Domesticating AI in medical diagnosis

Technology in Society 76 (2024) 102469

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Domesticating AI in medical diagnosis

Robin Williams [a,*], Stuart Anderson [b], Kathrin Cresswell [c], Mari Serine Kannelønning [d], Hajar Mozaffar [e], Xiao Yang [a,b]

[a] Institute for the Study of Science, Technology and Innovation, The University of Edinburgh, United Kingdom
[b] School of Informatics, The University of Edinburgh, United Kingdom
[c] Usher Institute, The University of Edinburgh, United Kingdom
[d] Department of Archivistics, Library and Information Science, Oslo Metropolitan University, Norway
[e] The University of Edinburgh Business School, The University of Edinburgh, United Kingdom

ARTICLE INFO

Keywords:
Domestication
Social learning
Artificial intelligence
Machine learning
Medicine
Health
Diagnosis

ABSTRACT

We consider the anticipated adoption of Artificial Intelligence (AI) in medical diagnosis. We examine how seemingly compelling claims are tested as AI tools move into real-world settings and discuss how analysts can develop effective understandings in novel and rapidly changing settings.

Four case studies highlight the challenges of utilising diagnostic AI tools at differing stages in their innovation journey. Two 'upstream' cases seeking to demonstrate the practical applicability of AI and two 'downstream' cases focusing on the roll out and scaling of more established applications.

We observed an unfolding uncoordinated process of social learning capturing two key moments: i) experiments to create and establish the clinical potential of AI tools; and, ii) attempts to verify their dependability in clinical settings while extending their scale and scope. Health professionals critically appraise tool performance, relying on them selectively where their results can be demonstrably trusted, in a de facto model of responsible use. We note a shift from procuring stand-alone solutions to deploying suites of AI tools through platforms to facilitate adoption and reduce the costs of procurement, implementation and evaluation which impede the viability of stand-alone solutions.

New conceptual frameworks and methodological strategies are needed to address the rapid evolution of AI tools as they move from research settings and are deployed in real-world care across multiple settings. We observe how, in this process of deployment, AI tools become 'domesticated'. We propose longitudinal and multisite 'biographical' investigations of medical AI rather than snapshot studies of emerging technologies that fail to capture change and variation in performance across contexts.

1. Introduction

The recent striking achievements of Machine Learning have reinflated expectations and driven substantial public and private investments in AI for healthcare and especially diagnosis [1–3]. A cornucopia of promising applications has arisen, often emerging from enthusiastic local collaborations between clinicians and academic AI specialists. This has resulted in the formation of very large numbers of ambitious start-ups - high-risk ventures with initially limited resources - alongside the strategic investments being made by big IT and health vendors moving into the area [3].[1] Expectations about the transformational potential of AI are, however, being called into question by the brittle and variable performance of early clinical AI applications [4,5] In particular, the performance of AI tools in the field has not matched laboratory demonstrations on curated data sets (Idem. 2020 [6,7]; and fluctuates as tools are applied across different settings [8]. As a result, developers and early adopters find themselves engaged in a struggle not only to develop AI-based tools but also, crucially, to demonstrate their

* Corresponding author. Institute for the Study of Science, Technology and Innovation, The University of Edinburgh, Edinburgh, EH1 1LZ, United Kingdom.
E-mail address: r.williams@ed.ac.uk (R. Williams).

[1] In the current period major investments are being made by the big health technology companies. https://www.philips.no/healthcare/resources/landing/ai-manager. https://www.gehealthcare.com/products/edison. https://www.siemens-healthineers.com/digital-health-solutions/artificial-intelligence-in-healthcare.

https://doi.org/10.1016/j.techsoc.2024.102469
Received 14 August 2023; Received in revised form 20 January 2024; Accepted 20 January 2024
Available online 26 January 2024

clinical effectiveness and sustain them at scale across different technological and social contexts.[2] These efforts constitute a real-world social experiment [9]. However, notwithstanding commercial and political push towards accelerated implementation timelines, little attention has been paid to conducting high-quality evaluations of AI system performance [10–12].[3] In many of these real-world experiments, potentially valuable learnings are not being systematically collected, let alone applied. Governments worldwide have accordingly begun to extend the remit of their policies and strategic interventions on AI in healthcare, beyond promoting new tool development towards mobilising and guiding various actors to contribute to enabling and accelerating the implementation processes on a broader scale (see, for example, Danish Ministry of Finance & [13–15].).

The research team have been exploring empirically these issues arising in early, highly-experimental stages in the development and adoption of diagnostic AI tools in different settings. By comparative analyses of four case studies we seek to develop an effective evolutionary understanding, drawing on concepts from Science and Technology Studies. Our analysis (and this paper) focuses on the interactive processes of sensemaking and negotiation involved, characterised by Ref. [16] as social learning. These discovery processes are particularly intense in the early stages of establishing a promising emerging technology like medical AI as those involved struggle to get artefacts to work and to be useful in contexts of early adoption. In this paper we distinguished two key moments in this process with differing features: i) initial experimentation, applying machine learning techniques to data sets to create effective models and demonstrate their potential value; and ii) subsequent implementation and use of AI tools within and across real-world settings, scaling-up from initial contexts to other sites, and collecting the evidence needed to demonstrate their clinical robustness and commercial viability. In this research we also observed unfolding initiatives to extend the scope of AI applications – moving from single point applications to deploying suites of AI tools through platforms. Our analysis highlights how, in this process, technologies and expectations arising "in the wild" of research settings are tamed and technologies and their uses domesticated [17,18] to meet the exigencies of use by professional specialists and other stakeholders in health service settings.

We reflect upon the challenges of studying a rapidly changing emerging technology like diagnostic AI. Applying methodological insights from the Biography of Artefacts and Practices perspective [19] we highlight the need for longitudinal and multi-site studies of adoption and widespread use.

1.1. Social learning and the domestication of technology

[16] drew attention to the collective processes of experimentation, tinkering and sensemaking involved in developing and applying novel technologies in contexts of work and everyday life, captured by the term social learning [16]. For Sørensen these are processes of negotiation and conflict between the diverse actors involved and implicated as well as narrow cognitive processes of knowledge exchange. Sørensen highlighted the practical trial-and-error activity involved in the struggle to implement an artefact and make it useful in contexts of use (*learning by doing);* how local knowledge and experience may be transferred elsewhere *(learning by interacting)* and how players seek to order other players and establish the "rules of the game" for the appropriate use of the new technology *(learning by regulating)*. In this process they may *domesticate* - establish norms and practices for appropriate use of - technologies coming in from *'the wild'* of R&D [18].

We apply this conceptual framework to explicate the social learning processes involved in applying diagnostic AI. We explore how they vary across two key moments and settings, contrasting: i) the 'upstream' stages involving experimental innovation of hopeful applications through research and development of AI tools and coupling them with clinical settings, and, ii) the 'downstream' stages of embedding, scaling, and extending the scope of established AI tools. We argue that tool development and implementation involves ongoing experimentation [20]. Innovation continues after an artefact leaves the settings of research and development as actors involved in the down-stream roll-out stages struggle to deploy it in contexts of implementation and use [21]. However we note that the focus of social learning is changing across these evolving settings. Local learning-by-doing by clinical and computational researchers developing AI tools in the upstream cases (1 and 2) gives way to more distributed learning-by-interacting and learning-by-regulating as players collaborate to establish institutional mechanisms and infrastructures to support the reliable operation at scale and sustainable procurement of a range of tools. The latter endeavour brings in a wider range of players involved in governance and procurement (especially case 4).

2. Research design and methodology

2.1. Our sample of four cases

We explore the processes of social learning empirically through four independently initiated studies involving various combinations of the research team (and other collaborators, see acknowledgements), united by our shared conceptual framework. Two cases - a research project developing AI for recognising brain lesions and a project on embedding AI for Atrial Fibrillation (AF) prediction in clinical practices - explore upstream stages in the emergence of medical diagnostic AI. The other cases- a project rolling out an AI tool for lung nodules and a procurement project for AI image diagnostics - focused 'downstream' on embedding and expanding the scale and scope of adoption of established tools. Table 1 summarises this.

The cases are presented in detail in the result section as parts of vignettes summarising emerging findings from each study.

2.2. Data collection

The case studies, conducted by different researchers (respectively: XY, HM and KC, KC and RW, and MK), primarily involved qualitative interviews with a selection of key interviewees directly involved in the AI implementation from adopter organisations, vendors and broader stakeholders (except for case 1, which was a participant observation study). Data collection methods for each case are summarised in Table 2. For all cases, written informed consent was obtained from the interviewees, and all data were anonymised to protect participant confidentiality. Institutional approval for each study was secured using the investigators' University Research Ethics processes.

Case 1. Automatic Recognition of Brain Lesions - research project

The author, XY, conducted an ethnographic study of an academic

Table 1
The focus of the studies.

Upstream: creating and coupling tools with clinical settings	Downstream: embedding established tools
Case 1: Automatic Recognition of Brain Lesions – research project	Case 3: Rolling out Computed Tomography scanning tool for Lung Nodules
Case 2: Digitising ECGs and AI-enabled Atrial Fibrillation (AF) Prediction	Case 4: Hospital AI procurement strategies: From AI apps to platform

[2] AI tools operate through complex assemblages of algorithms, training data, interconnected devices/infrastructures eg scanners, divisions of labour amongst professional users which vary in ways that influence how the tool performs.

