketing surveillance – manufacturers will have a greater obligation to gather a broader range of evidence (user feedback, clinical follow-up, scientific literature) proactively and regularly on products in use and apply this not only to ensure product safety but also to improve useability and performance.
- Information and traceability – all products on the market will be registered on a unified database (EUDAMED), and each individual product will have to carry a Unique Device Identifier (UDI) for traceability,
- Clinical studies – clearer guidance on the standards expected in clinical investigations

On software, the EU has published new guidance [4] on scope and risk classes. More software will be assessed within higher-risk categories in future. Software "*intended to provide information which is used to take decisions with diagnosis or therapeutic purposes*" is now in Class II(a). Products supporting decisions on severe conditions or life-critical implications are in either Class II(b) or III.

United States of America

The US Food and Drug Administration leads on regulating medical devices and software in the USA. It handles products in three risk classes:

- Class I products, the lowest risk group, are mainly exempt from the need to notify the FDA before marketing, but must still comply with general regulatory requirements and quality standards.
- Class II products, which include most diagnostics, monitoring, and interventional devices and corresponding software, usually need premarketing notification to the FDA. The most common route used is the 510(k) in which a manufacturer shows that their product is substantially equivalent to one on the market[xi]. The premarket notification submitted might need to include lab or clinical evidence on functions, and evidence that its mode of operation is either similar to previous products, or differs technologically but introduces no new safety or effectiveness questions. The producer also needs to show it meets general and type-specific regulations and quality standards.
- If there is no previous equivalent, but a new product is very likely to be classed as low- or medium risk, a special De Novo Classification can be sought.
- Class III products – the highest risk group, including defined high-risk product types (e.g. active neurological implants) and some very novel products – must have a Premarket Approval, the most stringent assessment, and robust clinical evidence is always required.

In contrast to the UK and EU systems which rely on third-party assessments, the FDA practice has been for FDA officials to assess submissions (supported by advice from expert panels). However,

there is increasing use of a new route allowing accredited Third Party Review for low- and medium-risk devices.

The FDA has a new Digital Health Centre of Excellence to support high-quality digital health innovation, which among other roles, explores new regulatory models for software as a medical device (SaMD") and AI/ML – work which is discussed further in Section (4).

3. Product assessment for use - digital tech and AI/ML in the NHS

The level of evidence needed to justify large scale use in healthcare is often much higher than is required to simply place technology on the market. The National Institute for Health and Care Excellence (NICE) has a crucial role in (a) issuing thorough evidence-based guidance on specific technologies[xii], which is only possible in a small number of cases each year and (b) providing quality standards for decision-makers in the UK health and care services.

For Digital Health Technologies, including AI-based products (using fixed algorithms only), NICE offered standards on the evidence needed on effectiveness and economics [5] in 2019, updated and simplified in 2021, which divides technologies into three Tiers:

> Tier A – mainly system services, such as EHR platforms and ward management systems
>
> Tier B – mainly concerned with communication and understanding, for example helping support users understand health topics, lead healthy lifestyles or stay in touch with their GP.
>
> Tier C – mainly about intervention, preventing or managing disease. This tier includes specific preventive technologies (smoking, sexual health etc.), patient self-management . diagnosis support, treatment advice etc.

Many of the Tier A and B products might not be classed as medical devices in UK law.

In the Tier C, the minimum evidence standard would require a high-quality intervention study showing improvement in diagnostic accuracy, clinical outcomes, or behaviour change. Best practice would be a randomised controlled study (or meta-analysis) in a setting relevant to the UK system. In addition, evidence would be needed on both health professional and user credibility and acceptability, reliability of the information, safeguarding, inequalities risks, and arrangements for ongoing monitoring.

Innovators and adopters of AI-based products are also expected to meet UK government good practice guidelines [6], which set expectations on data transparency and ethics, usability, technical assurance, clinical safety, cybersecurity, interoperability, open standards, etc.

App Library

With well over 100,000 health-related mobile Apps available, primarily not regulated as medical devices, the NHS has offered an Apps Library to simplify and support decisions on which products to

use. Having the "NHS stamp of approval" has had a large impact on trust and uptake for the selected Apps. For the future, however, the NHS envisages moving away from the single library and instead highlighting Apps supported by policy experts within the NHS information on particular health topics and services.

Some private sector organisations now also offer App libraries and/or other digital health testing and assurance services – such as ORCHA.

