Transparency and product safety regarding medical diagnostic systems

Transparency and product safety regarding medical diagnostic systems

Daria Onitiu

University of Edinburgh, United Kingdom

donitiu@ed.ac.uk

Abstract. A promising research area, revolutionising the early detection of diseases, entails the use of medical diagnostic systems for clinical decision support. However, we need to conduct a careful balance between a system's performance and its usefulness in informing patient outcomes. This paper is focusing on the ethical challenges of medical diagnostic systems on patient autonomy, as well as notions of shared decision-making, and evidence-based medicine, suggesting how to address these issues in a regulatory landscape. It intends to scrutinise broader efforts by the U.S Food & Drug Administration on Artificial Intelligence and Machine Learning software as medical device and claims that the quality of oversight ultimately informs the safety of medical diagnostic systems as clinical decision support. Additional guidance needs to establish the conditions of a medical diagnostic system's continuous alignment of patient values and stimulate causal effects in clinical reasoning.

Keywords: medical diagnostic systems, ethics, transparency

1 Introduction

Medical diagnostic systems transform healthcare. By way of illustration, imagine a medical diagnostic system that uses machine learning (ML) to classify between mild or advanced diabetic retinopathy. The U.S Food and Drug Administration (FDA) has recently approved such a device, which employs an algorithm to analyse images of a patient's retina, providing an important contribution in managing a common disease in diabetic patients that leads to vision loss (FDA, 2018). Yet, medical diagnostic systems currently operate in a regulatory lacuna, and we need to think about how disease classification using ML will shape the role of a healthcare professional and the patient when engaging with such a system on the ground. This paper intends to uncover some ethical problems of medical diagnostic tools, focusing on the interplay of performance modifications and the need for transparency regarding the certification of these Artificial Intelligence (AI) tools.

We need a nuanced discussion on the impact of medical diagnostic systems on ethical principles, including patient autonomy, as well as notions of shared decision-making, and evidenced-based medicine. For instance, individual autonomy can conflict with notions of shared-decision-making when conflated by algorithmic constructions on disease classification. Moreover, I highlight that ML approaches do not necessarily improve the quality of decision-making, including the healthcare professionals acting on best available evidence. There is a risk that a medical diagnostic system, if not defined appropriately within ethical principles, can blur the line of human intervention, and disturb the role of ML-approaches as clinical decision-support.

The FDA intends to address some challenges of medical diagnostic systems and envisages a product lifecycle approach suitable for adaptive algorithms in a healthcare environment (FDA, 2019, p. 10; FDA, 2021, p. 3). Yet, we need to substantiate some aspects of this regulatory proposal, including the interplay between performance specifications and transparency requirements, focusing on medical diagnostic systems. I suggest a differentiated picture of how transparency goals, including so-called post hoc explainability methods in medical diagnostic systems, can complement FDA proposals that go beyond a system's intended use and will consider the impact of ML approaches on clinical decision-making.

An important step to verify the interplay ML-approaches with the users, healthcare professional and the patient in a clinical environment is to understand the role of potential and causal effects of medical diagnostic systems. That is, we introduce a "language" in the model to allow us to formalise the knowledge and assumptions in the data based on underlying cause-and-effect relationships (Miller, 2021). Modelling causal effects is one aspect of ensuring effective human oversight, and we need to consider this as an important safeguard to channel ethical principles within the role of verification and validation of medical diagnostic tools.

2 An outlook of the ethical challenges of medical diagnostic systems

The study of medical ethics allows us to discover the principles and processes to justify a particular course of action, including a practitioner's communication of evidence and considering patient involvement in the decision-making process (Laurie, Harmon, Porter, 2013; p. 2). Some research enumerates the ethical challenges of medical AI systems holistically (for example, Luxton, 2022). Building on this work, I provide a more nuanced picture showing how algorithmic processes can produce tensions with some core ethical principles within shared decision-making and evidence-based medicine.

2.1 ML approaches do not readily fit with patient autonomy

The principle of patient autonomy entails the healthcare professional's duty of negative and positive action to enable patients making an informed decision about their medical

care (Beauchamp and Childress, 2019, p. 104; Holm, 2022, p. 183). As argued by Christine and Kaldjian (2003, p.10), 'communicating information about prognosis and treatment is recogni[s]ed as one of the clinical cornerstones of respecting patient autonomy' as well as shared decision-making. With medical diagnostic tools, the position for defining patient autonomy lies in the system's *data* about the individual patient and the model's information-processing technique describing the healthcare professional's associative process, judging between the benefits and risks regarding the system's output.