[3] Even in the most well-established area of computer-aided image analysis, in mammography, outcome and performance evidence is incomplete. See Ref. [35].

Table 2
Overview of research methods.

	Interviews	Meetings observed	Participant observations
Case 1	–	–	Daily for 5 months
Case 2	15	–	–
Case 3	39	–	–
Case 4	6	14, approx. 60 h altogether	–

research group working on the automatic segmentation of WMH (White Matter Hyperintensities) brain lesions. The interdisciplinary group comprised three Biomedical AI doctoral students, one supervisor in machine learning (ML), and one supervisor in medicine. XY conducted participant observation in the project for five months in the role of one of the ML researchers.

Case 2. Digitising ECGs and AI-enabled Atrial Fibrillation (AF) Prediction

The authors, HM and KC, conducted 15 in-depth semi-structured qualitative interviews with healthcare professionals currently using ECGs from six health provider settings across UK. Using a purposeful snowball sampling approach, participants were recruited with the following criteria: use of ECG in daily practice, maximum variation in relation to use cases, geography, seniority and demographics of health provider settings. Interviewees included medical consultants, cardiology consultants, emergency department consultants, cardiac physiologists, junior doctors, and nurses.

Case 3. Rolling out Computed Tomography scanning tool for Lung Nodules

The authors, KC and RW, were funded to conduct a qualitative formative evaluation of the roll-out of the AI tool for pre-cancerous lung nodules across various UK hospitals – alongside a quantitative study of the effectiveness and cost-effectiveness of the tool. The authors and two research assistants (listed in acknowledgements) conducted 39 interviews, primarily with clinicians involved, but also other hospital stakeholders, members of the vendor organisation, and technical and medical experts in the field.

Case 4. Hospital AI procurement strategies: From AI apps to platform

The author, MK, conducted non-participant observations of 14 digital meetings for approximately 60 h altogether and six semi-structured interviews during the procurement process. The digital meetings observed were meetings between the project team of the AI procuring health service provider and five vendors invited to participate. The focus of the observations was the verbal communication unfolding during the meetings. After the meetings were completed, semi-structured interviews were carried out with the four vendors qualifying for making a final offer and with the procurement team's project manager.

2.3. Data analysis

Interview transcripts from cases 2, 3 and 4 and field notes from cases 1 and 4 were coded and analysed individually by the respective researchers. For case 3, results were coded using ethnographic software. Periodic discussion between the research teams, some with overlapping membership (cases 1, 2 and 3) harmonised coding and interpretation between cases.

The analysis for each case focused on examining the social learning process and extracting the primary insights gained in each instance. This is presented is the results section. We then used our biographical lens to compare the findings across the four cases and map different parts of AI innovation journey identifying three key themes which are presented in the analysis section.

3. Results: four vignettes

3.1. Case 1 Automatic Recognition of Brain Lesions research project

3.1.1. Introduction

This case study examines an academic research project on the automatic segmentation of White Matter Hyperintensities (WMH) – a kind of brain lesion that can be detected by MRI. Identifying and quantifying WMH is crucial for clinical treatment and brain disease research. Segmentation of MRI is considered time-consuming and labour-intensive, and thus the introduction of automated mechanisms is imperative. Machine learning models for WMH segmentation have become an important subfield of diagnostic AI.

3.1.2. The case

WMH segmentation is a pattern recognition problem. However, the team still needs researchers with medical backgrounds to explain the data (annotating WMH images) and the medical issues involved.

The project used a public dataset and also a local dataset. The public dataset, comprising MRI images of 60 patients from five scanners of three institutions, was provided by a competition (WMH Challenge (https://wmh.isi.uu.nl/), organised by the MICCAI (Medical Image Computing and Computer Assisted Interventions) which in 2017, targeted "the segmentation of white matter hyperintensities of presumed vascular origin on brain MR images" [22]. Participants were able to download the training data by registering for the conference. Models with good performance were accepted for the conference. After the conference, these models could be further refined for publication in journals. The Challenge dataset is still available for future research. There are many similar medical image "Grand Challenge" competitions (see https://grand-challenge.org/challenges/) which have helped drive interest in applying AI in medical images, especially in radiology [23]. As well as the public datasets, studies require further data to train models. An additional local dataset was provided by the medical researcher [24].

The data provided by the Challenge were already pre-processed. However, the images usually need to undergo a series of pre-processing operations before they can be fed into ML models, such as registration (unifying the images under different scanners to the same coordinates), brain extraction (removal of the shadows of brain bones from the images), and sometimes resampling normalisation (reducing non-significant variations in images – in this case, normalising the distribution of the grey values or intensities of the background). When pre-processing these files, integrated-designed toolkits are often more efficient than code written directly by researchers. These toolkits[4] are developed by professional institutions and are free to use [25]. The Challenge data has also been pre-processed with these toolkits.

3.1.3. Trial-and-error – cycles of optimisation and evaluation

For WMH segmentation, statistical metrics are used for presenting the performance of the models. The WMH Challenge had five metrics to automatically evaluate the models, which are aimed at scoring the similarity between the predicted and annotated segmentation [26]. For example, the most widely used metric Sørensen-Dice Similarity Coefficient (DSC) is defined as the harmonic mean between sensitivity and precision. Different evaluation methods can be used to evaluate different aspects of WMH segmentation from the perspective of accuracy, robustness, consistency, and reliability. Therefore, it is advisable to use multiple evaluation methods to obtain a comprehensive and fair assessment of WMH segmentation algorithms.

During the experiments, images are subjected to evaluation methods which generate numerical values. ML scientists are committed to

[4] Such as the FMRIB Software Library (FSL) and Statistical parametric mapping (SPM).

improving the performance of their models as measured through these evaluation metrics. Researchers tend to start with well-known ML models. For example, in recent years, most research groups used U-Net. It was introduced in 2015 [27] and is widely used for medical image recognition because of its mathematical attributes, suitable for capturing critical information in a relatively small amount of image data. U-Net is often used as a baseline first, followed by some adjustments, such as the number of convolution layers and the value of the hyper-parameters.

The models used in ML required constant tuning of the parameters to obtain an optimal solution. They read the data into the model, output the results, and evaluated the model with the statistic, a numerical value. Next, the scientists changed the model's parameters and trained and tested the model again. The process was lengthy and sometimes required serendipity.

Incremental improvements to the model increase the model quality, reflected in numerical evaluation values. The results trained by ML models were compared with these standard results. WMH Challenge solutions are publicly available, and many ML researchers upload their code to open-source platforms, such as Github. Every improvement leads to a publication, which has led to a surge in the number of papers and citations in this field. In contrast to the confidentiality surrounding commercial developments, research in academia is presented in a more open and transparent approach to improving models and publications. The medical imaging challenge has also led to the creation of a decentralised, shared medical AI infrastructure. Data collections and ML models built by researchers are shared via open-source platforms, which facilitate the further scaling up in a real-world model. Regarding being built on the incentives of publication, researchers in academia have a limited focus on practical applications.

3.1.4. Concluding remarks

In the automatic segmentation of WMH, images are annotated by medical experts. Medical expertise has not vanished but is becoming increasingly reliant on technology and software in complex ways [28]. In contrast to the public discourses on the power of ML and other AI tools, this study emphasises the often overlooked practical work involved in developing and improving effective AI models, starting with the establishment and cleaning of data sets for training and validation [29]. ML researchers undertook a series of trial-and-error, 'learning--by-doing' [16] activities, adjusting the parameters of the models during training to improve the statistical values used to assess the models' performance.

Researchers in academia are subject to powerful incentives for publication and have limited focus on practical applications. However, the ML models built by researchers are open sourced in open-source platforms, which facilitates further scaling up in a real-world model. The medical imaging challenge has also led to the creation of this decentralised, shared medical AI research infrastructure.

3.2. Case 2: digitising ECGs and AI-enabled Atrial Fibrillation (AF) prediction

3.2.1. Introduction

This case explores the views of health professionals about embedding an AI enabled clinical decision support system (CDSS) for electrocardiogram (ECG)-based prediction of Atrial Fibrillation (AF) in clinical settings. The tool had been developed in 2020 by a team of developers, including clinicians and informaticians working across several academic and health settings. After successful testing of the tool in laboratory settings, it was presented to its future potential users through use case scenarios to evaluate its potential use for practice.

3.2.2. The case

ECGs are simple heart activity diagnostics tests frequently conducted by hospitals and other healthcare providers to investigate cardiac conditions of patients. Paper-based ECGs continue to be in common use in many health provider organisations, with limited digitisation and limited links to digital patient records [30]. Despite increasing uses of CDSS in some health specialties, and the availability of knowledge-based CDSS in ECG machines, the decision support information produced by machines are rarely trusted and used by health professionals. Recent research, however shows that, AI has great potential in cardiac screening such as detection of left ventricular dysfunction and diagnosis of episodic AF [31].

These two elements shaped the opportunities for development of an AF-prediction tool using AI by the developer team including cardiologists and health informaticians. Prediction of AF is an under-developed field in the context of AI, but deep learning techniques are increasing the accuracy of predictions and detection. There are two potential contexts for use of AI prediction for AF: a) Consumers (wearable devices), and b) clinical health settings (patient pathways in hospitals). This particular tool was developed for the second context in the clinical health settings.