Digital Technology Assessment Criteria (DTAC) (2021)

To support good, timely, and proportionate decision-making across the vast range of digital technologies, software, and mobile apps that could be considered for use in the NHS, in 2021, the NHS issued new Digital Technology Assessment Criteria [7]. These are intended for developers, managers, and policy experts, to cover both products that are MHRA registered and those that fall outside medical devices legislation. They align with the NICE Evidence Standards Framework (3.2), addressing value propositions and over thirty assessment aspects in the areas of:

- Clinical safety
- Data protection
- Technical security
- Interoperability
- Usability and accessibility

4. Product regulation – software and AI/ML as a special case

For at least two decades, regulators have been considering where and how medical software needs a different regulatory approach to other medical devices. Policies have reflected that where software is concerned:

- Classification as a *medical* device with a specific risk category can be challenging since the uses and impacts of software and outputs vary between users and over time
- Updates *can be* made more often and faster than for other products, and sometimes security or compatibility issues mean they need to be made urgently
- Completeness in testing is problematic – even test regimes covering every line of code cannot be relied on to show how advanced software will behave in every situation
- Software is usually used as part of a complex system: most failures relate to untested interactions with other elements or later updates to parts of the system.
- Long-term monitoring of performance in use is easier than for other products

Since 2018, the US Food and Drug Administration has been exploring a software precertification model to make it easier to launch, improve and update medical software, and thus accelerate digital healthcare. The idea is that where producers' processes over the lifecourse of the product – from the early definition of user needs to long-term in-use monitoring – can be shown to be of the

highest quality, products and updates could be released without FDA approval, or with just a streamlined assessment. Streamlined FDA assessments would be retained for new products in medium or higher risk products (depending on the rating given to the producer's processes), and for major upgrades in the higher risk groups only. All minor changes from any assessed producer could be released directly. In September 2020, this framework was still in a development phase, with work ongoing on the type of evidence needed to assess the producer processes, define health benefits, and define what the streamlined FDA assessments would involve.

Artificial Intelligence

Artificial intelligence can present further regulatory challenges across a similar range:

- Complete testing of how a model will respond to all conceivable inputs can be impossible, and subtle variations in upstream parts of the systems - which can seem insignificant to operators – can significantly affect performance.
- The ways people use AI when making decisions can drift towards over-reliance or over-interpretation, rather than sticking to what was intended by producers or approved by regulators.
- Monitoring in-use can be difficult if the workings of the AI-enabled system are not well understood by users
- AI systems based on machine learning can need more regular updates as new data accumulate to provide the best achievable performance, and uniquely, AI has the potential to be set up so that updates are wholly or partly automatic.

Regulatory policies and/or plans in several jurisdictions recognise that for AI (or software in general), there may be a need to allow producers to release updates, or systems that learn autonomously, without changes being approved again by regulators, but only if this is adequately controlled and the changes stay within predefined product boundaries. For example:

- The European Union, in its April 2021 proposals[xiii] for future legislation on AI, envisaged that for regulated high-risk uses of AI, later changes to the algorithm or its performance that were pre-determined and addressed in the original technical documentation before approval would not count as substantial modifications needing a new pre-release assessment.

- The Japanese Ministry of Health, Labour and Welfare is moving to enable faster cycles of in-use monitoring and improvement for all medical devices[xiv], with lighter checks, and envisages that minor changes could be made post-marketing as part of AI-based devices of a pre-agreed 'Improvement Process' without additional approvals [8] [9]

- The Korean Ministry of Food and Drug Safety envisages [10] a need for approval and certification for changes to medical software, but with exemptions for AI-based learning when the change is only about improving accuracy (without design changes) within a quality management system, pre-set policies on training data, and pre-set functional boundaries.

These policy positions are quite new and not yet accompanied by detailed guidance.

However, the US FDA has already developed a detailed framework for handling modifications for AI/ML based software. It consulted on a framework [11]in 2019 and issued a plan [12]early in 2021.

The FDA policy is built on the idea of greater emphasis on a Total Product Lifecycle Approach in regulation. Assurance on the quality of manufacturers' monitoring practices, communication with users on performance on updates, and 'Good Machine Learning Practice' will all help ensure lighter-touch regulation and faster innovation. The approach remains risk-proportionate, with more direct regulatory oversight in higher-risk product classes.

When new products are approved within the FDA framework there would be an accompanying 'Pre-determined Change Control Plan' which would include:

> *Software Pre-Specifications* – i.e., the anticipated changes to algorithm performance, types of input data and patient groups covered, and/or changes in intended use, to define a reasonable 'region of potential change'. This might include using data from a wider range of instruments from different suppliers, or a wider range of patients, for instance.
>
> *Algorithm Change Protocol* - which would include, for each type of change envisaged:
>
> - Data management – collection, quality, audits etc
> - Re-training – aims, ML methods, data pre-processing, criteria
> - Performance evaluation – metrics, statistics, evaluation triggers, expert involvement
> - Update procedures – including software testing, timing, implementation of updates, user communication and transparency.