Let us elaborate on this, considering the meaning of a consequentialist notion of contemporary medical ethics, which examines 'the effect of a decision on individuals' and whereby the individual creates 'their own decisions as far as possible' (Laurie, Harmon, Porter, 2016, p. 6). We are interested in how the use of algorithms for disease classification can support a patient's welfare, such as satisfaction or wellbeing, including the actions leading to the best possible outcome. Therefore, we might want to focus on consequentialist theory as it puts us in the position to judge the relative effects of actions and importance of outcomes (Card and Smith, 2020, p.4).

Imagine a medical diagnostic system that can analyse chest x-rays and provide a prediction that is associated with the individual patient suffering from pneumonia. How do we evaluate the possible benefits and harms of this predictive process and how much weight should we place on the probabilities? Do we consider the degree of benefit and harm as a statement of the system's automatically testing chest x-rays scans, or do we need to go further and look at the algorithm's functional representations informing both, the doctor's, and individual's own value judgement? There is a risk that an agent's actions and probable outcomes are rather associated with the system's probabilistic account of disease classification, than related to the individual's calculation of risk management and communication. That could, in turn, distort the way the patient perceives the system's performance specifications and articulates own choice.

Moreover, healthcare professionals not only need to quantify shared decision making, considering the probability of benefits and harms regarding disease classification using AI, but also streamline the contribution of medical diagnostic systems in a clinical environment. This is certainly not an easy task, as ML approaches can disturb constructed relationships in reasoning.

2.2 ML approaches can disturb constructed relationships

A healthcare professional needs to conduct a delicate balance between knowing a patient's best interests, including values and beliefs, and appraising the system's reliability in individual circumstances (Grote, 2021, p. 337). We need to note (i) a healthcare professional's consideration of a patient's interests and choice (Christine and Kaldjian, 2003, p. 13) (ii) a healthcare professional's positive action promoting patient wellbeing and utility (Beauchamp and Childress, 2019, p. 217), whereby the relationship between the doctor and the patient resembles a process of shared decision-making.

Nevertheless, it is important to note that ML approaches operate as a 'black box', whereby the algorithmic decision-making processes are not comprehensible to the

average user (Mittelstadt, Allo, Taddeo et al, 2016, p. 6). A healthcare professional may be impaired to have 'a realistic understanding of the system', beyond disclosing the system's specificity and sensitivity (Holm, 2021, p. 183). However, is there a moral imperative for the healthcare professional to rely on a predictive model that can outperform human judgement and possibly override a patient's demand for alternative treatment options? What this shows is that medical diagnostic systems can conflict with patient autonomy and beneficence, endorsing a notion of soft paternalism based on the system's intended use as a disease classification tool.

Imagine a scenario where a medical diagnostic system detects pneumonia in an image of the patient's chest x-ray and the healthcare professional must decide on the underlying factors that influenced the classification outcome. What follows is that medical diagnostic systems could give away new interpretations of patient autonomy and beneficence in clinical decision-making. A medical diagnostic system that operates as a 'black box' could minimise a healthcare professional's discretion in exercising clinical judgement, which includes those intuitive assumptions about the appropriate treatment recommendations that reflect his or her experience and the patient's best interests. There is a risk that the use of AI can steer positive action to the degree that the value attached to probabilistic judgements remains unverified by both the healthcare professional and patient.

We need to determine what are the harm-inducing conditions that go well beyond a system's functional use, and which shape a healthcare professional's risk communication and management in a process of shared decision-making. I suggest that a healthcare professional must judge a system's confidence level based on the degree of information that allows exercising clinical discretion to balance ethical principles, as well as utilising clinical expertise based on scientific evidence. Nevertheless, another issue concerning the use of AI for decision support is that ML approaches do not necessarily promote evidenced-based outcomes, as I will show below.

2.3 ML approaches do not necessarily improve evidenced-based outcomes

Evidence-based medicine concerns the process of decision-making that combines 'clinical expertise and patient values' with 'best available evidence' (Burlacu, Iftene, Busoiu et al, 2020, p. 191). For example, a medical diagnostic system is argued to provide evidenced-based modalities in imaging, supporting clinicians 'in integrating ever-increasing loads of medical knowledge and patient data into routine care' (Scott, 2018, p. 44). However, medical diagnostic systems do not necessarily promote evidence-based outcomes. Google Health conducted an interesting study examining the impact of deep learning approaches on the detection of diabetic retinopathy within 11 clinics across provinces in Thailand (Beede, Baylor, Hersch et al, 2020). The report illustrates a discrepancy between the system's accuracy in a laboratory and real-life environment and highlights that the deployment of medical diagnostic systems needs to be tailored to the clinical workflow, such as considering data quality and socio-economic factors in specific healthcare contexts (Beede, Baylor, Hersch et al 2020; Douglas Heaven 2020).

3 Medical diagnostic systems and FDA proposals: performance versus transparency

The previous discussion identified some challenges of medical diagnostic systems regarding decision-making, and the need for more guidance in considering patient autonomy, the notion of shared decision-making, and the deployment of these tools within evidenced-based statements. Therefore, we should specify the kind of verification and validation requirements required to tackle these challenges.

The FDA proposals focusing on the role of ML approaches in software as medical device establish a suitable starting point to investigate these issues. It intends to provide an approach regarding the so-called 'update problem' of medical devices by using ML/DL approaches that are iterative and adaptive to a healthcare environment (Gilbert, Fenech, Hirsch et al, 2021, p. 2; Gerke, Babic, Evgeniou et al, 2020, p. 1; FDA 2019, p. 3).

The FDA examines the safety and effectiveness regarding software changes in their Discussion Paper and the Action Plan, which I shall call the "FDA guidance" in the remaining part of this discussion (FDA, 2019; FDA 2021). The FDA guidance, providing an important premarket assurance regarding software modifications, illustraes the Predetermined Change Control Plan that includes Pre-Specifications (SPS) and Algorithmic Change Protocols (ACPs) (FDA, 2019, p. 10; FDA, 2021, p. 3). The SPS and the ACPs both illustrate a shift from dealing with a software's significant changes to types of modifications that are 'anticipated' (FDA, 2019, p.10; FDA, 2021, p.1). In addition, the FDA guidance stipulates enhanced transparency requirements based on manufacturer's real -world performance monitoring (FDA, 2019, p. 9; FDA, 2021, p. 1).

The FDA guidance, endorsing a notion of safety and performance of medical AI devices, including medical diagnostic systems, stands in sharp contrast to transparency. Transparency is commonly defined as the scrutability of algorithmic decision-making, being closely aligned to explainability (Mittelstadt, Allo, Taddeo et al, 2016, p. 6). As acknowledged by Mittelstadt, Allo, Taddeo et al (2016, p. 6), the debate of transparency of algorithms is not new. Nevertheless, I assume to provide a different perspective on transparency in medical diagnostic systems, which includes an extension of a system's intended purpose to the application of ethical principles within clinical decision-making. Whilst we can agree that diagnostic skill can be examined in standard settings, such as comparing clinician judgement and the system's specificity and sensitivity, there are important variations in clinical judgement that go well beyond academic ability (see also, Chan, Gentzkow, Yu 2021). Take the example of a medical diagnostic system that offers the same classification to similar cases, but obliges the health care professional to act differently, based on the patient's own needs, values, and preferences. What I intend to show is that further guidance needs to specify that disease classification is a coordinated process that shapes how the doctor and the patient perceive the system's computational interpretation of an underlying disease.

3.1 Patient autonomy and ex ante usability

An important aspect of the FDA guidance is the connection between transparency and usability using a 'patient-centered approach' (FDA, 2021, p.4). A system's usability can be defined as 'the extent to which a [ML] system can be used to achieve specified goals with effectiveness, efficiency, and patient satisfaction in multiple healthcare environments' (Cutillo, Sharma, Foschini et al, 2020, p. 1).

Imagine a medical diagnostic system with improved performance in classifying images of diabetic retinopathy after the manufacturer retrained the algorithm on real-world data (see FDA, 2019, p.18). The first aspect for manufacturers is to ensure the system's usability and document performance improvement, including analytical and clinical validation in the ACP regarding the system's use (FDA, 2019, p. 18). Moreover, manufacturers need to ensure that any modifications and anticipated changes in the system's performance are transparent to the users (FDA, 2021, p. 14). However, there is a discrepancy between the role of usability in performance modifications and usability and transparency, whereby the former is constrained to the system's analytical and clinical performance *within* the system's intended use. A manufacturer documenting a modified algorithm that can determine high-confident cases based on real-world data, proves the medical diagnostic system's 'quality of use', considering the requirements regarding device labelling and real-world performance monitoring (Bevan, 1995). We need a more structured approach to define unanticipated impacts of medical diagnostic systems when interacting with the patients, including his or her expectations about future treatment, perception of symptoms and the role of AI in clinical judgement.

The FDA needs to consider another important aspect of usability that intends to investigate whether the medical diagnostic system promotes patient-centered outcomes. Cabitza and Zeitoun (2019, p. 161) illustrate this aspect of usability very well and distinguish between 'statistical validity' and the system's usability to verify 'the extent to which physicians can relate to the AI, attach some clinical meaning to its advice and integrate its use in their daily workflows and routines'. This interpretation of usability is appropriate to 'fill the semantic gap' regarding the expert's interpretation of medical imaging using AI (Holzinger 2016, p. 122). I would add that FDA guidance might need to safeguard the system's alignment with patient values, including autonomy to offer a comprehensive statement about a system's usability. Accordingly, my view of usability envisages how the healthcare professional can attach different values based on the system's probability of harm and benefits (see also, Savulescu and Wilkinson, 2019). In other words, how does the algorithm's outlook help healthcare professionals to act in favor of the patient's interests and values when considering the recommendation of treatment options? Does the algorithmic decision-making support the patient to act within his or her best interest?

If we want to tackle these questions, we could argue that the degree of human oversight can shape a system's intended use beyond a medical diagnostic tool's clinical and analytical performance. For example, a healthcare professional who is dealing with a system that classifies specific complications of diabetic retinopathy needs to weigh up different values informing risk communication and management to the patient,

rather than when dealing with mild forms of diabetic retinopathy that require the patient to come back for re-screening in the future. The FDA considers this scenario as entailing the manufacturer's real-world performance monitoring and updating device labelling, provided that the user's degree of human oversight is not based on a system's limitations or unchanged performance requirements (see also, FDA, 2019, p. 16). However, what is missing is that the manufacturer should also consider the way AI tools shape diagnostic decisions for various stages of disease classification tasks and how these findings can comply or conflict with ACPs and SPS.

The outcome is the need for a requirement of ex ante usability documenting potential risks that goes beyond the manufacturer documenting anticipated changes, such as updating ACPs using performance data. A requirement of ex ante usability would require manufacturers to update performance modifications, as well as technical safeguards, enabling different agents to foresee the degree of intervention with a medical diagnostic system, maximising an individual's choice. I will articulate the role of technical safeguards in this context in Section 3.2.

3.2 Shared decision-making and post hoc explainability

Another aspect I mentioned in the previous section is that manufacturers need to ensure that any modifications of the system's performance are transparent to the users (FDA, 2021, p. 14). This notion of transparency, whilst tied to the system's modifications, should arguably include the inscrutability of the medical diagnostic tool when it is 'actionable to the user' (COCIR, 2021, p. 19; FDA, 2019, p. 10). However, we need to provide a more comprehensive picture of the role of performance and explainability of medical diagnostic tools and analyse the way that interplay should influence the FDA approach to AI medical device regulation.

An important question arising from the notion of shared-decision making is the practitioner's and patient's involvement to act upon a system's output. The FDA guidance seems to investigate this by focusing on the device modifications that a system needs to show an '[a]ppropriate level of transparency (clarity) of the output and the algorithm aimed at users' (FDA, 2019, p. 10). The FDA, the United Kingdom's Medicines, and Healthcare products Regulatory Agency (MHRA) and Health Canada paper on Good Machine Learning Practices (GMLP) explicitly mention the 'need for the human interpretability of the model outputs' (FDA, Health Canada, MHRA 2021, Principle 7).

Post hoc explainability methods in medical imaging offer an important perspective concerning the degree of insight regarding the algorithmic decision-making. For example, a popular post hoc explainability method entails saliency maps in medical imaging tasks, highlighting the input pixels that contribute to the output (Seçkin Ayhan, Kümmerle, Kühlwein et al, 2022, p. 1). By way of illustration, a saliency map operating on an AI model to detect pneumonia can localise the regions of the image made for the prediction (see also, Da Silva, 2020).

However, explainability methods are not a source of justifying the system's reliability in individual circumstances (Lipton 2016, p. 41-42). For example, Imagine the saliency map focused on the pixel value in the image, rather than the aspects of the

image correlating with the underlying disease (Ghassemi, Oakden-Rayner, Beam, 2021, p. 746). Research by Arun, Gaw, Sing et al (2021, p.1) revealed that saliency maps do not assist clinicians to localise the underlying factors for disease classification. Accordingly, it would be naïve for us to assume that transparency safeguards and explainability ensure a system's continuous use regarding anticipated changes from the outset. However, we need then to ask ourselves how these technical safeguards correspond to 'the need for a manufacturer's transparency to users about the functioning of AI/ML-based devices [and] the benefits, risks, and limitations of these devices' and how can user perspectives inform the evaluation of the medical diagnostic system (FDA, 2021, p.5-6).

We must move away from a conception of post explainability methods enabling human observation of the algorithmic process and emphasise its role to support a healthcare professional's positive action towards an observable result. Let me elaborate on this using an example of a saliency map which outlines *some* features in a localised region in an image. Suppose now that the practitioner acknowledges that the model's localization of an area within the chest x-ray is the *deciding factor* describing the patient's pain and suffering of a disease. What this shows is that the model's relative feature importance for the output only gains significance when exhausted by the mutual interaction with the user, the healthcare professional and patient.

Future work needs to inform the role of post hoc explainability methods to inform the model's reconciliation with clinical decision-making, rather than concordance with human judgement based on the system's functionality in a clinical context. FDA guidance needs to establish post hoc explainability as a process-based verification method that requires specific training for clinicians when dealing with visualisation methods including saliency maps and translates medical diagnostic systems within the ambit of shared-decision making regarding diagnostic tasks.

Another aspect related to the example above is the representation of knowledge and expertise regarding an underlying disease as being a deciding factor for the model's output. The healthcare professional's current expertise about disease classification, including the diagnosis and treatment of pneumonia would contribute to his or her positive action to observe the algorithms' classification. Accordingly, I am going to analyse how clinical evaluation is another aspect for assessing probabilities for the system's alignment with evidenced-based medicine in Section 3.3.

3.3 Evidenced-based outcomes and clinical evaluation

The FDA, MHRA, and Health Canada's GMLP document recognises that datasets need to be 'representative of the intended patient population' and users need to be informed about the 'performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model' (FDA, Health Canada, MHRA, 2021, Principle 3, Principle 9).

Having said that, many FDA- approved medical AI devices are based on retrospective studies, making it more difficult for operators to understand the system's

nuances applied to individual cases (Wu, Wu, Danshjou et al, 2021, p. 582). Hence, it is argued that manufacturers need to consider validation studies considering prospective randomised studies to assess actual clinical outcomes, as well as comparisons between clinicians' performances with and without the medical diagnostic system (Wu, Wu, Danshjou et al, 2021, p. 582-583). The idea of randomized control trial is to 'unmask vulnerabilities' such as generalizability of the system's performance and inherent limitations, such as overfitting, to define patient-centered goals and outcomes (Wu, Wu, Danshjou et al, 2021, p. 582-583).

Therefore, the first aspect defining the role of clinical validation of AI systems is to achieve the 'empirically rigor' 'to maximize the benefits' of a medical diagnostic tool, whilst 'offsetting potential effects' regarding a system's implementation (McCradden, Anderson, Stephenson et al, 2022, p. 9). This requires the FDA to establish the parameters for constituting a sufficient demonstration of evidence and confirming the system's reliability in a clinical setting. The work done by researchers in developing reporting guidelines on clinical trials on AI tools, such as the SPIRIT-and CONSORT AI extension (Liu, Cruz Rivera, Moher et al, 2020; Cruz Ribera, Liu, Chan et al, 2020) is helpful to establish transparency of AI-based recommendations but more good practices need to deal with evaluation considering adaptive algorithms (Liu, Cruz Rivera, Moher et al, 2020, p. 1371). Moreover, manufacturers need to ensure that the clinical validation of an AI system considers how the information collected in a prospective study is contextualised in a clinical setting (Liu, Glocker, McCradden, Ghassemi et al 2022, p. 385). We need to address this question from the perspective of clinical evaluation as well as the post-market monitoring, to ensure that the system's performance ensures clinical outcomes are respected (Wu, Wu, Danshjou et al 2021, p. 583).

However, another issue is the inherent risks related to the misuse of medical diagnostic tools when interacting with a patient. By way of illustration, imagine a medical diagnostic system trained within a specific sub-population in San Francisco, United States and used on a different population with different socio-economic factors. What this shows is that we must pay attention to the 'more insidious failures, such as an algorithm that gives racially biased recommendations because it was trained with subtly biased data' (Kaushal and Altman, 2019, p. 62). I believe that the FDA needs to build up its statement in Principle 9 of the GMLP and elaborate on the risks of how medical diagnostic systems can give rise to "off-label" use (FDA, Health Canada, MHRA, 2021, Principle 9). Accordingly, we need more guidelines on how users need to be informed about the risks of a system's performance bias in a clinical setting and how unintended uses can be monitored during a system's deployment, considering the role of practitioners to ensure effective risk management and communication of probability of harm when these AI tools operate on the ground.

4 Setting the tone for the role of causal effects in medical diagnostic systems

An important point, which surfaced in my discussion concerning the verification and validation of medical diagnostic systems, is that clinicians and patients need to have more control of the tool's operation as decision-support. I established that we need a requirement of ex ante usability, which entails technical safeguards to enable a degree of human intervention regarding potential and unanticipated changes in the system's operation. In addition, I elaborated on the role of technical safeguards, including post hoc explainability, and how the practitioner and patient can realistically assess the system's confidence within a specific setting. Finally, I highlighted that the manufacturer needs to consider how different social actors contextualise a system's predictions to ensure evidenced-based outcomes. I will be closing the argument by specifying how we should discuss the degree of human control in the future.

Modeling causal effects and so-called 'what if' scenarios (Holzinger, 2021) can safeguard the role of medical diagnostic systems in clinical decision-making. Causal effects are not inherent in the algorithmic process (Pearl and Mackenzie, 2018). For example, a common argument regarding ML- predictions in healthcare is that the healthcare professional should rather know *why* the tool is making certain predictions than relying solely on algorithmic correlations (Holzinger, Carrington, Müller, 2020). We can investigate whether a model's simulation of causal effects can stimulate decision-making when acting within the interests of the patient. Whilst a comprehensive discussion would exceed the scope of this paper, I suggest extending my investigation by discussing the FDA approach and making three recommendations, based on the role of human control and intervention with medical diagnostic systems:

First, we need to identify the role of patient autonomy to justify certain actions when a ML system is performing certain tasks. Ex-ante usability is a requirement that needs to shape information duties, as well as device changes. In doing so, we must not underestimate the degree of one's own action to assess the contours of individual choice. That means a practitioner can *create* the patient's manifestation of a disease based on the system's association with the patterns in the data. Ex-ante usability is a requirement for manufacturers to engage with some boundary work and investigate how users, healthcare professionals and patients engage with a different set of scenarios. One way in doing so is the use of counterfactual explanations in medical imaging, which allows the individual to stimulate decision-making focusing on alternative scenarios (Vermeire, Brughmans, Goethals et al 2022). Counterfactuals allow us to quantify likelihood of harm by isolating individual circumstances, including preconditions (Glocker, Musolesi, Richens et al 2021, p. 3). This allows us to scrutinise and define those outcomes which are most closely related to the actions promoting wellbeing and the patient's interests.

Second, we need to steer a healthcare professional's positive action to achieve reconciliation concerning the use of medical diagnostic tools as clinical decision-support. Therefore, an important step for us would be examining the way healthcare professionals assess confidence levels with reference to the AI, including what parameters allow a healthcare professional to build up a reasonable justification for

positive action, such as recommending the patient treatment options. Here, I would specify that counterfactuals in medical diagnostic systems could direct the individual to engage with his or her perceptual judgement, based on possible reconstructions of the underlying factors, and navigating through the system's classificatory purpose.

Finally, we see the role of causal effects to define the need for expert knowledge in clinical evaluation (Glocker, Musolesi, Richens et al 2021, p.1). It is argued that randomized control trials allow the evaluation of causal hypothesis (Prosperi, Guo, Sperrin et al 2020, p. 369). Nevertheless, London (2019, p. 17) argues that the extent domain experts scrutinise those causal relationships 'derives from experience and precedes our ability to understand why interventions work'. What follows is that the aim in understanding causal effects in disease classification is based on dealing with uncertainties informing both, clinical evaluation, as well as risk communication regarding the use of AI on the ground. Hence, what we need are metrics for reasonable causal conclusions guiding randomized study.[1] The role of decision-making here is crucial to inform randomized study including the level of human expertise and '*how* the outputs of the AI system were used to contribute to decision-making or other elements of clinical practice' (emphasis added by author) (Liu, Cruz Rivera, Moher, 2020, p. 1371). What needs to be added; however, is that we need further multidisciplinary engagement that discusses the delicate process that should exemplify both, the benefits of explanations stimulating causal thinking between various users, healthcare professionals and novice operators and risks of the use of AI as decision support where one cannot untangle causation from correlation.[2]

5 Conclusion

The FDA is facing some big questions when dealing with the future of ML innovation in medical devices operating on the ground. This discussion aims to scrutinise the proposals and identify whether and how the FDA approach can leverage a patient-centered approach more broadly, considering the challenges of medical diagnostic systems implementing patient autonomy, as well as notions of shared decision-making, and evidence-based medicine. Defining transparency in how medical diagnostic systems stimulate clinical reasoning and patient values is the next step for regulators to ensure the role of medical AI tools as a reliable and safe decision-support on the ground.

[1] Indeed, another area concerns causal inference using observational data when a randomised study is not practicable (Hernán and Robins, 2016).

[2] London (2019, p. 17) uses the example of healthcare professionals prescriping aspirin as a drug that could relieve pain 'for nearly a century without understanding the mechanism through which it works'.

Acknowledgement

The work benefitted from the research undertaken at the UKRI Research Node on Governance & Regulation within the Trustworthy Autonomous Systems programme. The work also benefitted from the author's research stay at the Stanford Center for AI Safety.

References

Arun, N. Gaw, N. Singh, P. Chang, K. Aggarwal, M. Chen, B. Hoebel, K. Gupta, S. Patel, J. Gidwani, M. Adebayo, J. Li, MD. Kalpathy-Cramer, J. (2021). Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence, 3* (6), pp. 1-12.

Beauchamp, TL. Childress, JF. (2019). *Principles of Biomedical Ethics*. Oxford University Press.

Beede, E. Hersch, F. Iurchenko, A. Wilcox, L. Ruamviboonsuk, P. Vardoulakis, L. (2020). A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-12.

Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal, 4* (2), pp.115-130.

Burlacu, A. Iftene, A. Busoiu, E. Cogean, D. Covic, A. (2020). Challenging the supremacy of evidence-based medicine through artificial intelligence: the time has come for a change of paradigms, *Nephrology, Dialysis, Transplantation, 35* (2), pp. 191-194.

Cabitza, F. Zeitoun, J-D. (2019). The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence, *Annals of Translational Medicine, 7* (8), pp. 1-9.

Card, D. Smith, NA. (2020) On Consequentialism and Fairness. *Frontiers in Artificial Intelligence, 3*, pp.1-11.

Chan Jr, DC. Gentzkow, M. Yu, C. (November 2021, revised May 2021). *SELECTION WITH VARIATION IN DIAGNOSTIC SKILL: EVIDENCE FROM RADIOLOGISTS*. www.nber.org/system/files/working_papers/w26467/w26467.pdf.

Christine, PJ. Kaldjian, LC. (2013). Communicating Evidence in Shared Decision Making. *The Virtual Mentor, 15* (1), pp. 9-17.

Cruz Rivera, S. Liu, X. Chan, AW. Denniston, AK. Calvert, MJ. The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, SPIRIT-AI and CONSORT-AI Consensus Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine, 26* (9), pp-1351-1363.

Cuttilo, CM. Sharma, KR. Foschini, L. Kundu, S. Mackintosch, M. Mandl, KD. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency, *NPJ Digital Medicine, 3* (1), pp. 1-5.

Da Silva, M. (8 October 2020 GitHub). *Interpretable Deep Learning Part II: Visual Interpretability with Attribution Methods*. https://metrics-lab.github.io/2020/10/08/visual-interpretability-with-attribution-methods.html.

Douglas, HW. (27 April 2020 MIT Technology Review). *Google's medical AI was super accurate in a lab. Real life was a different story*. https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/.

FDA. (11 April 2018 FDA Press Release). FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.

FDA. (2019). 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. https://www.fda.gov/media/122535/download

FDA. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based: Software as a Medical Device (SaMD) Action Plan. https://www.fda.gov/media/145022/download

FDA., Health Canada., MHRA. (2021). Good Machine Learning Practice for Medical Device Development: Guiding Principles. https://www.fda.gov/media/153486/download

Gerke, S. Babic, B. Evgeniou, T. Cohen, IG. (2020). The need for a system view to regulate artificial intelligence/ machine learning-based software as medical device. *NPJ Digital Medicine, 3* (1), pp. 1-4.

Ghassemi, M. Oakden-Rayner, L. Beam, AL. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet, 3* (11), pp. 745- 750.

Gilbert, S., Fenech, M., Hirsch M., Upadhyay S., Biasiucci A., Starlinger J. (2021). Algorithm Change Protocols in the Regulation of Adaptive Machine Learning-Based Medical Device. *Journal of Medical Internet Research, 23* (10), 1-8.

Glocker, B. Musolesi, M. Richens, J. Uhler, C. (2021). Causality in digital medicine. *Nature Communications, 12* (1), pp- 1-6.

Grote, T. (2021). Trustworthy medical AI systems need to know when they don't know. *Journal of Medical Ethics, 47* (5), pp. 337-338.

Hernán, MA Robins, JM. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology, 183* (6), pp. 758-564.

Holm, S. 2022. Handle with care: Assessing performance measures of medical AI for shared clinical decision-making. *Bioethics, 36* (2), pp- 178-186.

Holzibger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?.*Brain Informatics, 3* (2), pp. 191-131.

Holzinger, A. (2021). Explainable AI and Multi-Modal Causability in Medicine. *De Gruyter Oldenburg, 19* (3), 171-179

Holzinger, A. Carrington, A. Müller, H. (2020). Measuring the Quality of Explanations: The System's Causability Scale. *Künstliche Intelligenz (Oldenburg), 34* (2), pp. 193-198.

Laurie, GT. Harmon, HE. Porter, G. (2016). *Law & Medical Ethics*. Oxford University Press.

Lipton, ZC. (2016). The Mythos of Model Interpretability. *Communications of the ACM, 61* (1), pp. 36-43.

Liu, X. Cruz Rivera, S. Moher, D. Calvert, MJ. Denniston, AK. The SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine, 26* (9), pp. 1364-1374.

Liu, X. Glocker, B. McCradden, MM. Ghaessemi, M. Denniston, AK. Oaken-Rayner, L. (2022). The medical algorithmic audit, *The Lancet, 4* (5), pp. 384-397.

London, AJ. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *The Hastings Center Report, 49* (1), pp. 15-21.

Luxton, DD. (2022). AI decision-support: a dystopian future of machine paternalism?. *Journal of Medical Ethics, 48* (8), pp. 232-233.

McCradden, MD. Anderson, JA. Stephenson, EA. Drysdale, E. Erdman, L. Goldenberg, A. Zlotnik Shaul, R. (2022). A Research Ethics Framework for the Clinical Translation of Healthcare Machine Learning, *The American Journal of Bioethics, 22* (5), pp. 8-22.

Miller, K. (16 March 2021 Stanford Hai). *Should AI Models Be Explainable? That depends*. https://hai.stanford.edu/news/should-ai-models-be-explainable-depends.

Mittelstadt, BD. Allo, P. Taddeo, M. Wachter, S. Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society, 3* (2), pp. 1- 21.

Pearl, J., Mackenzie D. (2018). Mind over Data. *Significance. 15* (4), 6-7.

Prosperi, M. Yi, Guo. Y, Sperrin, M. Koopman, JS. Min, JS. He, X. Xing, R. Shannan, W. Mo, B. Iain, E. Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence, 2* (7), pp. 369-375.

The European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR). (2021). Artificial Intelligence in EU Medical Device Legislation. www.cocir.org/media-centre/publications/article/cocir-analysis-on-ai-in-medical-device-legislation-may-2021.html.

Scott, IA. (2918). Machine Learning and Evidence-Based Medicine. *Annals of Internal Medicine, 169* (1), pp. 44-46.

Seçkin Ayhan, M. Kümmerle, LB. Kühlwein, L. Inhoffen, W. Aliyeva, G. Ziemssen, F. Berens, P. (2022). Clinical validation of saliency maps for understanding deep neural networks in ophthalmology. *Medical Imaging, 77*, pp.1-29.

Vermiere, T. Brughmans, D. Goethals, S. Barbossa de Oliveira, RM. Martens, D. (2022). Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications, 25*, pp. 315-335.

Wu, E. Wu, K. Daneshjou, R. Quyang, D. Ho, DE. Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals, *Nature, 27* (4), pp. 582- 584.