After successful experimental development of the tool in laboratory settings, the developer team decided to employ an evaluation team to present the tool and its use case to health professionals to initially determine the availability and potential production of learning data for the AI tool and, secondly, to evaluate the perception of future users on embedding the AI tool into clinical practice.

3.2.3. Health professionals' vision of use and reliability

As the AI-enabled CDSS tool for ECG was explained to the health professionals by the evaluators, we identified three areas of concern in moving from experimental settings to develop effective models for embedding predictive diagnostic AI into practice.

First of all, there was a lack of differentiation between AI and conventional knowledge-based CDSS operating mechanisms and, hence, a lack of differentiation between the resulting interpretations. While health professionals appreciated the overall potential of AI for health diagnosis, they believed that digital predictive interpretations of ECG might have limited value and, hence, might be ignored. This view was largely based upon their experience of knowledge-based CDSS in current ECG equipment. This lack of trust in data interpretation led to further scepticism as to whether predicting AF was accurate "Because I mean predicting the future seems a little bit odd". Secondly, due to the lack of digital ECGs more generally, professionals found it hard to imagine having additional information and ways of integrating this into the existing workflow: "You're talking about something that we don't have at this moment in time, and therefore we don't know what to do with it as well." This led to many health professionals stressing the need for digitalisation of ECGs rather than incorporating AI for complex prediction: "we need digital copies of ECG that we can connect to the right patient". Thirdly, there was uncertainty about changing the course of treatment, even if AI predictions were valid, as predictions were only perceived to be useful if they would result in changes to the course of treatment, which was not perceived as viable in the case of AF prediction: "… in terms of change my management of AF, it would be something about rate control and rhythm control and anticoagulation. But I'm not going to do any of those things unless they [the patient] are actually in AF.".

3.2.4. Conflicts between different groups about perceived value generated by AI

Health professionals valued AI differently depending on their seniority and background. For instance, more senior and experienced clinicians were less likely to value an additional decision support aid as they were used to interpreting ECGs themselves. There was also a difference in value of AI as perceived by different specialties. For instance, cardiologists had a preference to interpret ECGs by themselves, whereas other specialities (e.g., nurses) mentioned possible benefits from having "valid" interpretations when cardiologists were not present: "often you'll get ECGs where you're not sure whether [it is] AF or not."

Most respondents expressed difficulty in envisioning predictive AI use for decision support, or scepticism about its operation and outcomes. However, some professionals envisaged possible future uses for AI predictions. These included different ways of using AI from those anticipated by the development team, including alerts to issues that may have otherwise been missed - particularly for special/edge cases rather than for routine predictions - which could strengthen decision making. Similarly, AI could be used to help identify analyse risk factors to be presented to health professionals e.g., when deciding whether to offer procedures to particular patients, to support clinical decision making.

Another way of envisioning value from AI-enabled prediction was to change the focus of the health concern. Some health professionals believed that AF prediction is not necessarily useful in offering better treatment to patients – as AF on its own may not be necessarily harmful. Rather AI can be more effective if it could predict likelihood of certain health conditions such as stroke or heart attack: "Identifying/predicting a stroke or heart attack may be more useful than predicting AF, as this will change the course of treatment and may also be helpful for those not used to reading ECGs".

3.2.5. Perceptions of unintended consequences

Health professionals perceived different types of unintended consequences in use of AI-enabled predictions of AF. Firstly, AI prediction of AF was perceived to introduce clinical uncertainty which was perceived to potentially impact on patients (e.g., through unnecessary investigations) "they can create clinical uncertainty and patients may in consequence end up having unnecessary investigations or being kept in for longer than they would have otherwise." Others mentioned safety issues that might result from unnecessary interventions and patient treatment pathways: "I'd be cautious that … there'll be a danger that … either medication or intervention would be put in place that might actually result in an unforeseen harm." "Where the diagnosis or the proposed diagnosis is inaccurate, that can lead to inappropriate investigations being ordered, which will indirectly cause harm to patients."

Secondly, health professionals raised concerns about overreliance on technology leading to deskilling of junior health professionals "If AI takes over, you think that the future generation of doctors will not be very good at interpreting ECGs. And that's of course a concern as well. And we do have groups of doctors now who do not normally interpret ECGs". Further concerns were raised about additional workload that was required in order to generate AI friendly data for use by the algorithms: "Right now if you asked someone to scan an ECG, I think you'd get a look of horror."

3.2.6. Concluding remarks

The development of technological innovations for health diagnostics necessitates an understanding of the diverse knowledge foundations within different medical domains. Our study of AI-based CDSS in predicting AF illustrated that there were conflicting visions of applications of AI in health diagnostics, which were influenced by differing professions and backgrounds as well as past experiences of use of technology. Developing diagnostic tools can be challenging, especially when the relationship between symptoms and diagnosis or outcomes is not well-established. Without a well-established relationship between signs and diagnosis, there may be a lack of consensus among medical professionals regarding the interpretation and significance of certain signs, which may lead to variations in diagnostic criteria as well as decision-making. It is difficult to apply AI techniques in such uncertain settings, let alone to develop CDSS. Clear and practical use cases are essential to inform the design and validation of technologies and demonstrate their effectiveness in real-world scenarios. This requires a multidisciplinary approach involving clinicians, social scientists, statisticians, and data scientists. Otherwise, the adoption of such technologies and their potential benefits may be limited.

Furthermore, the use of AI in clinical settings for risk predictions raises specific issues regarding trust among clinicians. In particular, where existing digital tools are perceived to be inaccurate in specific areas of health diagnosis, clinicians' perception of risk prediction is more negatively impacted. AI algorithms often operate as black boxes, meaning that the inner workings of the algorithms and the reasoning behind their predictions are not easily understandable by humans. This lack of transparency can increase mistrust among clinicians, as they may not fully comprehend how the AI system arrives at its predictions. The opacity of AI systems can be seen as a risk factor, as it limits the ability to validate or explain the results. Whilst experts can scrutinise AI-based interpretations by comparison with their own judgements (e.g., as we see in the next case, #3, in the detection of lung nodules by radiologists), this cannot readily be applied to algorithmically-predicted long-term outcomes. Trust in predictive tools may need to be established through the conduct and dissemination of formal assessments rather than pragmatic experience. Lack of trust is a potentially serious barrier to clinical acceptance, adoption and use of diagnostic AI [32].

3.3. Case 3: Rolling out a Computed Tomography scanning tool for lung nodules

3.3.1. Introduction

This case explores the early roll-out of a commercial AI tool to support review of Computed Tomography (CT) lung scans across various UK NHS hospitals (Farič et al., 2023). The tool was originally developed in the Netherlands and was already licensed for use in the UK. The developer, Aidence, received an award/" title="https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/">AI in Health and Care Award from the National Health Service (NHS) AI Lab, a UK initiative focused on integrating AI technologies into healthcare services provided by the NHS. Aidence was thereby funded to implement and evaluate the tool for incidental scans conducted in routine clinical care (it is already used in screening at risk patients) to determine its effectiveness, and clinical and economic impact with a view to generating the evidence required for the National Institute for Health and Care Excellence (NICE) to recommend large-scale deployment.[5]

3.3.2. The case

Radiologists are grappling with increasing volumes of scans. There has been a long history of applying computerised decision support (CDS) to help improve the speed and reliability of screening starting notably in mammography [33] and extended to other major areas of cancer screening (lung, prostate cancer etc). The Aidence Veye Lung Nodule (VLN) tool is already widely used to identify and measure potentially cancerous lung nodules through Computerised Tomography (CT) for screening at risk patients (with symptoms or nodules). In this trial VLN is being considered as a means to monitor 'incidental' CT chest scans undertaken as part of routine care for other purposes. The large numbers of scans involved with lower incidence of disease augments the potential benefits of CDS. Though there have been numerous claims from technology suppliers about the performance of AI in comparison to humans, these are mainly based on test data and there is a lack of evidence about

[5] https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/ai-health-and-care-award-winners/.

performance of these tools in real-world settings [34])[6] – even in the long-established case of breast cancer screening [35].

The VLN tool is designed to integrate with existing radiology scanning workflows and infrastructure. Anonymised scans are screened in the cloud (to mitigate local Information Governance barriers). The results are delivered via the existing Picture Archiving and Communication System (PACS) as an overlay to the CT image on the reader in the radiology room.

3.3.3. Issues in implementation

Though the VLN tool was designed to work with Digital Imaging and Communications in Medicine (DICOM) standards-compliant PACS, compatibility problems were encountered. One PACS vendor was slow to resolve these, which delayed roll-out to some sites. Poor technical infrastructure (wifi/internet/server capacity) in other sites meant that scan results were only available the following day. This asynchronous processing had knock-on effects. For example, patients needed to be called back rather than being treated in a single visit, impeding attempts to create a smoother, more efficient workflows. In another site, problems with system configuration led to impaired performance (an incorrect safety/specificity setting which was picked up by radiologists!). Integrating the tool into the local infrastructure and maintaining it entailed not insignificant effort (and cost).

3.3.4. Issues in system use

Radiologists reported a positive experience of using the system and particularly appreciated the way it was designed to integrate within existing workflows and PACS systems. These highly skilled users rapidly became acquainted with the affordances of the tool. They pointed to its strengths in reliably identifying large numbers of nodules, speedily measuring nodule size – a process that was slow when done manually – and appreciated its tools for report generation. They also became quickly aware of instances where the machine performance was less reliable – for example in identifying nodules at the edge of the lung or adjacent to heart vessels. These experts routinely measure their own performance on standard training sets and subjected the tool to similar scrutiny. In this way they assured themselves of instances where they could rely on the machine – which they were happy to delegate to the tool as it freed their time to focus on more complex presentations which required human judgement where the algorithm proved less reliable. In this way the radiologists were able to subject VLN performance to forensic scrutiny even though they were not able to see how the system derived a particular recommendation. Thus, the radiologists were able to quickly detect the aforementioned problems with the tool's performance due to inappropriate configuration.

The radiologists in domesticating the technology established selective ways of using the technology that they felt were accountably appropriate and reliable. The radiologists retained responsibility for decisions. They used the tool as a second or concurrent reader - a "second pair of eyes in the search for lung nodules on CT scans".[7] This could be seen to constitute a model for responsible use of the technology – based on robustly understanding its performance rather than an ability to scrutinise how the algorithm works and how particular interpretations were made. This reinforces observations that in medical practice, accuracy may be a more important warrant of trustworthiness than explainability [4,36]. Our highly expert respondents – who came mainly from specialised tertiary treatment centres – also articulated concerns that users with less experience/expertise (e.g. general radiologists from district hospitals; radiographers) would be less well-equipped to scrutinise the performance of the tool and more likely to align uncritically with the tool's recommendations, right or wrong. This leads on to a further implication, which the project we studied hopes ultimately to explicate, that tool performance will vary between sites depending upon the specific division of labour, existing knowledge, and workflows.

As well as informal appraisal of the tool by health professionals, hospital organisations subjected the VLN tool to internal appraisal. For example, the Chief Information Officer in an inner-city hospital expressed concern that the training data set may not adequately represent the specific demographics of their multi-ethnic community.

Participants in the funded trial benefitted from free access to the VLN tool. However, they were aware that under the normal commercial contract, tool use was reimbursed on a fee-per-scan basis. Hospital managers were concerned that these charges might not be sustainable given the set fee they received under the NHS for each patient screened.

3.3.5. Concluding remarks

The rollout of even relatively mature AI tools like VLN constitutes an ongoing experiment in which the utility and performance of the tool need to be established (in the dual sense of being both achieved and demonstrated). The vendor and adopter involved found themselves in a joint process of learning how to implement across varying organisational contexts as they struggled to address challenges surrounding integration with existing technical infrastructures, tool configuration, information governance etc.

The performance of the tool was affected by a number of specific features including the division of labour and skill and flow of work in particular hospitals. However, there was only limited opportunity in the course of this study to explore in detail how the tool performance was influenced by the specific arrangements for human-AI interaction. This is a topic that requires further research.

This case does not support the widespread presumption that the black boxed character of current diagnostic AI would present problems for clinicians using these tools that could only be overcome by reliably explainable AI [37]. Instead, the expert radiologists we interviewed were able to scrutinise performance of tool in forensic detail. Pragmatic accuracy in real life seems to be accepted by clinicians as an acceptable warrant for the trustworthiness of an AI tool.

The fact that the rollout and evaluation in everyday use and at scale across diverse provider organisations to this tool and a number of similar AI implementations were conducted with public financial support from the UK NHS AI lab highlights the potentially challenging costs of evidence collection not just to meet safety regulatory requirements but also to satisfy procurement compliance guidelines. This tool was geared towards a large scale, screening process, meeting the formalised professional diagnostic requirements of the British Thoracic Society. If the roll-out of a tool to meet these large-scale standardised decisions needed public support, then serious questions would arise about how to cover the costs of developing, validating and sustainably providing stand-alone AI tools to support smaller scale treatments.

3.4. Case 4: Hospital AI procurement strategies: from apps to platforms

3.4.1. Introduction

This case explores an early AI procurement process at a large Norwegian public health service provider, including four hospitals [38]. forthcoming). The process was initiated by the image diagnostics

[6] Thus in the specific use case that this project addresses, the UK National Institute for Health and Care Excellence (NICE) had recently concluded that "There is not enough evidence to recommend AI-derived computer-aided detection (CAD) software alongside clinician review of CT scan images to detect and measure lung nodules in, or outside of, targeted lung cancer screening" [47]. AI-derived computer-aided detection (CAD) software for detecting and measuring lung nodules in CT scan images. National Institute for Health and Care Excellence. Retrieved 3. August from https://www.nice.org.uk/guidance/dg55

[7] https://www.aidence.com/veye-lung-nodules/.

department, and the aim was to procure up to six CE-marked[8] AI applications for radiology and image diagnostics. This ambition was motivated by a constant growth in examinations and radiologists struggling to keep up with the workload. In this context, AI technologies were seen as tools potentially easing the radiologists' work and making the diagnostic workflow more efficient. In the course of the procurement, several vendors proposed an AI platform as a more flexible and efficient way of adopting AI than selecting specific AI tools. The hospital eventually decided to adopt this, previously unanticipated, procurement strategy.

3.4.2. The case

In early 2020, the image diagnostic department started the procurement process by conducting pre-phase investigations, looking at AI applications available in the market and the deployment status of such applications in real-world clinical settings nationally and internationally. During this initial phase, they involved different stakeholders, such as the hospitals' radiologists, another health service provider, and the regional health ICT service provider. The radiologists were involved through workshops aiming to find and select use areas where AI could be useful in their work practices. Around 55 use areas were identified, from which six were chosen for procurement on the grounds of clinical importance and the prospects for applying AI: 1) CT Thorax for lung nodules, 2) Pulmonary emboli, and 3) lung metastasis, 4) Brain MRI for Multiple Sclerosis, and 5) Conventional X-ray for skeletal and 6) chest. These were included in the invitation to tender, along with a set of preliminary requirements, including that the AI applications should a) be fully trained, b) be validated and in clinical use in Europe, c) be possible to integrate into different Radiology Information Systems (RIS) and Picture archiving and communication systems (PACS) from various vendors, d) not use the hospitals' images for development or training the algorithms, and e) contribute to increased quality and efficiency of diagnostic work. Based on the responses to the invitation, five vendors were pre-qualified for participation in the further procurement process.

Eventually, a competitive dialogue[9] procurement project was carried out with a team of 16 members: two head physicians from local image diagnostic departments, two procurement officers (from both the national and the local departments), three IT officers (from both regional and local departments), five physicists (including the project manager and three representatives from the collaborating health service provider), and two officers working within personal data protection, legal issues, IT security and ethics. Additionally, a handful of dedicated radiologists were involved. However, as time-pressured healthcare professionals, they only attended sporadically and when most urgent. The procurement project also collaborated with the regional health authorities, the Norwegian Directorate of Health, the Norwegian Medicines Agency, and several other health service providers. The latter altogether covers over half of the Norwegian population. As the procurement project aimed for a framework agreement, the other health service providers showing interest had the opportunity to enter the contract and benefit from not needing to go through a procurement process themselves, as well as drawing on the experiences from the validation process carried out by the initiating organisation (e.g., their assessments of whether the platform or applications are compatible with existing equipment, information systems and workflows).

3.4.3. Issues in technology selection

The project's initial aim was to buy up to six AI applications for the above-mentioned 1)-6) imaging areas, as prioritised by the radiologists involved in the initial process. During the dialogue meetings with the vendors, assessments of the AI applications were undertaken according to the health service provider's needs and early requirements. As this process proceeded, lessons were learned, and the project's focus and aim changed accordingly. Most significantly, a shift occurred as the procurement team became increasingly aware that the performance of the AI applications, as presented by the vendors, would not necessarily materialise as expected. This uncertainty arose given the limited evidence available about actual benefits across differing contexts and conditions of use. Additionally, they realised that procuring and verifying stand-alone AI applications individually would take time, require many resources and be costly, as well as potentially locking them into early solutions that might quickly be outperformed. A member of the procurement team explained the situation and why they changed their mind as follows:

> What we found during the dialogue was that the algorithms that exist today … it's not sure that they are optimal for us. It's not sure that they will give us what we hope for. It's very uncertain. I know that the documentation concerning value is scarce. At the same time, we know that there is a rapid, ongoing development where new algorithms are developed and become available, so there is a potential in the future to implement new algorithms as time goes by. New and better algorithms that do not exist today or are further developments of the ones available today. (Procurement team member, MK's translation)

However, this realisation did not stop the procurement process from continuing, as another option had been introduced during the dialogue meetings. During the first round of meetings, although this had not been requested, four of the five vendors also promoted their more or less developed platforms for AI applications. These were described as cloud-based solutions that make AI applications from various third-party vendors available to users through a single contract. The vendors typically explained them as marketplaces, comparing them to well-known commercial platforms such as Apple's Appstore or Google Play. Given the uncertainty of whether specific AI applications would benefit local clinical practices within the health service provider, the procurement team became convinced that a platform, potentially enabling the adoption of a wider variety of current and future algorithms, was a more adequate procurement strategy:

> So, it is in many ways a strategic choice [buying a platform]: to be able to implement new, exciting things fast in the time to come. And if we implement one algorithm, and the infrastructure connected to one algorithm, it is necessary to have a technical set-up [for each application], we will have to make a ROS [risk and vulnerability analysis], we need to have a "solution design", we have to make a procurement [each time]. And the extra work to make a platform accessible is so small and limited. And as most of the vendors had a platform solution, we saw this as an appealing opportunity. (Procurement team member; MK's translation).

3.4.4. Concluding remarks

Through the platform option, the procurement team identified an opportunity for greater flexibility in their processes of choosing applications for implementation now and in future, as an increased range of AI tools and updates to existing tools were expected to become rapidly

[8] CE-marked products indicate that they meet certain requirements from the EU concerning safety, health and environmental protection. Getting products CE-marked can be necessary for entering the European market. AI applications intended for humans in healthcare (e.g., for diagnostics, treatment, or monitoring) are considered a Class II category within the Medical Device Regulations. This means that CE-marking can only be obtained through a 'notified body approval' (as opposed to other products where a self-declaration from the manufacturer may be sufficient for getting the CE-mark) [39]; p. 361; [16].

[9] A competitive dialogue is a public procurement procedure used in many European countries, beneficial when procuring complex, innovative solutions and requirements are not yet possible to specify. Through dialogue with tenderers, an overview of what is on the market can be gained, and solutions for procurement can be adjusted according to the clients' needs. See, for example, https://www.procurementjourney.scot/route-3/develop-strategy/procurement-routes/competitive-dialogue.

available. In addition, they viewed the procurement of a platform as a way of reducing the time, staff effort and costs involved in procuring future AI applications, which had proved to be time-consuming and expensive to conduct on a case-by-case basis. Furthermore, with the platform offering access to a larger variety of existing and new applications, a scaling opportunity emerged beyond what the team had initially envisaged. Instead of assessing single AI applications for specific imaging areas, the platform offered a solution that potentially made it possible to test various applications and keep only those that seem beneficial and applicable for the health service provider's different clinical workflows, patient data and so on. Thus, the outcome of the procurement process not only extended the original aims. It added an infrastructural dimension to the project with more far-reaching consequences for radiologists' present and future work practices (e.g., due to the availability and the potential of bringing more AI applications into the workflow). The initial focus on which AI applications to procure had evolved, apparently seamlessly, into an experiment to establish mechanisms to broaden the scope and increase the scale and pace of uptake of AI applications in the health service provider's hospitals. This shift in focus from implementing stand-alone AI solutions towards establishing new institutional and technical machinery for procuring, implementing and validating an evolving flow of tools highlights processes of 'learning by interacting' and 'learning by governing' [16] among these and other stakeholders.

4. Analysis

We distinguish ***three*** key (albeit interconnected and overlapping) processes of social learning for diagnostic AI, building upon Sørensen's typology:

1) **Coupling AI with clinical settings:** (learning by doing) encompassing early experimental efforts to create effective models and demonstrate their potential to enhance clinical practice; establishing ideas about use cases and potential clinical and commercial value. These efforts are typically based on collaboration between AI specialists and clinicians. Very large numbers of such hopeful experimental collaborations have emerged but very few of these culminate in successful, sustainable innovations [39].
2) **Embedding, scaling and extending scope** (learning by doing/learning by interacting) Successful applications must be implemented (and further developed and optimised) in contexts of everyday use to demonstrate their clinical effectiveness and their commercial viability. Large-scale adoption across several settings is required to collect the evidence needed to satisfy clinical governance (effectiveness and safety) requirements of regulatory bodies and health provider organisations, validate adopter's business models and satisfy procurement requirements.

Our investigations into embedding AI at scale also encountered subsequent efforts to move from 'discrete' applications of specific tools for particular conditions and extend the scope of AI applications through procurement of platforms offering multiple tools for diverse conditions. This led, in turn, to a third form of (second order) reflection and social learning:

3) **Developing collective frameworks for appropriate implementation and use** (learning by regulating)These early skirmishes around the domestication of specific AI tools have heralded in a broader period of social learning and experimentation about the institutional arrangements for effectively (safe and clinically effective, cost effective) applying AI in health settings [4]. Our studies highlight informal regulation/domestication by adopting health provider organisations/clinical professions seeking to verify the sustained performance of tools in real-world use. A separate paper discusses the reworking of national formal governance arrangements that is taking place alongside these developments geared toward prior approval of rapidly changing AI tools and ongoing Post-Market Surveillance of their everyday use [40].

4.1. Experimental innovation of hopeful applications: coupling AI with clinical settings

The huge generic expectations surrounding AI in health (and other fields) are driving a plethora of hopeful innovations [41,42]. However, there is as yet no general model for applying AI in health diagnosis. Those involved in experimental developments must navigate a challenging array of choices as illustrated by case 1 and case 2. The initial challenge is to build a model whereby an AI tool can help make distinctions that matter in clinical decisions. The initial search may be conditioned by practical exigencies, for example, by the availability of data sets or by the concerns of health professionals involved. In practice, however, the choice of target disease area may be driven by contingent factors – for example, the specific interests of clinicians drawn into collaboration with an AI specialist or the availability of data. The significant costs of collecting, cleaning and labelling data represent a key constraining factor for AI specialists. Most experimental projects, lacking the substantial financial resources required to create data sets, resort to already existing and readily accessible data repositories. A highly visible feature here has been the role of competitions which have been very successful in motivating the AI community to create algorithmic models using standard data collections.[10] However, the uptake of these models remains very low [43]. The current commercial and policy focus on these kind of short-term experiments arguably diverts attention away from the hard work needed to improve the performance of algorithms (for example, by focusing pattern recognition algorithms on clinically relevant parts of a scan) and get them to perform in ways that can support diagnostic decisions – let alone apply them in clinical practice.

Even though these 'upstream' experiments may generate multiple hopeful solutions, they still lack a clear pathway to sustain and scale. The creation of an algorithm that can support a diagnostic decision is only the start of a protracted innovation process. Start-ups, and especially SMEs with limited resources, struggle to stay in the game. They need to secure the much larger investments needed to progress from the initial proof of concept and integrate/deploy it in real-world clinical settings to demonstrate its safety and potential benefits. Much work has to be done to refine their embryonic solutions and bring them to the market –a challenge which, given the huge attrition rate, is described as the 'valley of death' [44].

Technology adopters are concerned regarding the clinical efficacy of the myriad of new offerings and their benefits for the quality and efficiency of care. Technology suppliers frequently articulate compelling claims about the capability of their products, often by producing numerical comparisons of their performance with physician performance. Claims of AI tools (trained on trial data sets) out-performing human judgement have become a legitimatory currency - apparently providing 'public-proof' of the capability of AI tools. However, experience shows that tool performance in laboratory settings is not a reliable indicator of real-world performance – which needs to be established in settings of everyday care. The very low rate of sustained uptake of new offerings in clinical practice has resulted in guidelines and regulations encouraging better external validation. However, there is to date a dearth of systematic evidence of the clinical and other benefits and challenges of diagnostic AI [45]. Thus, in a survey of 100 CE marked AI products in commercial use in Europe, 64 lacked evaluation data and only 18 had systematic evidence about application performance and efficiency/cost savings [46].

These issues have particular salience in relation to diagnostic AI. The

[10] For example https://www.med.upenn.edu/cbica/brats2020/data.html.

performance of an algorithm appears to be fragile. It may change as a product moves from its original context of development and initial implementation and is applied to different target populations in other settings – which may differ in terms of their demographics and health profiles [47]. In addition, as our third case highlights, the configuration of a tool within particular technological infrastructures (internet, scanners etc.) and organisations (e.g. care pathways, divisions of expert labour) also affect its performance. Since real-life tool performance cannot be verified in the developer's laboratory, we have seen a shift of attention in AI tool experimentation from development settings to contexts of early implementation, where developers and early adopters become involved in a joint process of *learning and demonstrating how AI tools may be productively and reliably deployed in health settings*.

4.2. Embedding and extending scale and scope

Creating a commercial product is thus just the opening stage in a struggle to create conditions where an AI tool will be widely and sustainably adopted and become commercially successful. Developers/vendors need to install their products and get them to work effectively in multiple real-world settings to create the evidence needed to convince regulators and potential adopters of the safety and effectiveness of their products [29] and establish viable business models to enable their wider adoption [48].

Our vignettes encompass two attempts to extend respectively the scale and scope of AI adoption: i) to embed and extend adoption/use of a tool across a number of settings to make these available at scale on a sustainable basis (case 3); and, ii) to extend the application of AI across a growing range of clinical areas and activities (case 4).

i) Embedding, demonstrating sustained dependable use of specific AI tools and extending the scale of adoption

Embedding AI tools in health settings involves not just their integration within an existing infrastructure of scanners, viewers, wireless and broadband networks but also the establishment of accountable and productive methods of using technologies in routine clinical care. After successfully implementing individual tools and demonstrating their effectiveness in routine care in initial locales, attention shifts to extending the scale of adoption and demonstrating sustained, dependable use. Health technology developers need to provide evidence about performance and safety of their products to various audiences: safety regulators, clinical governance bodies; health provider organisations may have additional clinical governance and procurement criteria; clinical specialists and their professional organisations. As well as demonstrating product safety, evidence is sought about the performance of applications required for clinical and informal professional governance but also to enable cost-benefit assessment (e.g. assessments conducted by the UK National Institute for Health and Care Excellence [NICE][11] and hospital business plans and fulfil health service procurement requirements.

ii) Extending scope – *Learning how various AI tools may be optimally exploited across health service settings*

The time and work required to develop data sets to train AI algorithms and then conduct multiple trials of their effectiveness to satisfy complex governance requirements presents a challenge to tool developers. This includes gathering the evidence required not only for initial approvals but also for continued scrutiny over product performance as the product diffuses and evolves and its use is optimised. Indeed, the continued evolution and refinement of AI tools will require sustained post-market surveillance to fulfil clinical and safety governance requirements, representing a significant ongoing cost. These costs may be prohibitively high in relation to available income streams – particularly for SMEs, accentuating concerns about reimbursement: how soon can new ventures secure sales/licensing income to recover their investments [44]. It may be hard for vendors to cover these ongoing costs simply with initial procurement/license costs. Here we note ongoing experimentation to establish viable business models for service delivery, including the growth of Software-As-A-Service models based on fee recovery per use.

Adopters face parallel challenges in evaluating, procuring, implementing and certifying diverse vendor offerings. Implementation and configuration of specific tools is made more complex by the need to integrate them and establish interoperability with other complementary components (e.g. scanning equipment, picture archiving and communication systems [PACS]). Diversity in the infrastructure and installed equipment across the estate can accentuate interoperability challenges and may create implementation problems and affect performance. Such problems may recur as a result of periodic upgrading of hardware and software. The high transaction costs in terms of time/effort of procuring AI tools on case-by-case as well as charges for commercial tools, may make it difficult for potential adopters to demonstrate a positive Return on Investment and establish a business case for adopting stand-alone specific applications. Our study also encountered concerns within health provider organisations about how to meet usage charges within their service funding models.

These challenges generate powerful incentives to achieve economies of scale and also economies of scope - twin vectors surrounding the implementation of AI tools.

Suppliers and adopters have begun, separately and together, to try to work out mechanisms for procurement, clinical validation and safety certification and establish viable business models for supply and ongoing scrutiny over their performance [49]. These underpin recent developments in the diagnostic AI sector, including:

a) Ongoing vendor mergers and acquisitions reflect pressures towards the convergence of capabilities within a specialist supply sector – creating vendors able to cross-sell solutions across multiple disease areas.
b) Concerted procurement and adoption of AI tools by hospitals/health service providers to reduce costs of search, evaluation and contract negotiation and ease implementation;
c) The emergence of platform providers – offering internally-developed and acquired solutions as well as, in some cases, allowing third-party providers to mount their solutions onto the platform [41]. We encountered in the course of this fieldwork various attempts to more radically extend the scope of application by the development of platforms supporting an expanding array of AI tools for various requiring investment of less time, effort and money than repeated stand-alone procurement/implementation of point solutions for particular conditions [38]. forthcoming).

Actors from different locations in the supply chain are seeking to position themselves as centres for this convergence. For example, major equipment makers are seeking to embed AI functionality and more generic AI toolkits in their products (e.g. Canon, GE,[12] Philips, Siemens) as are established Health IT vendors. Here we note the recent flurry of Merger and Acquisition activity as larger players seek to position themselves as suppliers of a range of diagnostic AI tools rather than single-point solutions. Thus, the supplier of the lung cancer scanning tool in case 3 merged with a prostate cancer and brain imaging tool supplier before being taken over by a large US outpatient diagnostic

[11] See for example https://www.nice.org.uk/about/what-we-do/digital-health/multi-agency-advisory-service-for-ai-and-data-driven-technologies.

[12] https://www.ge.com/news/press-releases/ge-healthcare-accelerates-ai-model-development-and-deployment-launch-edison.

imaging service provider which had a mammography tool.

Our study also highlights a more radical shift in modes of AI procurement and implementation away from the adoption of specific standalone tools towards 'platforms', variously conceived, through which a wide and changing array of specific tools may be procured and implemented. We note the emergence of new kinds of commercial intermediaries, notably AI marketplaces, open to diverse supplier offerings [50], some of which project themselves as non-proprietary platforms, as well as attempts by scanner and PACS providers to offer bundled solutions to their clients. These different models, as exemplified by case 4, represent differing ways of configuring the AI procurement and implementation arena. They offer radically different commercial opportunities for entrepreneurs, SMEs and large IT or pharmaceutical corporations with different levels of agency for health providers (individual hospitals or groups of hospitals). Other modes of supply may be possible, as well as oligopolies, closed platforms, open platforms and public platforms. At this stage, it remains unclear which modes of provision will prevail and which players will succeed in establishing themselves at the centre of the emerging diagnostic AI ecosystem. Further research is urgently needed of these developments (which our group seeks to contribute to).

4.3. Developing collective frameworks for appropriate implementation and use

Our comparative investigation has highlighted a trajectory in the ongoing social learning processes amongst clinicians, provider organisations and technical specialists. The trajectory begins with local *learning by doing* to create models that bear upon diagnosis and then demonstrating their effectiveness and reliability in everyday care contexts and moves on to include reflexive processes of *learning by interaction* to address questions about how to adopt AI tools more widely. This begins with a focus on scaling up: extending the use of these point solutions across multiple settings and ultimately moves on to extending scope beyond point solutions towards an increasing range of conditions which in turn draws attention to developing new institutional arrangements that can reduce the costs of procuring, implementing, validating and sustainably using diverse kind of AI across multiple conditions. The process culminates in attempts to establish new models of formal and informal governance better aligned with the challenges of implementing and further innovating AI tools (*learning by regulating*).

The success of AI innovation in the digital economy has been driven by rapid cycles of development and implementation. This innovation model may pose challenges to our existing precautionary system of evaluating and regulating medical devices originally designed for collecting evidence about the performance of specific physical artefacts. However, the properties and behaviour of software and especially AI-based artefacts, cannot be readily/reliably demonstrated, for instance, by inspection but must be proved in everyday use. Moreover, AI performance may change from one application setting to another. Health provider organisations also seek to verify (and periodically reappraise) the performance of tools they adopt [8]. We finally also note attempts to rework the existing regulatory apparatus in place for medical devices to effectively engage with the distinctive features of AI-based innovations, notably: i) their performance may be fragile and may decay and vary between contexts of use; ii) they can evolve rapidly and improve through multiple cycles of development and operations). These features mandate in favour of a shift from precautionary *a-priori* regulation to evolutionary risk governance frameworks catering for anticipated changes in the model following on initial safety/effectiveness certification and extending to post-market-surveillance [40].

5. Discussion

5.1. Expectations misaligned with recent achievements

The recent successes of AI in the digital economy have dramatically (re)inflated expectations about the scope to apply AI in health and biomedicine. Driven by the convergence of clinician hopes, innovator ambitions and public health policies and industry strategies, a generic vision has emerged of the applicability of machine learning and AI tools across the full range of health activities from diagnosis to health system management and biomedical research [46,48,51]. However, these generic and over-hyped generic expectations are out of alignment with the uneven specific achievements of current AI tools implemented in real-world settings [32,38]. forthcoming). As a result they do not provide a good model of how to develop and apply medical AI let alone guidance about the challenges that need to be overcome. Health service planning and public health, which offer potentially important early benefits, have been somewhat eclipsed by intense interest in AI tools for clinical diagnosis and treatment [52]. The application of AI tools in healthcare delivery and in biomedical research has been somewhat conflated. Thus, it is frequently suggested that everyday use of diagnostic AI tools will enable the detection of clinically relevant features not apparent to clinical specialists [1]. Though clinical use of AI may indeed open up longer-term opportunities for such biomedical research, these will not emerge spontaneously through healthcare delivery but will require longer timeframes and particular organisations. Successes to date, notably in diagnostic AI, have largely been in "Narrow AI" [42] Kumar, Chauhan & Awasthi, [29]).). Thus, early applications of diagnostic AI notably in mammography, later extended to other cancer screening programmes and other major disease areas (neurology and cardiology) have offered specific 'point' solutions in the care of specific conditions [53]. Much tool development activity hitherto has focused on areas where medical signs are well characterised and their diagnostic implications well established with clearly designated disease management pathways. In addition, contemporary diagnostic AI applications largely seek to emulate some parts of currently established clinical workflows and have particularly targeted common diseases with large-scale screening programmes where a division of expert labour has enabled specialist occupations such as radiologists and radiographers to emerge. In addition to providing candidate expert decisions that can be amenable to AI this model offers convenient opportunities to build cost-justifications by promising to increase the productivity of extremely busy, well-paid specialists, projecting cash-releasing benefits for the business modelling of provider organisations. These settings provide a way for developers to resolve the key challenge they face of securing reimbursement for the costs of developing/sustaining applications [54]. AIs perform well with relatively closed systems – for example, in medicine, in areas where there are clear clinical signs well-correlated with specific diagnoses, courses of treatment and medical outcomes. In contrast, AI struggles to deal with more open-ended settings where existing medical knowledge offers less clear correlation between signs and outcomes. Where there are more variables, escalating levels of training data may be required to reliably resolve this relationship. Epistemic open/closed-ness varies substantially between different medical conditions/fields of medicine, shaped by the current state of medical knowledge.

5.2. The domestication/appropriation of AI apps within clinical settings

Much existing social science and ethical literature focuses on the lack of transparency of 'black-boxed' commercial AI solutions and the difficulties this poses for lay users, unable to interrogate algorithmic decisions [36,55,56]. In contrast, our findings – especially of the adoption/use of the lung nodule scanning tool (case 3) - suggest that emergent AI technology is being appropriated and domesticated by health professionals and provider organisations in the course of its

procurement, implementation and clinical validation. In this process the performance of AI applications is being subjected to critical scrutiny and is being used in an accountable and de facto responsible manner.

Our cases focus on the implementation of AI tools in elite medical centres and teaching hospitals amongst specialist expert communities which exhibit a high level of expertise and job control. These AI tools (like other new clinical technologies) are subjected to detailed scrutiny by the provider organisations and professional groups using them. This was especially marked in the case of radiologists (see case 3). Though not computer scientists, radiologists and other specialised occupations (radiographers, medical physicists) have grown up around the application of new medical diagnostic technologies and their subsequent digitisation. This strictly regulated field of medical expertise is already subject to very high levels of internal quality control. The radiology profession already exercises a high level of numerical etc. scrutiny over members' professional performance. They are therefore very well equipped to subject the performance of AI tools to forensic scrutiny. Thus, they assessed the performance of AI tools on test data sets in the same way they assess their own performance. These confident and highly expert users rapidly become familiar with the strengths and limitations of current AI, identifying cases where AI tool is unreliable and those tasks which they may reliably delegate. They are in consequence relatively well-equipped to exercise a level of control over the way in which technologies are deployed (at least in the European public health settings we studied) – and retain full responsibility for judgements made with their assistance [57]. Paradoxically, our experts were concerned that users with lower levels of expertise may be less confident and less well-equipped to assess the performance of an AI tool and, thereby, more likely to rely on the decision-support it provides and less able to repair its deficiencies! This kind of *pragmatic trustworthiness* may not emerge, the ECG case (case 2) suggests, where clinicians experience existing tools as unreliable or are unable to directly verify an AI tool's future performance (e.g., future predictions) [58].

Though much debate about algorithms emphasises the informational asymmetry between AI specialists (with their black-boxed algorithmic systems) and lay users, this appears to be mitigated in the medical domain, whose extensive bodies of expertise are embedded in ancient and strongly entrenched institutional/professional structures. These professional groups have not immediately ceded ground/authority to vendor claims but have instead subjected the performance of vendor offerings claims to careful scrutiny. These experts anticipate benefits from applying these tools but insist on retaining responsibility (and liability) for diagnostic decisions. In this process, we suggest, technologies and associated promises articulated 'in the wild' of technology developers are being tamed and domesticated into the world of clinical practice [59]. Radiologists and other highly expert users are, we suggest, not necessarily 'performed' by AI. *Au contraire,* they work to domesticate AI – they seek to establish ways of appropriate, accountable and responsible use of AI-based tools (Farič et al., 2023, 2023).

The aggressive technology supply rhetorics – including highly publicised claims advanced by AI specialists like Geoffrey Hinton,[13] often emphasising the superior laboratory performance of AI tools over humans, have arguably diverted attention away from examining and providing evidence of the real-world performance of AI-based systems. Their zero-sum conception of the relationship between human and machine intelligence has inhibited crucial enquiry into how to establish productive and reliable combinations of machine and human capabilities that can make best use of their different affordances (strengths and failure modes). The careful use of the lung nodule scanning tools in case 3 suggests that clinical professionals will be willing to increasingly delegate to the machine those tasks where the machine proves to be effective and dependable. In this way, they may be expected to develop practices and conventions for a form of accountable, indeed *responsible* use of AI. However, supplier rhetorics have arguably inhibited productive feedback from these key experimental settings where health professionals have learnt how to make the best use of AI tools. The current roll-out and large-scale use of AI tools could provide an important de facto 'living laboratory' for learning how to design and implement tools that support reliable and responsible use.[14] However, this valuable learning needs to be systematically collected and shared (e.g. through effective evaluation of deployment and continued post-market surveillance of the use and performance of AI tools).

6. Conclusions

This investigation was made possible by the fortuitous alignment of a diverse sample of studies of early diagnostic AI amongst our research community. We have been able to chart the opening moves of a range of projects in various medical fields and at different stages of their innovation journey. We have tracked their often difficult and unpredictable journey. We show initially how AI diagnostic tools are constructed by applying Machine Learning and other techniques to available data sets and then optimised to make distinctions relevant to clinical decisions. We next explore how these tools subsequently need to be implemented and assessed in everyday clinical use before they can be validated and adopted at scale. Our vignettes have highlighted the experimental character of both upstream emergence and downstream embedding of AI tools as the actors involved struggle to improve, demonstrate and sustain tool performance. Their arduous and uneven progress stands in contrast to widespread expectations of rapid, pervasive adoption of AI across the health sector. One of our observations is that current promissory visions of medical AI are out of alignment with actual achievements to date and do not provide an effective or useful account of the challenges involved.

This brief overview has highlighted significant variations in the state of AI development across different fields of medicine. This includes the focus of investor interest and clinical enthusiasm on diagnostic AI rather than, for example, health service organisation (which arguably offers bigger health benefits) and, within this, a focus on areas where the state of medical practice provides a foothold for developing effective tools particularly where large-scale screening programmes offer scope for profitable commercial provision.

We have highlighted the brittleness in performance of AI tools as they move from their initial application setting to other contexts of use, which differ, for example, in their technical infrastructure, professional and organisational arrangements and patient populations, which inter-alia underpins the need for ongoing post-market surveillance.

We have highlighted the 'domestication' of AI. This was observed as a specific tool was rolled out for identifying and measuring potentially cancerous lung nodules in CT scans. Health professionals subjected tool performance to critical scrutiny – judging it with the same criteria they apply to their own work. By establishing areas where the tool performed reliably and freeing up time, the attention could be focused on the ambiguous edge cases where the algorithm was less reliable. This constitutes a form of de facto *responsible use* of AI-based tools upon scrutinised empirical performance rather than making algorithm operation transparent or explainable. Finally, we noted a shift in the focus of social learning activities from the struggle to develop and prove the effectiveness of specific AI tools towards the creation of institutional mechanisms and infrastructures to support the sustainable procurement,

[13] In 2016 deep learning pioneer Geoffrey Hinton famously said *"people should stop training radiologists now - it's just completely obvious within 5 years deep learning is going to do better than radiologists. It might be 10 years, but we've got plenty of radiologists already"* (cited in Ref. [15].

[14] The recent UK NHS *Artificial Intelligence in Health and Care Awards,* which supported qualitative and qualitative analysis alongside large-scale tool adoption, can be considered exemplary in this respect. https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/.

implementation and maintenance at scale of a burgeoning array of AI applications.

6.1. Strengths, limitations and scope for further research

Social scientists find themselves repeatedly invited to come to some conclusion about the character and implications of novel developments before the technology matures and its impacts can be established. This is particularly important in relation to medical AI – a rapidly changing field that is subject to powerful competing visions and expectations. How can researchers make useful assessments of the prospects and implications of still unfolding innovation processes in the face of these conflicting claims? Existing research frameworks are not well-equipped to assess the prospects and consequences of rapidly emerging technology. Short-term snapshot studies of particular settings of early implementation can generate partial or misleading accounts of the prospects and social implications of technology [60]. To avoid these pitfalls, we have sought an evolutionary understanding of how technology and its uses are shaped, drawing upon Sørensen's Social learning (1996) and domestication (1994) framework.

This preliminary investigation has exploited the opportunistic alignment of a number of relatively modest studies within our research group. We were able to derive comparative insights from these independently initiated investigations by utilising our broadly shared theoretical perspectives and concerns. However our sample only covers a limited array of early initiatives To understand the potential and (uneven) performance of diagnostic AI, we need to extend this work across a wider range of settings and also longitudinally, as AI tools become more widely adopted and embedded. Studies need to encompass the full range of actors involved and affected groups (notably of patients and carers, not studied by us and currently under-represented in research). A more detailed focus is needed upon the integration of AI tools within particular workflows, professional practices and divisions of labour [12]. Our finding of the responsible use of AI tools poses the question of whether we can identify specific combinations of human and machine intelligence that prove efficient and reliable (offering a positive-sum model of human-machine interaction rather than the crude zero-sum game anticipated by Hinton and similar rhetorics of technology supply).

AI-based tools are evolving rapidly in their emergence and early adoption and will continue to change across multiple development-operations cycles. We need to examine these artefacts and their associated practices of development and use in multiple points of time and space. Our investigation picked up the early stages in the emergence of new arrangements for large-scale deployment of a broad and changing array of diagnostic AI tools. The picture changed significantly in the course of our research and it will be important to address the competing pressures shaping these developments and the different trajectories that may emerge. We note that our studies focused on research-oriented and highly expert elite settings of public medicine in Northern Europe. Different conceptions about AI deployment may emerge in other settings (e.g. contexts of commercial medicine where pressures to fragment and commodify services may dilute or by-pass the control currently being exercised by skilled clinical specialists). Further work is needed across a wider range of international contexts to address the influence of differing institutional, technological and cultural settings [61].

We conceive these studies as the initial stage of a longer-term enquiry into medical AI, informed by the Biography of Artefacts and Practices perspective [62]. This offers methodological guidelines and conceptual frameworks to go beyond short-term snapshot case-studies and extend detailed ethnographic enquiry longitudinally and across multiple locales. It highlights the need in this case i) to examine the spread of technologies across multiple sites and, ii) to conduct detailed studies across the technology lifecycle from development to initial implementation and subsequent extension and scaling of adoption. Over time, multiple studies can yield a more adequate biography of AI artefacts and practices over an extended timeframe.

As well as these medium-term developments, scholarship should engage with longer-term and systemic changes in health systems as AI tools become more widely adopted. Longer-term investigation is needed to address ways in which health service delivery and, for that matter, medical expertise and practice may evolve alongside the widespread adoption and embedding of AI tools.

6.2. Implications for policy and practice

This paper has direct implications for public policies to promote the development and uptake of AI tools. Public support, currently focused on initial innovation, should be extended to the whole life-cycle – including full-scale and sustained uptake and use. Experimentation involving technical and clinical specialists is not limited to initial model building but continues as tools are rolled out in real life and as technology use is scaled up. However, though many real-world experiments are currently happening as AI is deployed, the valuable learnings from these experiences are impaired as these distributed experiments are often not systematically evaluated [46]. It is 35 years since Fleck highlighted the importance of innovation in the arenas of technology implementation, and warned that painfully and expensively acquired adoption experience was not being systematically captured and applied elsewhere [21]. This is better recognised today, particularly in relation to AI, as exemplified by the NHS award/" title="https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/">AI in Health and Care Awards. However, collecting and sharing such implementation experience remains the exception rather than the rule. New forms of evaluation may be required to track the rapid evolution of these emerging technologies. Crucially evaluation needs to continue as a technology is rolled out and further innovated in use [63]. This will require new methodologies and infrastructures for efficiently and reliably achieving post-market surveillance [8].

We have documented some key early moves in the establishment of new institutional arrangements for the development, procurement, implementation, spread and dependable operation of AI tools and their extension across multiple conditions and forms of activity. Further research is needed to track the large-scale social experiment of embedding artificial intelligence, in conjunction with human expertise, in our health systems.

Funding sources

Case 1 was supported by the United Kingdom Research and Innovation (UKRI) Centre for Doctoral Training in Biomedical AI at the University of Edinburgh (grant number UKRI EP/S02431X/1).

Case 2 has drawn on a program of independent research funded by the Wellcome Institutional Partnership Award (Tier 1 Discretionary Fund) (grant number WT 209710/Z/17/Z).

Case 3 was undertaken with funding from the UK NHS Artificial Intelligence in Health and Care Awards (grant number 2119C25043) and UKRI Trustworthy Autonomous Systems Node in Governance and Regulation (grant number EP/V026607/1).

CRediT authorship contribution statement

Robin Williams: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Stuart Anderson:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Kathrin Cresswell:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Mari Serine Kanneløønning:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Hajar Mozaffar:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Xiao Yang:** Writing – review & editing, Writing – original draft, Methodology, Investigation,

Conceptualization.

Data availability

Data will be made available on request.

Acknowledgments

We also acknowledge the contribution of the wider team contributing to the fieldwork for case-study 3, Dr. Nusa Faric and Dr Sue Hinder.

References

[1] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthcare Journal 6 (2) (2019) 94–98, https://doi.org/10.7861/futurehosp.6-2-94.
[2] P. Rajpurkar, E. Chen, O. Banerjee, et al., AI in health and medicine, Nat. Med. 28 (2022) 31–38, https://doi.org/10.1038/s41591-021-01614-0.
[3] A. Zahlan, R.P. Ranjan, D. Hayes, Artificial Intelligence Innovation in Healthcare: Literature Review, Exploratory Analysis, and Future Research, Technology in society, 2023 102321, https://doi.org/10.1016/j.techsoc.2023.102321.
[4] Y.Y.M. Aung, D.C.S. Wong, D.S.W. Ting, The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare, Br. Med. Bull. 139 (1) (2021) 4–15, https://doi.org/10.1093/bmb/ldab016.
[5] M.G. Seneviratne, N.H. Shah, L. Chu, Bridging the implementation gap of machine learning in healthcare, BMJ Innovations 6 (2) (2020) 45–47, https://doi.org/10.1136/bmjinnov-2019-000359.
[6] M.L. Gordon, K. Zhou, K. Patel, T. Hashimoto, M.S. Bernstein, The disagreement deconvolution: bringing machine learning performance metrics in line with reality. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021.
[7] L. Pumplun, M. Fecho, N. Wahl, F. Peters, Adoption of machine learning systems for medical diagnostics in clinics: qualitative interview study, J. Med. Internet Res. 23 (10) (2021) e29301.
[8] J. Feng, R.V. Phillips, I. Malenica, A. Bishara, A.E. Hubbard, L.A. Celi, R. Pirracchio, Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare, npj Digital Medicine 5 (1) (2022) 66, https://doi.org/10.1038/s41746-022-00611-y.
[9] I. Van de Poel, Society as a laboratory to experiment with new technologies, in: D. M. Bowman, E. Stokes, A. Rip (Eds.), Embedding New Technologies into Society: A Regulatory, Ethical and Societal Perspective, Pan, 2017, pp. 61–68.
[10] C.J. Kelly, A. Karthikesalingam, M. Suleyman, et al., Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (2019) 195, https://doi.org/10.1186/s12916-019-1426-2.
[11] M.Y.C. Nagendran, et al., Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, BMJ 2020 (2020) 368, https://doi.org/10.1136/bmj.m689, m689.
[12] P. Nilsen, et al., Realizing the potential of artificial intelligence in healthcare: learning from intervention, innovation, implementation and improvement sciences, Frontiers in Health Services (2) (2022), https://doi.org/10.3389/frhs.2022.961475.
[13] Danish Ministry of Finance, & Danish Ministry of Industry, B. a. F. A., National Strategy for Artificial Intelligence, 2019. https://en.digst.dk/media/19337/305755_gb_version_final-a.pdf.
[14] Meld. St. 7, National health and hospital plan 2020–2023. Summary, Retrieved from, https://www.regjeringen.no/contentassets/95eec808f0434acf942fca449ca35386/en-gb/pdfs/stm201920200007000engpdfs.pdf, 2019–2020.
[15] NHS. (n.d.). The National Strategy for AI in Health and Social Care. NHS England - Transformation Directorate. Retrieved February 21st from https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/the-national-strategy-for-ai-in-health-and-social-care/.
[16] K.H. Sørensen, Learning Technology, Constructing Culture. Socio-Technical Change as Social Learning, 1996, pp. 18–96 [Working Paper].
[17] J. Brosveet, K.H. Sørensen, Fishing for fun and profit? National domestication of multimedia: the case of Norway, Inf. Soc. 16 (4) (2000) 263–276.
[18] K.H. Sørensen, Technology in Use: Two Essays in the Domestication of Artefacts, 2/29, 1994 [Working Paper].
[19] N. Pollock, R. Williams, Software and Organisations: the Biography of the Enterprise-wide System or How SAP Conquered the World, Routledge, 2009.
[20] R. Williams, J. Stewart, R. Slack, Social Learning in Technological Innovation: Experimenting with Information and Communication Technologies, Edward Elgar Publishing, 2005.
[21] J. Fleck, Innofusion or Diffusation?: the Nature of Technological Development in Robotics, University of Edinburgh, 1988 (4)[Edinburgh PICT Working Paper].
[22] J.M. Wardlaw, E.E. Smith, G.J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T O'Brien, F. Barkhof, O.R. Benavente, Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration, Lancet Neurol. 12 (8) (2013) 822–838, https://doi.org/10.1016/S1474-4422(13)70124-8.
[23] L.M. Prevedello, S.S. Halabi, G. Shih, C.C. Wu, M.D. Kohli, F.H. Chokshi, B. J. Erickson, J. Kalpathy-Cramer, K.P. Andriole, A.E. Flanders, Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions, Radiology: Artif. Intell. 1 (1) (2019) e180031, https://doi.org/10.1148/ryai.2019180031.
[24] M.D.C. Valdés Hernández, F.M. Chappell, S. Muñoz Maniega, D.A. Dickie, N. A. Royle, Z. Morris, D. Anblagan, E. Sakka, P.A. Armitage, M.E. Bastin, Metric to quantify white matter damage on brain magnetic resonance images, Neuroradiology 59 (10) (2017) 951–962, https://doi.org/10.1007/s00234-017-1892-1.
[25] J. Ashburner, SPM: a history, Neuroimage 62 (2) (2012) 791–800, https://doi.org/10.1016/j.neuroimage.2011.10.025.
[26] D. Müller, I. Soto-Rey, F. Kramer, Towards a guideline for evaluation metrics in medical image segmentation, BMC Res. Notes 15 (1) (2022) 1–8, https://doi.org/10.1186/s13104-022-06096-y.
[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, 2015. Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.
[28] H. Stevens, Life Out of Sequence: a Data-Driven History of Bioinformatics, University of Chicago Press, 2019.
[29] P. Kumar, S. Chauhan, L.K. Awasthi, Artificial Intelligence in Healthcare: Review, