the digital health society
The data arm of ECHAlliance Group

iHD

AI Club

Creating the environment of trust and acceptance that maximises the benefits of AI in healthcare

September 2023

Introduction & Executive Summary

The Digital Health Society (DHS) and the European Institute for Innovation Through Health Data (I~HD) formed the AI Club in September 2022 in order to explore the challenges and opportunities of leveraging the benefit of AI in health and care inspired by attendees at the DHS and I~HD Round Tables and the Calls to Action[1]* The group is formed of experts from academics, policy makers, regulators, industry, life sciences, health systems and policy backgrounds who all work with AI in Healthcare. This paper is intended to inform those working in health and care as to the current global progress to date and the areas where work and collaboration are needed.

The club has been formed to examine the baseline of AI in healthcare and the thematic areas that require further progress via expert consensus. The group have examined the current landscape, especially in light of the pace of development of generative AI and have annotated this in terms of four high priority areas:

Theme 1 - **Use Cases and Evaluation for Healthcare AI**

Theme 2 - **AI Explainability for Citizens**

Theme 3 - **AI Competencies and Education in Healthcare**

Theme 4 - **Safety and Bias in Healthcare**

The Club found that Implementing AI applications in healthcare necessitates rigorous and continuous evaluation, acknowledging that a “one-size-fits-all" is unlikely to work. Evaluations must be at the same time context-specific, application-specific and determined by predefined objectives. Therefore, to optimize AI's utility in healthcare, a methodologically sound, context-sensitive, and historically informed approach is paramount. A focus on creating evidence frameworks for efficacy are required.

Greater investments into research and standards specifying requirements for the design, implementation, evaluation, and documentation of AI explainability (understandability) must be made globally. Legal frameworks and ethical frameworks particularly required to achieve safe scale and pace of deployment.

Both professional groups and citizens require education in order to leverage, understand and trust in AI solutions. Much of the work to date globally has focused on professionals working in health related AI but those working in policy, leadership, and the citizen themselves must not be forgotten.

---

[1] There are 7 Round Table published reports which took place between 2020 and 2023 all focused on health data related matters.https://thedigitalhealthsociety.com/calls-to-action-on-health-data-ecosystems-recommendations-from-multi-stakeholder-round-tables/

AI holds great potential but it must be governed to ensure fairness, privacy, transparency, responsibility and patient safety for it to be trusted by all patients regardless of gender, race, location or demography. Allowing the peer review and open critique of algorithms and the data they are trained upon would inform future developers of AI to be aware of the pitfalls of existing biased training datasets.

In summary the AI club members have identified a need for International collaboration on standards around AI including the governance of AI, safety regimes, empirical evidence frameworks and best practice for practical implementation. It was also noted that competencies and skills are likely to be a rate limiting factor in progress unless opportunities to gain practical skills are created for professionals and citizens alike. It has outlined six recommendations below to enable AI in healthcare to progress more rapidly and safely in a global setting. These actions will need the co-operation of governments, standards bodies and those who work in health policy to drive convergence in AI governance and standards.

In 2024 the United Nations will hold its Summit of the Future and in view of the need for international co-operation this event would be an ideal opportunity to convene in order to make global progress. Our 6 recommendations were presented at the United Nations General Assembly Science Summit in New York on Tuesday 26th September 2023.

Rachel Dunscombe

September 2023

For More information about the Digital health Society and the European Institute for Innovation Through Innovation see here and here.

Recommendations

The AI clubs's recommendations have implications at a global, national and regional level. Delivering these recommendations will require policy makers, non governmental organisations, health systems, academia, the technology industry and life sciences to collaborate to create a conducive environment for safe healthcare AI to deliver benefit. Where possible these recommendations should be undertaken at scale and there is a need for global initiatives led by international organisations. Recommendations such as those regarding skills may be undertaken at country level.

Each of the recommendations will have short, medium and long term endeavours associated with them as the landscape further matures in health and care AI. The creation of this landscape must emerge at a pace that allows health systems to benefit from the technologies but mitigates risks associated with technological progression.

The following recommendations have been identified as the key initiatives required to allow the safe global progress of AI in health at scale and pace :

R1. Creation of a global framework for standardised but flexible evidence-based impact and process evaluation for AI in health that also allows global comparison and continuous monitoring of AI performance in real world deployment.

R2. Creation of a global evidence based evaluation framework for risk and opportunity for use cases that allows for local sensitivities such as demographics and healthcare priorities and personalisation and including ethics.

R3. Creation of a global standard to define and implement explainability that meets the needs of all stakeholders including the citizen, professionals and leaders.

R4. Development of evidence based competency frameworks for the skills required by the health and care workforce and national alignment to education interventions needed to achieve them.

R5. Development of an exemplar global health literacy skills framework for the citizen to empower them to engage with use of AI in their health system and other related services.

R6. Creation of a global data transparency framework ensuring AI training data is initially fit for purpose, continues to be fit for purpose, can be benchmarked globally and mitigates bias.

AI Club Members

Professor Rachel Dunscombe (Chair), CEO OpenEHR, Visiting Professor Imperial College London, and Director of the Digital Health Society.

Dr. Richard Baumgartner is a Senior Principal Scientist with Biometrics Research Department, Biostatistics and Research Decision Sciences, Merck and Co.

Sara Boltman Founding Director Butterfly Data, Data Scientist, and lead for Theme 4 Safety & Bias in Healthcare AI.

Dr Kathrin Cresswell, Senior Lecturer, Usher Institute, University of Edinburgh, Old Medical School, Teviot Place, Edinburgh and lead Theme 3 AI Competencies and Education in Healthcare.

Elisabeth Gatti, Policy Analyst Digital Health & Taxation Advocacy EMEA Johnson & Johnson.

Professor Jordi Piera-Jiménez, Director of the Digital Health Strategy for Catalonia, Catalan Health Service, Barcelona, Spain, Digitalization for the Sustainability of the Healthcare System DS3-IDIBELL, L'Hospitalet de Llobregat, Spain, Faculty of Informatics Multimedia and Telecommunications, Universitat Oberta de Catalunya, Barcelona, Spain.

Gordon Johnston, Digital Health & Innovation, Engineer at Johnson & Johnson.

Professor Dipak Kalra, President, the European Institute for Innovation through Health Data, Visiting Professor at the University of Gent, Belgium and lead Theme 2 the Importance of AI Explainability for Citizens.

Dr Mehdi Khaled, Managing Partner SEHA and lead for Theme 1 Application Fields and the Need for an Evidence-Based and Theoretically Informed Evaluation Framework.

Jesper Kjaer, Director of Data Analytics Centre Danish Medicines Agency.

DR Nathan Lea, Senior Research Associate UCL and Information Governance Lead at I~HD.

Angel Martin, Senior Director Digital Health & Taxation Advocacy EMEA Johnson & Johnson.

Stuart McCann, Regional Digital Business Manager Roche.

Bleddyn Rees, Chair Digital Health Society & Deputy Chair ECHAlliance

Dr Isobel Taylor, Consultant Butterfly Data

Professor Paul Timmers, Research Associate Oxford Internet Institute, Professor KU Leuven, European University Cyprus. CEO iivii BV.

Dr Philip Webb, CEO Respiratory Innovation Wales.

Other Club Members took part in four quarterly meetings but did not take part in the theme working groups. Rachel, DHS and I~HD would like to acknowledge the considerable contributions made by all AI Club members to the meetings and this report. Individuals attended in a personal capacity so the views of this report do not necessarily reflect the views of any of their employing organisations. This report is in the form of consensus papers from the 4 Themes convened and edited neutrally and independently by DHS and I~HD. The AI Club will continue its work in 2024.

Index

THE EXPANDING ROLE OF ARTIFICIAL INTELLIGENCE IN HEALTH AND CARE:

Theme 1 - Applications Fields and the Need for an Evidence-based and Theoretically-informed Evaluation Framework

Sub Group: Dr. Mehdi Khaled, Dr. Kathrin Cresswell, Bleddyn Rees, Philip Webb, Professor Jordi Piera-Jiménéz

The proliferation of Artificial Intelligence (AI) promises transformative potential within health and care, most notably within clinical and administrative domains. Due to the emergent properties of AI and the fast-changing development and implementation landscape, new and agile approaches to evaluation are needed. However, these approaches need to be both evidence-based and theoretically-informed, in order to ensure that they build on and learn from existing experiences.[2] Grounding evaluation methodologies in sound theoretical foundations ensures that potential risks, benefits, and ethical implications are comprehensively considered. Marrying empirical evidence with theoretical insights allows stakeholders to make informed decisions, guide investments, foster transparent AI implementation, and drive the continuous improvement of AI-driven healthcare solutions that prioritize patient well-being while navigating the complexities of modern medical practice. Overall, evidence-based evaluation frameworks can help to guide these efforts. They can, for example help to identify risks to adoption and scaling of systems proactively and thereby maximise chances of successful implementation.

An evaluation framework can also help to assess the performance, impact, and ethical considerations of AI systems, encompassing criteria such as accuracy, fairness, transparency, safety, and interpretability.

In this chapter, we first delve into AI's distinct use-cases within healthcare, exploring both clinical and non-clinical applications. Afterwards, we outline components that we believe a comprehensive AI evaluation framework in healthcare needs to include.

[2] Cresswell K, Rigby M, Georgiou A, Wong ZS, Kukhareva P, Medlock S, De Keizer NF, Magrabi F, Scott P, Ammenwerth E, Williams R. The need to strengthen the evaluation of the impact of Artificial Intelligence-based decision support systems on healthcare provision. Health Policy (in press).

1. AI in Clinical Settings

AI solutions in healthcare have the potential to act as clinical decision support tools, leveraging their computational prowess to assist healthcare professionals in making more informed clinical decisions. These AI-powered tools, ranging from diagnostic aids, treatment recommendation systems to patient monitoring and predictive analytics, which may help to elevate the quality of care while enhancing patient safety. However, these systems can also have unintended consequences and the variation of contexts and integration into existing workflows is not well-understood.

These AI tools have the ability to ingest vast amounts of patient data, encompassing medical records, imaging studies, genetic profiles, and real-time sensor data. By applying advanced machine learning and pattern recognition techniques, AI can identify subtle correlations and trends within this data that might elude human perception. This ability empowers healthcare practitioners to detect potential diseases earlier, predict patient outcomes, and personalize treatment plans based on an individual's unique characteristics.

Furthermore, AI decision-support tools can augment healthcare professionals' expertise by providing evidence-based recommendations from extensive medical literature and clinical guidelines. This can help to assist in reducing diagnostic errors, minimizing variability in treatment approaches, and enhancing overall clinical efficiency.

In an ideal situation, integrating AI into clinical decision-making does not replace human judgement but amplifies it. It offers a symbiotic relationship where AI provides data-driven insights and suggestions, allowing healthcare professionals to make well-informed choices. This synergy is crucial in complex cases, rare diseases, and scenarios where up-to-date medical knowledge is essential.

Through the following use cases, we will illustrate the complexities surrounding AI implementation and adoption, which we hope will help us to show that we need better approaches to evaluation.

Use Case 1: AI in Diabetic Retinopathy Detection

Recent studies illustrate the potential of deep learning algorithms in detecting diabetic retinopathy, a leading cause of blindness worldwide [3]. Training AI models with large datasets of retinal images, researchers have achieved diagnostic accuracy rivalling that of human ophthalmologists. This use-case demonstrates how AI can enhance patient care through early detection, thereby improving prognosis and reducing the burden on healthcare providers.

[3] Gulshan V, et al. (2020). Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India. JAMA Ophthalmology. 138(9), 945-953.

However, potential risks include misdiagnosis, which, although statistically low, remains a significant concern [4]. Ensuring the model's accuracy requires continuous validation and monitoring, which can be resource-intensive. Moreover, integrating such AI systems in resource-poor settings can be challenging due to technological constraints and limited accessibility.

Use Case 2: AI in Suicide Risk Predictions

AI algorithms have made notable strides in mental health, specifically in identifying patients at risk of suicide. By analysing various factors, these AI systems provide a novel approach to suicide risk assessment, enabling early interventions and potentially saving lives [5].

Nevertheless, such applications are not without risks. The algorithms' accuracy, particularly concerning false positives and negatives, is critical due to the severe implications of missed or unnecessary interventions [6]. Ethical considerations surrounding the handling of sensitive mental health data pose additional challenges. Safeguarding patient privacy and ensuring data security is paramount and necessitates robust data management practices.

AI's role in clinical healthcare can be transformative, when evidenced by its substantial contributions to disease detection and diagnosis. However, alongside its myriad benefits, the application of AI still presents potential risks and challenges. Effective integration of AI in clinical settings requires a thorough understanding of these factors and the development of robust evaluation frameworks to ensure optimal patient care.

2. AI in Non-Clinical Applications

AI is demonstrating transformative potential not only within the clinical domains but also in non-clinical, operational, and administrative areas of healthcare. We explore AI's specific non-clinical use cases in healthcare, shedding light on the associated benefits, risks, and challenges inherent to each.

Use Case 1: AI in Operating Room Scheduling

AI has been employed effectively in the management of Operating Room (OR) schedules, resulting in optimised patient flow and resource utilization [7]. Here, machine learning

[4] Topol EJ. (2020). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 25(1), 44-56.

[5] Walsh CG, et al. (2020). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. Clinical Psychological Science. 6(3), 398-411.

[6] Cheniaux E, et al. (2020). Sensitivity and specificity of the Zung self-rating depression scale for depression diagnosis in outpatients. Clinical Neuropsychiatry. 17(1), 32-38.

[7] Epstein RH, Dexter F. (2020). Optimizing Sequences of Operating Room Start Times for Surgery Centers. Journal of Medical Systems. 44(1), 1-9.

algorithms analyse multiple variables including surgical durations, patient recovery times, and surgeon preferences, among others, to devise optimal schedules. Such intelligent scheduling can significantly enhance operational efficiency, minimize patient wait times, and lead to improved patient satisfaction [8].

However, the risk of inaccurate predictions could disrupt OR schedules, leading to inefficiencies. Furthermore, staff resistance to AI-induced changes and the need for continuous training pose challenges to successful implementation and utilization of such systems.

Use Case 2: AI in Managing Outpatient No-Shows

AI has made noteworthy contributions to outpatient services, particularly in predicting appointment no-shows. Employing predictive analytics, AI helps identify patients who are likely to miss their appointments, allowing for effective countermeasures like reminder systems or schedule adjustments, consequently saving costs and improving operational efficiency [9] [10].

Despite these benefits, predictive models might not always be accurate, leading to potential scheduling disruptions. The risk of over- or under-booking, reliance on the quality and comprehensiveness of input data, and potential privacy concerns pose significant challenges [11].

Use Case 3: Operations Research and AI in Healthcare Delivery

Operations Research (OR), characterized by the application of mathematical methods to decision-making, has been enhanced by the integration of AI, leading to more effective healthcare delivery. AI-supported OR has facilitated solutions to complex problems such as resource allocation, patient flow management, and service delivery optimisation [12].

However, developing and implementing OR models that integrate AI can be complex and resource-intensive. Inaccurate predictions or unsuitable models could lead to suboptimal decision-making. Moreover, successful implementation requires a high level of

[8] Van Houdenhoven M, et al. (2020). Fewer intensive care unit refusals and a higher capacity utilization by using a cyclic surgical case schedule. Journal of Critical Care. 24(2), 304-9.

[9] Huang Y, Hanauer DA. (2020). Patient no-show predictive model development using multiple data sources for an effective overbooking approach. Applied Clinical Informatics. 4(3), 367-76.

[10] Valero-Bover, D., González, P., Carot-Sans, G. et al. Reducing non-attendance in outpatient appointments: predictive model development, validation, and clinical assessment. BMC Health Serv Res22, 451 (2022).

[11] Zheng K, et al. (2021). Web-Based Just-in-Time Information and Feedback on Antibiotic Use for Village Doctors in Rural Anhui, China: Randomized Controlled Trial. Journal of Medical Internet Research. 20(2), e57.

[12] Van Essen JT, et al. (2021). Reducing waiting time and raising outpatient clinic efficiency using computer simulation. Archives of Disease in Childhood. 101(7), 620-625.

multidisciplinary cooperation and data sharing, which could be challenging in daily practice [13].

3. Processes and Impact:

An evidence-based and theory-informed evaluation framework is essential to harness the potential of AI in health and care effectively. This framework should encompass evaluating impacts (e.g. effectiveness, impact on patient outcomes, cost-effectiveness and practitioner/organisational performance), and processes (e.g. integration with work practices and organisational functioning, unintended consequences, scalability and transferability).

The cornerstone of AI process evaluation is the systematic and comprehensive assessment of the procedures, mechanisms, and strategies used during the AI system's development, implementation, and functioning. This involves examining the data collection and pre-processing methodologies, the choice and training of algorithms, the validation and testing processes, the interpretation of results, and the broader deployment and adoption context. Ensuring transparency, fairness, robustness, and accountability throughout these stages is critical to gaining trust and ensuring the successful integration of AI systems in real-world applications. It may also include exploring:

1. Integration: AI's ability to integrate into existing healthcare systems and workflows is a crucial part of the evaluation. Consideration should be given to the need for user training and the adaptability of the AI system to ensure a smooth transition [14].

2. Workforce impact: The impact on the healthcare workforce has illuminated both transformative benefits and significant challenges. On the positive side, AI-driven tools have the potential to automate routine tasks, enhance diagnostic accuracy, and personalize treatment plans, thereby augmenting the efficiency and effectiveness of healthcare professionals. This augmentation can lead to reduced workload and human error, fostering a more patient-centered approach. However, there's also a concern that AI could displace certain job functions, necessitating workforce retraining and a shift in job roles. Furthermore, the introduction of AI tools requires healthcare professionals to acquire new skills in data interpretation and technology interaction. Overall, while AI promises to reshape the healthcare landscape, its impact on the workforce necessitates careful planning and continuous education.

AI impact evaluation may be assessed across the following domains:

[13] Royston G, et al. (2021). Using system dynamics to help develop and implement policies and programmes in health care in England. System Dynamics Review. 18(3), 373-385.

[14] Shortliffe EH, Sepulveda MJ. (2020). Clinical Decision Support in the Era of Artificial Intelligence. JAMA. 322(23), 2283-2284.

1. Effectiveness: The effectiveness of AI systems is typically evaluated using validation datasets. Assessing AI system accuracy, sensitivity, and specificity is key. It is also crucial to evaluate AI performance within real-world healthcare contexts for a more accurate understanding of their potential [15].

2. Patient impact: The impact of AI on patient outcomes needs to be evaluated but this is difficult to evidence. There is a growing body of evidence highlighting the positive influence of AI on diagnosis, treatment, and overall patient care, further emphasizing the importance of continuous monitoring and assessment [16].

3. Cost-effectiveness: Economic implications of AI adoption, including potential cost savings through improved efficiency and better resource allocation, should be considered. Cost-effectiveness analyses are vital in providing insights into the economic viability and potential return on investment of AI interventions [17].

4. Ethical implications: Ethical considerations, including data privacy, informed consent, and algorithm transparency, should be integral to evaluation . In the digital age, concerns surrounding data security and privacy are paramount, necessitating the assessment and addressing of these ethical issues [18].

5. System optimisation: Evaluating AI's potential to optimise system performance and improve healthcare delivery is essential. It enables an understanding of AI's capability to drive systemic changes and lead to improved healthcare outcomes [19].

6. Liability: Liability considerations are increasingly becoming crucial in AI evaluation frameworks. In case of errors or harm caused by AI applications, determining responsibility can be complex. Guidelines need to be in place to ensure accountability and address potential legal implications [20].

4. Tensions surrounding the evaluation of AI in healthcare

In relation to evaluation, various interest groups hold specific agendas, and it is essential to define what specific evaluation activities aim to achieve and identify the intended audience. Audiences may include:

- Suppliers: these are motivated to showcase the effectiveness and impact of their systems to facilitate sales and market their products.

[15] Fleuren LM, et al. (2020). Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Medicine. 46, 383–400.

[16] Topol EJ. (2020). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 25(1), 44-56.

[17] Verghese A, Shah NH, Harrington RA. (2020). What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. JAMA. 323(1), 29-30.

[18] Coiera E, et al. (2020). The Digital Scribe. NPJ Digital Medicine. 1, 58.

[19] Bates DW, Auerbach A, Schulam P, Wright A, Saria S. (2020). Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence. Annals of Internal Medicine. 172(11_Supplement), S137-S144.

[20] Price WN 2nd, Gerke S, Cohen IG. (2020). Potential Liability for Physicians Using Artificial Intelligence. JAMA. 322(18), 1765-1766.

- Governments seek to understand the benefits and returns on investments made in digitalization. They conduct evaluations to build business cases, improve processes, and manage risks.
- Service delivery organizations are interested in assessing the benefits and returns on their investments to develop business cases. They may also use process evaluations to refine their workflows and mitigate risks.
- Service users and professionals may wish to know where healthcare organizations stand concerning digitalization. They may also wish to understand the impact of digitalization on healthcare processes and outcomes.
- Patients and Patient organisations striving to improve clinical outcomes

Several key tensions exist within the AI evaluation landscape:[21] [22]

- The rapid evolution of AI technology surpasses the pace at which evaluations can keep up. Policy makers and organizations require evidence to make informed procurement decisions but this evidence takes time to gather and is expensive. To address this, new pragmatic evaluation and agile methods are necessary but these need to be evidence-based and theoretically-informed to ensure rigour.
- Realizing benefits from complex digitalisation initiatives can take a substantial amount of time, sometimes up to decades. Measuring these benefits within the scope of short-term digitalisation projects (which are the norm in AI) is therefore challenging.
- Contextual variations can significantly influence evaluation outcomes, making it vital to consider and account for different contexts when conducting evaluations. There is, however, a trade-off between capturing contextual differences and extracting common lessons across contexts.
- Evaluators are often called in only when issues arise, resulting in missed opportunities to assess and learn from ongoing processes. This limited learning approach hampers the potential for improvement and learning.
- Establishing effective baselines is crucial, but often overlooked. The transformation of systems and organizational processes makes before-after comparisons challenging. Nevertheless, such assessments are important to demonstrate progress, as it is akin to comparing apples with oranges.
- Identifying the benefits to assess can be difficult, especially when unexpected benefits occur in unforeseen areas. The current focus lies on efficiency and time

[21] Cresswell K, Sheikh A, Franklin BD, Krasuska M, Hinder S, Lane W, Mozaffar H, Mason K, Eason S, Potts HW, Williams R. Theoretical and methodological considerations in evaluating large-scale health information technology change programmes. BMC Health Services Research. 2020 Dec;20(1):1-6.

[22] Cresswell K, Williams R, Sheikh A. Developing and applying a formative evaluation framework for health information technology implementations: qualitative investigation. Journal of medical Internet research. 2020 Jun 10;22(6):e15068.

saved, but the evaluations also needs to measure impact on quality and safety. Additionally, impact studies can be expensive, and directly attributing effects to specific factors is challenging in complex systems such as AI.

5. Conclusion

Implementing AI applications in healthcare necessitates rigorous and continuous evaluation, acknowledging that a “one-size-fits-all" is unlikely to work. Historical failures and successes underscore the criticality of leveraging established theoretical frameworks in the assessment process. Evaluations must be at the same time context-specific, application-specific and determined by predefined objectives. These assessments should consider both the resultant impact and the procedural aspects of the AI application and adoption, ensuring relevance to individual contexts. Therefore, to optimize AI's utility in healthcare, a methodologically sound, context-sensitive, and historically informed approach is paramount.

Theme 2 - **The importance of AI explainability for citizens**

Sub Group: Professor Dipak Kalra, Stuart McCann, Angel Martin, Elizabeth Gatti, Richard Baumgartner

1. The importance of trust

Patients and healthy citizens will increasingly be exposed to AI contributing to the delivery of their healthcare, such as the early detection of risks and deteriorations, the personalisation of their treatments and for learning through their data to advance healthcare for the future. They might be the direct users of software or devices that are controlled by or incorporate AI, for example purchasing prevention apps, using monitoring or disease management tools issued by their healthcare provider, or indirect users through discussion with their clinician about diagnostic or treatment recommendations arrived at through AI tools in the consulting room.

It is therefore important for patients and citizens to be in a position when they can know the extent to which they can trust and rely upon AI reasoning and AI outputs, irrespective of whether they are direct users or not. They need to know why AI is being used in their care and how it applies to them personally, what its influence is and the overall impact of AI solutions for healthcare and for research[23,24]. People also need to know how and by whom their health data will be used and safeguarded when AI is being developed and improved.

Trust, acceptance and explainability are interdependent.

The importance of explainable or interpretable AI has been articulated for many years, in order to counter the very real challenge, if not impossibility, for most users to understand how AI reasoning has arrived at its output. This differs from classical clinical decision support, which uses algorithms that are static, almost always implementing published evidence and traceable to that evidence. There is a strong obligation, for example recently enforced in Europe through EU AI Act, for developers to provide explainable AI, along with transparency, and to adhere to other ethical principles in addition to privacy and security standards as part of regulatory compliance.

[23] Medrano I. Artificial Intelligence in healthcare: Separating facts from fiction. Healthcare Transformers 16th August 2023. Available at https://healthcaretransformers.com/digital-health/artificial-intelligence/ai-precision-medicine-facts-fiction/

[24] McKendrick J. Healthcare May Be The Ultimate Proving Ground For Artificial Intelligence. Forbes 22nd February 2023. Available at https://www.forbes.com/sites/joemckendrick/2023/02/22/healthcare-may-be-the-ultimate-proving-ground-for-artificial-intelligence/?sh=4755c9ef2b55

However, the discipline of explainability is still evolving. One aspect of explainability is towards a computer literate expert who may need to work with an AI component within a larger software or hardware solution. A greater challenge is to explain AI as a methodology, and specific AI solutions, to patients and citizens whose health and care is going to be influenced by these products. Being able to provide clear and meaningful explanations of an AI system's outcomes is crucial to building and maintaining users' trust[25]. Explainability needs to be simple in describing the AI component's purpose, its process steps and fit within the healthcare delivery continuum, and its validation/safety evidence. The value of the AI also needs to be part of its explainability: the impact on a patient receiving a clear diagnosis in an efficient and timely manner from the AI tool and the impact of clinical decision-making.

This is where effort is urgently needed to develop good practice and to put this into practice.

Whilst the focus of this paper is on explainability to citizens, health professionals also need to understand where AI is working inside their digital ecosystems, why they should be confident to rely upon its outputs and convince their patients to be similarly confident, how to use AI-based solutions responsibly and when to know that they should not follow its advice. OECD identified the key concerns regarding a lack of trust in AI reported by health professional associations to be the black box and evolving nature of AI systems, that algorithms can reflect the biases implicit in their training data, and that there are tensions between the potential to improve accuracy and reduce biases through greater data access versus the importance of protecting personal data [26].

This last area of tension is not only held by health professionals, but by data protection officers and other stakeholders responsible for health data. It is important to address, because better access to fine grained health data will improve the accuracy and safety of AI, and reduce the problems associated with bias and inequity. Understanding the benefits of AI in general, and of each proposed new solution, is therefore essential to countering negative perceptions and enlisting greater stakeholder support for health data use for training and validation.

Whilst focusing here on citizens, a greater effort towards explainability to multiple non-technical stakeholders (users and decision makers) will therefore improve the quality and value of AI as well as increasing trust in its use and in its outputs. It must be recognised, though, that value is perceived differently by different stakeholders, who must all gain their dimension of value in order for AI solutions to thrive.

---

[25] Longo, L., Goebel, R., Lecue, F., Kieseberg, P., Holzinger, A. (2020). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In: Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science(), vol 12279. Springer, Cham. https://doi.org/10.1007/978-3-030-57321-8_1

[26] Socha-Dietrich K. Empowering the health workforce to make the most of the digital revolution. OECD Health Working Paper No. 129, 2021. OECD Publishing, Paris. https://doi.org/10.1787/37ff0eaa-en

2. What is explainability?

Explainability is the ability to accurately describe the mechanism, or implementation, that led to an algorithm's output. Interpretability refers to whether a human can derive meaning from a system's output for a specific use case [27]. Transparency, in contrast, relates to communicating to users of a system that its action or recommendation is based on or has used AI. Transparency does not equate to communicating how the decision was arrived at (interpretability) or how the AI performs its reasoning (explainability).

Several high-level policy-setting bodies have published ethical principles for the development and use of AI. Whilst these do not all use the above three terms in the same ways, they all call for developers to adopt practices that embrace these concepts.

The OECD Principles for responsible stewardship of Trustworthy AI on Transparency and explainability, adopted by the G20 nations at a meeting of Trade Ministers and Digital Economy Ministers in Japan in 2019 [28], states:

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

> **i.** *to foster a general understanding of AI systems;*
>
> **ii.** *to make stakeholders aware of their interactions with AI systems, including in the workplace;*
>
> **iii.** *to enable those affected by an AI system to understand the outcome; and,*
>
> **iv.** *to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.*

The WHO includes transparency, explainability and intelligibility as one of its six principles, published in 2021, that should serve as the basis for AI regulation and governance [29]:

> *Ensuring transparency, explainability and intelligibility: transparency requires that sufficient information be published or documented before the design or deployment of an AI technology. Such information must be easily accessible and facilitate*

[27] Schwartz, R. et al. (2021), Proposal for Identifying and Managing Bias in Artificial Intelligence (SP 1270), https://www.nist.gov/artificial-intelligence/proposal-identifying-and-managing- bias-artificial-intelligence-sp-1270

[28] Principles for responsible stewardship of trustworthy AI, ratified by the G20 Trade Ministers and Digital Economy Ministers, Japan, June 2019. Endorsing OECD Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. OECD 2022

[29] World Health Organisation. WHO issues first global report on Artificial Intelligence (AI) in health and six guiding principles for its design and use. WHO 2021. Available at: https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use

meaningful public consultation and debate on how the technology is designed and how it should or should not be used.

The EU Ethical Guidelines (2019) include transparency and explainability within its seven key requirements that AI systems should meet in order to be deemed trustworthy [30]:

The data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

3. Formalising explainability

Although the many published reports from national and multinational policy setting bodies, including those examples above, stress the importance of explainability and transparency, these largely express the concepts and what they mean at a relatively high level. They are not precise enough to enable a developer of AI to know with confidence how to conform to those expectations, nor to enable an assessment body or an adoption decision-maker to have confidence that good practices have been followed. A literature review conducted last year by Kargl et al concluded that there remains a need to make AI ethics more tangible for practical implementation, to include practically useful formulations and explanations of AI principles (also specifically for the biomedical domain), the development of concrete requirements for AI systems, and standardization initiatives [31].

Amongst the most structured specifications of good and ethical AI development practice are the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, published in 2020 by a High-Level Expert Group on Artificial Intelligence appointed by the European Commission[32], and the DECIDE-AI checklist[33]. The former checklist focuses in particular upon development practices that enable compliance to the EU Ethics Guidelines mentioned above. The latter covers scientific as well as ethical good practices, including methodologies for generating evidence of the accuracy and clinical effectiveness of the AI algorithm. These two checklists decompose broad concepts like explainability and transparency into contributing components, presented as headings with a definition of what these each mean and require.

---

[30] The European Commission High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI, 2019. Available from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[31] Kargl M, Plass M, Müller H. A Literature Review on Ethics for AI in Biomedical Research and Biobanking. Yearb Med Inform. 2022 Aug;31(1):152-160. doi: 10.1055/s-0042-1742516

[32] The European Commission High-Level Expert Group on Artificial Intelligence. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, 2020. Available https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[33] Vasey B et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022;377:e070904

Another well structured analysis, in this case focused on trustworthiness, was published in 2023 from the Center for Long-Term Cybersecurity at UC Berkely[34]. This report examines the characteristics of trustworthiness across the seven life-cycle points of AI development that are defined in the NIST AI Risk Management Framework [35]. These characteristics, including "explainable and interpretable", "with harmful bias managed", "accountable and transparent", are elaborated as multiple self-assessment questions for each of the seven life-cycle points. Although covering more than explainability, this report probably contains the most comprehensive and fine grained set of checklist questions that a developer should answer in order to assure and demonstrate the explainability of their AI solution.

The Coalition for Health AI (CHAI)[36] is a predominantly US collaboration of academic, healthcare and governmental agencies developing guidelines for reporting AI development and responsible use, including a focus on equity and bias. Its 2023 Blueprint for trustworthy AI implementation guidance and assurance for healthcare[37] sets out an agenda to develop an assurance lab, registries of AI developments and a value demonstration framework, all of which will contribute to trusted future adoption.

4. Main components of explainability

The explainability of AI can probably be broadly considered as comprising two kinds of explanation:

1. The AI developer community needs to document and share within its community the computer modelling methodologies that have been utilised, how the algorithm has been developed and validated, what kinds of data have been used as initial training data and as testing/validation data, and what development quality processes have been used.

2. The adopting community, including approvals decision-makers such as regulators and HTA, decision-makers about utility and value such as HTA, health ministries and health insurers, and end users such as health professionals and patients, all need a kind of "understandability", which needs to be non-technical and related to health and care needs, services, care pathways, safety and effectiveness including impact on health outcomes.

The former category of explainability is rapidly becoming mature, reflected in standards such as the AI ethical principles and the EU AI Regulation. There is a progression towards harmonised ways of describing this kind of explainability.

---

[34] Newman J. A taxonomy of trustworthiness for artificial intelligence. Center for Long-Term Cybersecurity, UC Berkely, 2023. https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/

[35] National Institute of Standards and Technology (NIST). AI Risk Management Framework, 2023. Available from https://www.nist.gov/itl/ai-risk-management-framework

[36] Coalition for Health AI (CHAI). Please see https://www.coalitionforhealthai.org/

[37] CHAI. Blueprint for trustworthy AI implementation guidance and assurance for healthcare. 2023. Available from https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf.

The latter category of adopter understandability remains variable in the extent to which this is covered as part of overall explainability. There is as yet no consensus good practice that enumerates the aspects of explanation that would best enable user understanding, nor yet a consensus on the information needs of different adopting stakeholders to instil confidence in using the AI and in relying upon its advice.

Examples of the adopter and end user understandability characteristics that are detailed in the structured checklists outlined above, which patients, citizens, health professionals and decision makers are likely to require, include the following.

- A clear statement of the health and care objective: the health condition being addressed, the challenging clinical or patient decision the AI helps with, including if its intended role is diagnostic, risk or care pathway stratification, personalisation of treatment, early detection of the need for care escalation etc.

- The patient profiles of the data that was used for AI development and for validation, which should largely dictate the scope of patient populations on whom there is likely to be reliable evidence of its safety and effectiveness, such as the age range, ethnicity, geography, health condition(s), severity, kinds of treatment included etc.

- Assurance that the AI been trained on good quality and unbiased data, how quality, bias and representativeness (equity) have been assessed and what mitigations and corrections have been applied (or recommended limitations of use) to compensate for biases that could not be eliminated.

- The degree of autonomy of the software, device, robotics incorporating the AI algorithm e.g. providing advice to the clinician or patient, issuing an alert or warning, taking an action or controlling an instrument such as a medication delivery closed loop system, and if its advice is normally going to be co-interpreted with other decision influencing information that a clinician will utilise in order to arrive at a final decision. It will be preferable to adopt a formal classification of autonomy such as that proposed by Bitterman et al [38].

- What approvals have been obtained per jurisdiction, such as EU Medical Device Regulation certification and AI Regulation certification, HTA approvals.

- What evidence has so far been accumulated about patient safety, clinical effectiveness, impact on patient outcomes and health economic value.

- Clear guidance to healthcare organisations about how to deploy the AI-containing solution including what input data flows (e.g. EHR data) it will require to perform its reasoning, the format of its outputs and how these may be audit logged and persisted by the adopting organisation, and what data flows are needed back to the developer to continue the machine learning cycles.

[38] Bitterman D, Aerts H, Mak R. Approaching autonomy in medical artificial intelligence. The Lancet 2020:2(9);E447-E449. https://doi.org/10.1016/S2589-7500(20)30187-4

- Clear guidance to healthcare organisations about how to use the AI solution within care pathways, use by whom, what aspects of training they will need, use in which patient situations, and how to override the AI output if it has some degree of autonomy.
- Where liability and accountability lie when users follow AI advice or give it serious weight in their decision-making but the advice proves to have been incorrect, or conversely what liability would exist for users choosing not to follow AI advice if it subsequently transpires that it would have been correct.

5. Conclusion and a call to action

There are huge potential benefits to healthcare from AI-driven applications and devices. There is a need for transformation towards quicker, smarter, faster diagnostic decision making = right patient, right time, receiving the right treatment = higher probability of better patient outcomes = positive impact on the wider patient journey and healthcare ecosystem.

To scale, AI has undoubtedly to be better understood by all actors across health and care, especially the public. Users of explainable AI systems benefit from being able to understand and challenge or contest an outcome, seek redress, and learn through human-computer interfaces[39].

The European Patients' Forum (EPF) conducted an investigation amongst its members (patient advocacy organisations across Europe) into the patient perspective on AI[40]. This involved two webinars, a survey and sixteen interviews undertaken during 2021. Their main recommendations were:

1. *Greater advocacy is needed to ensure patient and staff involvement in AI development, research, and policy projects that address the needs of these groups.*
2. *Stakeholders could use clearer and more practical guidance about whether, when and how to involve patients, and other health system stakeholders as well as how to mobilize the appropriate stakeholders to guide and implement the project.*
3. *Training patients and staff in AI principles is not a standard practice currently and there is a need for expert guidance on the curriculum that would be most helpful to specific projects.*
4. *Involving patients early in AI development projects is critical to ensuring the projects' vision, scope and requirements are rooted in the needs and perspectives of the people intended to benefit from, and use them.*

[39] OECD Digital Economy paper 349. Advancing accountability in AI, 2023. Available from https://www.oecd-ilibrary.org/science-and-technology/advancing-accountability-in-ai_2448f04b-en

[40] del Castillo J, Nicholas L. Artificial Intelligence in Healthcare from a Patient's Perspective. European Patients' Forum, 2022. Available from https://www.eu-patient.eu/globalassets/report-ai-1612---del-castillo-and-nicholas-2.pdf

To support these EPF recommendations, exemplary use cases spanning different healthcare systems and care pathways need to be collected and disseminated, with evidence of positive experiences and benefits: how well has the AI worked, why and what was the outcome?

AI adopters need education and guidance on the current standards and practices they should expect and require conformance to, and on how best to implement the AI in their context.

To these recommendations, it now seems vital to add a call for greater investments into research and standards specifying requirements for the design, implementation, evaluation and documentation of AI from an explainability (understandability) perspective, ideally at a global level.

Theme 3 - **AI competencies and education in healthcare**

AI Club | Sub-group: Dr. Kathrin Cresswell, Jesper Kjær, Bleddyn Rees, Professor Rachel Dunscombe

1. Competencies and Education Summary

There is currently a significant digital skills shortage in healthcare settings. Artificial Intelligence (AI) competencies in particular are lagging behind, as the area is emerging and rapidly evolving. The development of digital competencies in AI is important for a variety of stakeholders so that emerging applications are developed, implemented, used, and scaled appropriately and informed by scientific evidence. Previous research has shown that, if this is not done, there are important risks associated with the adoption of new technologies (such as AI), which can lead to adverse consequences in relation to safety, quality, and efficiency of care. We here give a broad overview of existing competency frameworks developed for AI in healthcare, identify gaps, and suggest ways forward to address these gaps. We conclude that, although competency frameworks building on digital skills have been developed for AI, the focus has to date been on particular stakeholder groups (mainly healthcare professionals). Other stakeholders, including patients, citizens, implementers (i.e. those responsible for putting a system into organisational environments), policy makers and developers have to date been neglected.

2. Concepts and definitions

Competence in healthcare has been defined as *"the ability of the practitioner to practise safely and effectively to a professional standard"* (page 4).[41] It goes beyond mere fulfilment of a particular role to include critical appraisal of evidence, solving problems, and working in a multi-disciplinary team.

[41] Storey L, Howard J, Gillies A. Competency in healthcare: A practical guide to competency frameworks. Radcliffe Publishing; 2002.

Competency frameworks can help to define and baseline competencies, to evaluate progress, and to develop competencies in specifically tailored training programs.[42] Here, tailoring to context is crucial. This may involve considering various stakeholder groups with different needs and existing skill levels, and also various settings (e.g. hospitals have very different requirements to primary care).[43]

3. Existing competency frameworks surrounding AI in healthcare

There are many different competency frameworks for digital skills in healthcare,[44] [45] [46] on which current efforts to develop AI competency frameworks are based. The general approach appears to be on helping healthcare professionals to develop an understanding what AI is and what it does, and on considering its limitations in various contexts (from direct clinical care to organisational and population-based performance). A more detailed understanding of these dimensions tends to represent higher levels of competency. Most frameworks focus on specific stakeholder groups and specific geographical areas, although some emerging international frameworks with the aim to represent a variety of contexts are currently under development[47] [48]

The majority of existing competency frameworks around AI in healthcare have been developed for medical education purposes.[49] [50] [51] Although frameworks vary in focus, common underlying aspects include:

- Understanding how AI operates and its strengths and limitations
- Understanding various use cases of AI

[42] Batt AM, Tavares W, Williams B. The development of competency frameworks in healthcare professions: a scoping review. Advances in Health Sciences Education. 2020 Oct;25:913-87.

[43] McGaghie WC, Sajid AW, Miller GE, Telder TV, Lipson L, World Health Organization. Competency-based curriculum development in medical education: an introduction. World Health Organization; 1978.

[44] Development of a digital competency framework for UK Allied Health Professionals. Available from: http://allcatsrgrey.org.uk/wp/download/informatics/Development-of-a-digital-competency-framework-for-UK-AHPs.pdf

[45] Australian Health Informatics Competency Framework For Health Informaticians. Available from: https://digitalhealth.org.au/wp-content/uploads/2022/06/AHICFCompetencyFramework.pdf

[46] Nazeha N, Pavagadhi D, Kyaw BM, Car J, Jimenez G, Tudor Car L. A digitally competent health workforce: scoping review of educational frameworks. Journal of medical Internet research. 2020 Nov 5;22(11):e22706.

[47] Capacity Development Network (CDN). Available from: https://www.i-dair.org/capacity-development-network

[48] Blueprint alliance for a future health workforce strategy on digital and green skills. Available from: https://bewell-project.eu

[49] Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: An e-Delphi study for competencies. Plos one. 2022 Jul 21;17(7):e0271872.

[50] Charow R, Jeyakumar T, Younus S, Dolatabadi E, Salhia M, Al-Mouaswas D, Anderson M, Balakumar S, Clare M, Dhalla A, Gillan C. Artificial intelligence education programs for health care professionals: Scoping review. JMIR Medical Education. 2021 Dec 13;7(4):e31043.

[51] Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Medical Education. 2020 Jun 30;6(1):e19285.

- Ethical and legal considerations surrounding the use of AI in healthcare
- Critical appraisal and evaluation of AI in healthcare

Less commonly mentioned are issues around how AI can be used as a decision tool in combination with professional knowledge, implementation considerations, and communication with patients and AI experts.[52] [53]

There are also some capability frameworks for implementers and practicing healthcare professionals, but these are often not tailored to specific roles.[54] [55] [56] Such frameworks tend to have a similar focus on understanding basic concepts, strengths and limitations, ethical implications, and evidence-based evaluation. They also include practical dimensions such as integration with workflows, ways to mitigate emerging risks and biases, secondary uses of data, and considerations surrounding post-market surveillance.

We only found one competency framework aimed at policy makers, the Artificial Intelligence and Digital Transformation Competencies for Civil Servants by the Unesco.[57] This includes understanding AI and current and potential future developments, identify and specify problems/use cases for AI, ways of addressing privacy and security issues.

4. The need for wider stakeholder representation in the development of AI competency frameworks

Although the development of AI-based competency frameworks for healthcare professionals is clearly important, there are many other stakeholders that need to develop AI competencies.[58] These include:

- Healthcare organisations procuring AI systems (e.g. for understanding what systems may or may not fit into existing business models and organisational practices)
- AI system developers and suppliers (e.g. to understand how systems may be tailored to contexts of use and scale across contexts)

---

[52] Çalışkan

[53] Sapci

[54] Russell RG, Novak LL, Patel M, Garvey KV, Craig KJ, Jackson GP, Moore D, Miller BM. Competencies for the Use of Artificial Intelligence–Based Tools by Health care Professionals. Academic Medicine. 2022:10-97.

[55] Artificial Intelligence (AI) and Digital Healthcare Technologies Capability framework. Available from: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/ai-and-digital-healthcare-technologies

[56] Competencies of Health Workforce in the age of Artificial Intelligence: A Conceptual Framework. Available from: https://agrh2021.sciencesconf.org/data/pages/Communication_AGRH_2021_Zaher_Vinot_1.pdf

[57] Artificial intelligence and digital transformation: competencies for civil servants. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000383325

[58] Matheny, M., S. Thadaney Israni, M. Ahmed, and D. Whicher, Editors. 2019. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. NAM Special Publication. Washington, DC: National Academy of Medicine.

- Patients and citizens (e.g. to understand how systems may impact on safety and quality of their care and to build trust, the current lack of AI literacy in the community is risking exacerbating of health inequalities)
- Policy makers (e.g. to ensure that AI strategies are evidence-based, responsible and cognisant of potential risks)

Understanding of how AI operates, and its limitations is likely to be key at all levels, but these stakeholder groups have to date been somewhat neglected.

We below introduce two case studies that illustrate potential ways to help develop AI-based competencies in healthcare for previously neglected stakeholder groups: 1) Elements of AI, and introductory AI course for beginners to help educate the general public; and 2) the recently developed National Health Service England Long Term Workforce Plan.

Use Case 1 Elements of AI[59] – an introductory course for beginners in Finland

Elements of AI is a creative commons licenced free 6-week course teaching basic concepts of AI, developed by the University of Helsinki and funded by the Finnish Ministry for Economic Affairs and Employment.[60] [61] It builds on Finland's national AI strategy that recognizes that education is crucial for adoption of AI for societal benefit and that lay people need to be engaged in discussions and in solving existing tensions. The course is intended for all audiences independent of their education, demographics, and background. It does not include complex maths and focuses on applications, teaching AI literacy and a critical informed perspective.[62]

Five years after launch, there are now close to one million users consisting of diverse learners (40% of users are female, 25% are over 45).[63] Future plans include extending the partner network outside Europe and developing an AI basics course for schools.[64] [65]

---

[59] Elements of AI free online course. Available from: https://www.elementsofai.com/

[60] Finland Releases Crash Course in Artificial Intelligence. Available from: https://www.unite.ai/finland-releases-crash-course-in-artificial-intelligence/

[61] Finland Releases Crash Course in Artificial Intelligence. Available from: https://sciencemediahub.eu/2019/06/26/a-scientists-opinion-interview-with-teemu-roos-about-ai-education/amp/

[62] AI Basics for Schools. Available from: https://ease-educators.com/ai-basics-for-schools/

[63] World's #1 AI MOOC w/ creator Prof. Teemu Roos: AI Course, AI Perception, Finland, Education & Learning, E24. Available from:https://appliedaipod.simplecast.com/episodes/worlds-best-ai-mooc-profteemuroos-course-perception-education-learning-fWaZ2PHd

[64] University Of Helsinki: Children In Preschool And Primary School To Explore Artificial Intelligence. Available from: https://indiaeducationdiary.in/university-of-helsinki-children-in-preschool-and-primary-school-to-explore-artificial-intelligence/

[65] Personal communication with Teemu Roos (developed the course)

Use Case 2 National Health Service England Long Term Workforce Plan[66]

The National Health Service England Long Term Workforce Plan was developed in June 2023. It highlights the urgent need to train healthcare staff to address current staffing shortages and highlights the important role of technological innovation in general and AI in particular. Investment in this area skills and education is planned to continue, with a focus on up-skilling and reskilling the workforce. A taskforce, building on the Topol review, will be established by NHS England to explore the best applications of AI and outline the necessary measures to ensure its effective support for NHS staff in the upcoming years.[67] The expectation is that AI will increase productivity of the workforce and improve service efficiency. In order to achieve this, investments will be made into training the NHS workforce in AI. This will draw on the Digital Healthcare Technologies Capability framework.[68] To what extend these plans will be translated into practice, remains to be seen.

6. Conclusions

Many stakeholder groups need to develop competencies to develop, sell, procure, use, and plan for AI-based applications in healthcare. Most efforts to date have focused on building on existing health information technology competency frameworks to develop competencies for medical education. There are also some frameworks focusing on implementers and practising healthcare professionals. These commonly include a focus on developing a critical empirically based understanding existing applications, their strengths and weaknesses, as well as ethical and legal frameworks.

Other stakeholders, including policy makers, system developers, citizens and patients have to date been somewhat neglected. However, these are likely to become increasingly important in developing AI-based competencies as existing applications need to effectively embed and scale. This is particularly important as many benefits of AI likely to occur in population health applications. There is therefore now a need to focus on re-skilling (acquiring new skills or knowledge in order to perform a different job or transition to a different career field) and upskilling (enhancing and expanding an individual's existing skills to keep up with evolving requirements) opportunities.

A key consideration in relation to competencies are temporal dimensions. The proposed EU AI Act promotes AI and digital literacy, but the implications of this are as yet unclear.[69] AI

[66] NHS Long Term Workforce Plan. Available from: https://www.england.nhs.uk/wp-content/uploads/2023/06/nhs-long-term-workforce-plan-v1.1.pdf

[67] The Topol Review. Available from: https://topol.hee.nhs.uk/

[68] Health Education England. Artificial Intelligence (AI) and Digital Healthcare Technologies Capability framework. Available from: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/ai-and-digitalhealthcare-technologies

[69] The Artificial Intelligence Act. Available from: https://artificialintelligenceact.eu/

applications are evolving fast, and competencies are likely to keep developing over time until there is a degree of stabilisation (i.e. when AI is becoming part of routine healthcare provision and population health). Whilst core principles should be able to stand the test of new developments, other competencies will likely be more temporary. For example, knowing how an application works may be important during the early stages of adoption when users are still establishing trust, but this may become less important as the application becomes widely accepted. Any strategic competency development therefore needs to take into account temporal dimensions as well as wide stakeholder engagement in order to ensure that AI will realise its potential in improving healthcare provision.

Investing in educational systems is crucial to ensure the integration of data science into future skill development and career paths across various expertise levels. This necessitates collaborative efforts between health and education ministries, fostering cross-ministerial collaboration. While it involves investing in digital, health, and AI literacy, there is a compelling need to invest in the upcoming generation.

Theme 4 - **Safety and Bias in Healthcare AI**

AI Club | Sub-group: Sara Boltman, Dr. Nathan Lea, Gordon Johnston, Dr Isobel Taylor

Oh, world of wonders, vast and wide, Where AI may soon take flight.
May we walk on with open eyes, and always seek the light.

[ChatGPT4, 2023]

1. Safety and Bias Summary

In this paper we consider the implications of the use of AI in healthcare with respect to Safety and Bias. We maintained there will always be human oversight with AI providing decision support depending on the context. As AI continues to advance, it is becoming increasingly ubiquitous in our lives, and different industries are finding new and innovative ways to use it. However, as we explore these new frontiers, it is essential that we consider the potential risks and dangers associated with these technologies and the difficulty of ensuring alignment[70,71]. In the field of European Health Data, from the perspective of the patient, the most relevant regulation to date is the EU AI Act[72].

It makes sense to begin with a definition of terms, as AI can be an extremely broad subject. While Artificial General Intelligence is still unachieved there have been many 'narrow' AI approaches where for a specific use case AI has been successfully adopted. For our shared understanding within this report, an AI system would be defined as a *"technological system that, autonomously or partly autonomously, processes data related human activities through the use of a genetic algorithm, a neural network, machine learning or another technique in order to generate content or make decisions, recommendations or predictions."*[73]

[70] https://aisafetyfundamentals.com/governance-blog/ai-alignment-approaches

[71] https://www.lesswrong.com/posts/2eaLH7zp6pxdQwYSH/a-brief-overview-of-ai-safety-alignment-orgs-fields

[72] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[73] Algorithmic Impact assessment and Responsible Use of AI
https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html

2. Safety and Bias Main body

Much of AI safety and ethics philosophy was proposed under the assumption that it would be developed in an 'air-gapped' or 'sandbox' environment and only when it had been comprehensively tested and proven to be 'safe' would it be released into the world in some controlled way, such that it could be disconnected if necessary. Events did not unfold that way, however, and now we see development and deployment at pace with examples such Generative AI models appearing seamlessly embedded into internet search engines and acting as a 'co-pilot' in our everyday office tools. Some companies have banned[74] the use of AI, others are taking a more pragmatic approach and include policy training for all their staff on responsible use of Generative AI.

While understanding of the underlying technology is not widespread, patients familiar with social media will have seen many examples of images and video created by generative AI, poetry and code written in response to prompts and may come to expect an encyclopaedic level of knowledge from every professional they interact with – if their optician uses augmented reality to show them what their new glasses will look like, why wouldn't their doctor know about the latest research on HRT[75] or nutrition[76]?

Some narrow applications of AI have been extensively trialled[77] and compared to human performance for tasks such as grading retina scans for Diabetic Retinopathy (DR) – two algorithms have US FDA approval. Deep learning has been used on breast cancer[78] cases to analyze genetic sequencing and histopathological imaging to help with diagnosis and treatment. But these are the ultimate applications of AI in healthcare – there are many less controversial areas the technology could be deployed.

The domains it could be applied to can be ranked in order of risk (from low to high and with some examples):

1) Automation of administrative tasks: letter and email templates, automatic adjustment of rotas when 'out of office', automated expense claim systems or reordering of out-of-stock items. The type of business application not specific to healthcare and in widespread usage in business.

2) Inform clinical management: optimising waiting lists, Operating Theatre scheduling and hospital flow logistics. Decision support for triage of non-critical patients.

[74] https://www.businessinsider.com/chatgpt-companies-issued-bans-restrictions-openai-ai-amazon-apple-2023-7?r=US&IR=T

[75] https://bjgp.org/content/70/700/e772.short

[76] https://www.sciencedirect.com/science/article/abs/pii/S0002934317312299

[77] https://diabetesjournals.org/care/article/44/5/1168/138752/Multicenter-Head-to-Head-Real-World-Validation

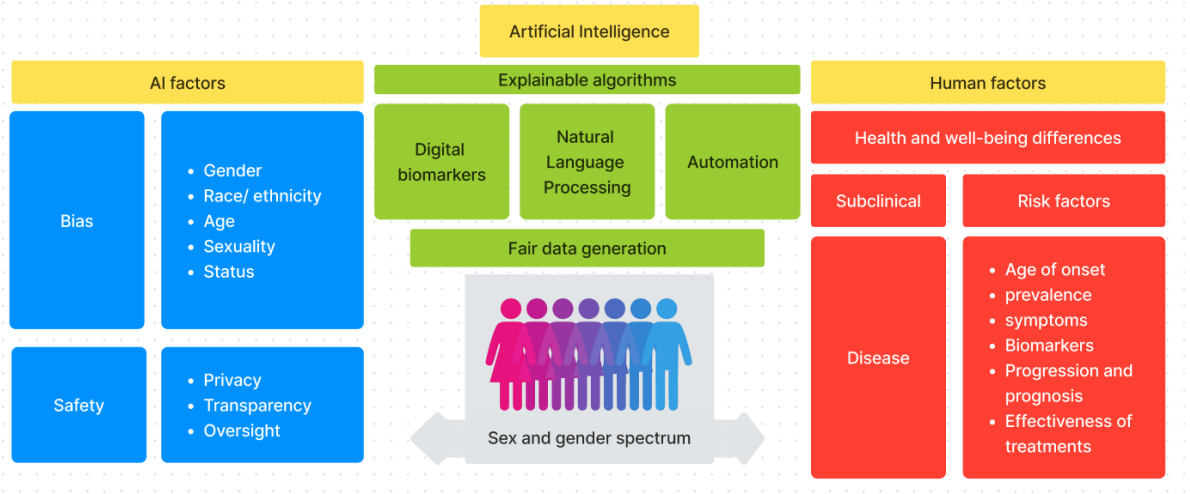
[78] https://www.sciencedirect.com/science/article/abs/pii/S0933365722000410

3) Drive clinical management: to match donors with patients on a transplant list, using an explainable decision tree or flow diagram.

4) Diagnose or treat: to determine whether an image of a tumour should be classified as malignant or benign, to measure the size of a foetus in utero,

3. Understanding Bias

Algorithms used for AI may not be inherently biased, but the outcome still could be. Machine learning models are trained on vast amounts of data, much of which has been collected from western white males. For example, the Large Language Models trained by US companies have English as their default language and the content of the internet to learn from, leading to a somewhat skewed view of the cultural record of the rest of the world.



There are many types of bias[79] and the diagram above shows only a few examples for illustration, although we mention 2 cases in more detail below there are many ways in which the data collected may become systematically biased – for example people who are physically fit may be more likely to participate in activities which collect data, likewise people who are well educated and technology literate.

Gender bias: many medical trials[80] are carried out on male participants, even when the subjects are mice rather than humans, to avoid interactions with different levels of female hormones during the menstrual cycle. This does mean certain pharmaceuticals may not be as effective in women as men or may not work at certain times of the cycle. Data collected from these trials and used to simulate synthetic data will perpetuate this blind spot.

[79] https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence

[80] Women's health and clinical trials, Londa Schiebinger https://www.ncbi.nlm.nih.gov/pmc/articles/PMC198535/

Racial bias: white patients are over-represented in most datasets[81]. For image recognition the Labelled Faces in the Wild (LFW) dataset assembled in 2007 by a team from UMass Amherst is 77% male and 83% white and has been used to build face-recognition apps.

Selection bias: This occurs when the dataset used to train an AI model does not accurately represent the population it is intended to serve. For example, if an AI system used to determine creditworthiness is trained only on data from one demographic group, it may make biased decisions against individuals from other groups.

4. Discussion and case studies

There have been many high-profile case studies cited where gender bias has been found in recruitment algorithms or racial bias resulting from the use of facial recognition software by law enforcement organisations. Our aim in this paper is to focus on the current 'state of the art' of AI in healthcare applications, from the perspective of the patient.

Exclusion of under-represented groups from medical trials is sometimes deliberate – for example following the Thalidomide scandal[82] the US FDA issued guidelines excluding the participation of women of childbearing age in drug trials. The infamous Tuskegee Study[83] has left a legacy of healthcare mistrust amongst people of colour and other groups who experience economic vulnerability or social deprivation.

The diagnosis, treatment and prevention of serious chronic diseases such as cardiovascular disease[84] continue to be based primarily on findings in men and sex-specific clinical guidelines are lacking. Spontaneous coronary artery dissection (SCAD), a potentially fatal condition that can cause a tear in blood vessels in the heart, predominantly affects women. Yet women are underrepresented in clinical trials[85] investigating SCAD. When developing algorithms, some biases relating to biomarkers[86] can be clearly identified within the data, prompting calls for a 'rebalancing' of the sample.

[81] Buolamwini, Joy. 2016. "How I'm Fighting Bias in Algorithms." TED Talk. Accessed June 11, 2021. https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms?language=en.

[82] https://www.sciencemuseum.org.uk/objects-and-stories/medicine/thalidomide#the-thalidomide-scandal

[83] https://www.tandfonline.com/doi/abs/10.1080/08964289.2019.1619511

[84] https://www.jacc.org/doi/abs/10.1016/j.jacc.2022.02.010

[85] Perdoncin E, Duvernoy C. Treatment of Coronary Artery Disease in Women. Cardiovasc J. 2 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5935279/

[86] https://www.nature.com/articles/s41746-020-0288-5

Use Case - Skin Cancer Diagnosis

For people of colour, introducing AI into the diagnosis of skin cancer could make an already inequitable[87] situation much worse. The standard image set[88] used for training of AI skin cancer diagnosis models has over-representation of Caucasian skin, leading to Racial bias[89]. Due to problems with edge detection and contrast, image enhancement techniques have not proven effective in correcting for this bias. Work on synthetic patient data[90] that captures the spirit of outliers and fills in gaps in particular profiles such as age and gender has produced statistically valid aggregated data preserving the relationships and nuances contained in the original dataset. In separate academic research, Generative Adversarial Networks have been used to expand the training set of images[91]. Both approaches show some promise but further trials will be needed.

5. Safety

Topics such as Privacy, Transparency and Training are covered in other chapters so our focus here will be on the prevention of actual harms to patients being caused by AI, either because the patient is disadvantaged by the automation processes put in place to smooth the administrative running of healthcare systems, or at the actual point of care where a mis-diagnosis could theoretically occur. Decision support algorithms are classified as 'devices' and as such can be CE marked in the UK by going through the appropriate process. In Europe the detailed guidance is still under discussion at the time of writing but US based tech companies are watching with interest as this legislation may affect their ability to market to European customers. US approval for Safety in AI healthcare comes through FDA, as with any technology in healthcare will need to meet medical device classification I to III, where I is the lowest risk and III is considered high risk. Much focus has been placed on ensuring technology companies only collect and retain data related to the primary aim of their product (the goods or services the customer receives) rather than building up consumer interaction data to train future machine learning models for new products. These so called 'dark patterns' can start with the best of intentions – for example a mental health app may make use of other data collected on the user's phone, such as how often they text or call people, so next time they feel low the app can prompt them to

---

[87] The Bias of Physicians and Lack of Education in Patients of Color With Melanoma as Causes of Increased Mortality https://assets.cureus.com/uploads/review_article/pdf/123676/20221219-28425-9ho777.pdf

[88] Characteristics of publicly available skin cancer image datasets: a systematic review https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00252-1/fulltext?ref=megabytesandme.com

[89] https://www.science.org/doi/10.1126/science.aax2342

[90] Synthetic data https://diveplane.com/geminai/

[91] Data as a tool to combat racial bias https://www.researchgate.net/profile/Daniel-Kvak/publication/365635174_Synthetic_Data_as_a_Tool_to_Combat_Racial_Bias_in_Medical_AI_Utilizing_Generative_Models_for_Optimizing_Early_Detection_of_Melanoma_in_Fitzpatrick_Skin_Types_IV-VI/links/63d15b42d9fb5967c204c470/Synthetic-Data-as-a-Tool-to-Combat-Racial-Bias-in-Medical-AI-Utilizing-Generative-Models-for-Optimizing-Early-Detection-of-Melanoma-in-Fitzpatrick-Skin-Types-IV-VI.pdf

reach out. Or for dementia patients, speed of typing whilst texting could indicate decline. For every interaction there is some form of justification for why the data ‘might be useful one day’ which leads to over-collection and tech-surveillance of users of most apps, with data breaches of behavioural data becoming potentially more valuable as the volume and sophistication of the data grows.

6. Conclusion

AI holds great potential for improving healthcare, but it must be governed to ensure fairness, privacy, transparency, responsibility and patient safety for it to be trusted by all patients regardless of gender, race or demography. Removing indicators from training data is ineffective as there are many proxies for protected categories, neither can auditing and reporting upon the distribution of the training data provide a panacea (though that is a necessary first step). Allowing the peer review and open critique of proprietary algorithms by researchers reporting to a central governing body would inform future developers of AI to be aware of the pitfalls of existing biased training datasets. Proactively encouraging under-represented groups to take their place at the table will be vital in securing equity of care in a data driven health landscape.

References and Abbreviations

EU AI Act - The AI Act is the European Union's draft legislation on the regulation of AI. Its classification system determines the level of risk an AI technology could pose to the health and safety or fundamental rights of a person. The framework includes four risk tiers: unacceptable, high, limited and minimal.

CE Marking - The EU and UKs current standard for medical device, AI and software as a medical device certification.
https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices
/https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk#overview

FDA Approval - US Food and Drug Administration who certify AI and software as a medical device.
https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd

IEEE - The IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems: This initiative, launched by the Institute of Electrical and Electronics Engineers (IEEE), brings together experts from a variety of fields to develop standards and guidelines for the development of ethical AI systems.
https://standards.ieee.org/industry-connections/ec/autonomous-systems/

OECD - The Organisation for Economic Co-operation and Development an intergovernmental organisation with 38 Member countries, founded in 1961 to stimulate economic progress and world trade.
https://www.oecd.org/science/laying-the-foundations-for-artificial-intelligence-in-health-3f62817d-en.htm

the digital health society

The data arm of ECHAlliance Group