The FDA has already made a first product approval that includes a 'Pre-determined Change Control Plan', and full guidance on what these plans should cover is in preparation.

5. International Standards

International standards bodies (ISO and IEC, CEN/CENELEC) and national bodies such as BSI have published or developed standards for AI-based products. A BSI report [13] provides a useful overview of how AI is being covered in specific standards and guidelines. Standardisation work includes areas such as neural network robustness, bias, ethics and human factors, trustworthiness, and software testing for AI. Most aim to set higher level standards that could be relevant to many AI application areas, not just health applications.

In addition, many well-established general standards – such as in Quality Management (ISO9000 series), medical software lifecycle standards (IEC 62304), software testing, health software safety and security (IEC 82304), human usability of medical devices (IEC 62366) and medical device risk

management (ISO 14791) – include some content that is highly relevant for AI-based product development and conformity assessments.

6. Publications evaluating AI in healthcare

Although there is a rapidly expanding volume of peer reviewed papers reporting on AI in healthcare, by 2020 commentators were pointed out a need for higher quality evidence on performance and value, and more complete and transparent reporting, to support wider use AI in healthcare.

There are good overviews in a 2019 paper from Google [14] and a 2020 collaborative academic paper [15]. Systematic reviews of specific fields often make similar points about the need for stronger evidence (e.g. recent reviews of deep learning in imaging [16] [17]).

The issues raised reflect the early stage of development of a lot of AI research in 2019 and 2020 and include:

1. The **data sources** relied on in evaluations need to reflect the full heterogeneity of real clinical cases, and realistic levels of data quality, incompleteness etc.
2. **Evaluation measures** need to reflect the clinical decisions the AI is intended to support, and compare AI performance with current best practice. For example, many papers focus on the area under the curve (AUC) for the Receiver Operating Characteristic (ROC) which compares false positives /true positives over a wide range and gives and averaged metric. But in clinical decisions other parameters (e.g. false negatives) are important, and the benefits or harms from some errors are much greater than for others, so an average may not be meaningful.
3. Few of the clinical AI studies published by 2020 were **prospective, randomised, multi-site**, studies. **Generalisability** across clinical populations and locations, and ability to cope with data drift over time, also needed more attention.
4. Evidence of **clinical value** is often a weakness – i.e. evidence of the health gains achieved when the AI is deployed in a real healthcare pathway. Better performance on a single task in the pathway does not always result in better outcomes for patients.
5. **Inequalities –** evaluations need to address AI's potential to either reduce or increase inequalities or racial biases in healthcare with new paradigms, or to lock-in old inequalities.
6. **Comparative evidence** contrasting AI methods or products needs to be more common. In early-phase research, the variety of different data sources and metrics used makes comparisons difficult, and in later stage research, very few clinical studies compare products.
7. Evaluative research needs to go further in addressing the **interpretability** of the AI, and its incorporation into **human decision making** systems.
8. **Logistical and safety aspects**, such as data quality control needed, adequacy of in-use monitoring and safety checks are rarely reported.

These issues are not unique to health AI. Across many fields of use of AI, there is a tendency for published work to focus on correctness and robustness on specific tasks, with less attention to efficiency, privacy, fairness, wider model relevance, and interpretability [18]. These aspects often

need more time and expense, and are less well supported by standard testing benchmarks or resources.

As part of the work of the Equator network, updated guidelines for complete and transparent reporting of clinical studies and protocols based on AI – CONSORT-AI [19] and SPIRIT-AI [20] - were published in 2020.

The technical details on AI that should be included in publications have also been debated. For example, in response to a 2020 Nature paper on an AI system in mammography, Nature published both an independent critique [21] suggesting standard details that should be routinely reported to ensure reproducibility (e.g. data pipeline, training hyperparameters and phases, software code / deep learning models used), and a large body of additional technical information from the developers [22].

Notes

[i] Other bodies have equivalent roles in Scotland (Health Boards), Wales (Division of Social Care and Health Research), and Northern Ireland (ORECNI)

[ii] Independent advice is provided by the Confidentiality Advisory Group in relation to research and uses beyond direct patient care. Across all health and social care data, the National Data Guardian has a border, independent statutory remit, including guidance for the Caldicott Guardians within each NHS organisation.

[iii] In Scotland, this is done by Healthcare Improvement Scotland, the Welsh Government works with NICE and expects its guidelines to apply in Wales.

[iv] This function is organised differently across the UK – for example, through the Care Inspectorate in Scotland

[v] The devices legislation took effect in May 2021, the separate In Vitro Diagnostics legislation takes effect in May 2022.

[vi] For more details see Medical devices: software applications (apps) - GOV.UK (www.gov.uk)

[vii] Replacing the EU's "Notified Bodies"

[viii] Source: Guidance: Medical device stand-alone software including apps (including IVDMDs) v1.06

[ix] In Vitro Diagnostics – some AI applications will be in IVDs, where there is a separate four-level risk-based classification.

[x] An AI product linked to IVDs would be regulated under these laws – for example, a system analysing Whole Genome Sequences to identify patterns of variation associated with disease risk

[xi] The requirement is to show equivalence to a product on the market in 1976, or a more recent product that has itself been judged equivalent

[xii] Specific guidance can only be offered on a small fraction of the technologies available

[xiii] Recital (6), Articles 13 and 43. However, these blanket proposals cover all applications of AI and are some years from becoming. It isn't clear when or how they will translate into changes in medical devices regulation.

[xiv] "IDATEN" or "Improvement Design within Approval for Timely Evaluation and Notice"

References

[1] UK Legislation : Medical Devices Regulations 2002, [Online]. Available: https://www.legislation.gov.uk/uksi/2002/618/contents/made. [Accessed August 2021].

[2] EU : Medical Devices Regulations, [Online]. Available: https://eumdr.com/. [Accessed August 2021].

[3] UK Legislation : Medicines and Medical Devices Act 2021, [Online]. Available: https://www.legislation.gov.uk/ukpga/2021/3/contents/enacted/data.htm. [Accessed August 2021].

[4] Medical Device Coordination Group Document MDCG 2019-11, October 2019. [Online]. Available: https://ec.europa.eu/health/sites/default/files/md_sector/docs/md_mdcg_2019_11_guidance_qualification_classification_software_en.pdf. [Accessed August 2021].

[5] National Institute for Health and Care Excellence, "Evidence Standards Framework for Digital Health Technologies," 10 December 2018. [Online]. Available: https://www.nice.org.uk/corporate/ecd7. [Accessed August 2021].

[6] Department of Health and Social Care, "A guide to good practice for digital and data-driven health technologies," 19 January 2021. [Online]. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology#generate-evidence-that-the-product-achieves-clinical-social-economic-or-behavioural-be. [Accessed August 2021].

[7] NHSx, "Digital Technology Assessment Criteria," [Online]. Available: https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/. [Accessed August 2021].

[8] Ministry of Health Labour and Welfare. Takanashi Fumihito (Medical Device Evaluation Division. [Online]. Available: https://www.pmda.go.jp/files/000234056.pdf. [Accessed August 2021].

[9] Chinzei K, et al "Regulatory Science on AI-based Medical Devices and Systems," *Advanced Biomedical Engineering,* vol. 7, pp. 118-123, 2018.

[10] Ministry of Food and Drug Safety, Republic of Korea, "Regulations: Guideline on review and approval of artificial intelligence (AI) and big data-based medical devices (for industry).," 04 11 2020. [Online]. https://mfds.go.kr/eng/brd/m_40/view.do?seq=72623&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=1. [Accessed August 2021].

[11] US Food and Drug Administration, "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," 2 April 2019. [Online]. Available: https://www.fda.gov/media/122535/download. [Accessed August 2021].

[12] US Food and Drug Administration, "Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan," 12 January 2021. [Online]. Available: https://www.fda.gov/media/145022/download. [Accessed August 2021].

[13] British Standards Institute, "BSI White Paper – Overview of standardisation landscape in artificial intelligence," [Online]. Available: https://www.bsigroup.com/en-GB/industries-and-sectors/artificial-intelligence/download-the-artificial-intelligence-whitepaper/. [Accessed August 2021].

[14] Kelly C, et al "Key challenges for delivering clinical impact with Artificial Intelligence," *BMC Medicine,* vol. 17, p. 195, 2019.

[15] Vollmer S, et al "Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness," *The British Medical Journal,* vol. 368, p. 16927, 2020.

[16] Liu X, et al "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health,* vol. 1, p. e271, 2019.

[17] Nagendran M, et al "Artificial intelligence versus clinicians: systematic review of design, reporting standards and claims of deep learning studies," *The British Medical Journal,* vol. 368, p. m689, 2020.

[18] Zhang J, et al "Machine Learning Testing: Survey, Landscapes and Horizons," *IEEE Transactions on Software Engineering,* February 2020.

[19] Liu X, et al "Reporting guidelines for clinical trial reports for interventions involving Artificial Intelligence : the CONSORT AI extension," *Nature Medicine,* vol. 26, p. 1364, September 2020.

[20] Cruz Rivera S, et al "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," *Nature Medicine,* vol. 26, p. 1351, 2020.

[21] Haibe-Kains B, et al "Transparency and reproducibility in artificial intelligence.," *Nature,* vol. 586, p. e14, 2020.

[22] McKinney S, et al "Addendum: International evaluation of an AI system for breast cancer screening.," *Nature,* vol. 586, p. E19, 2020.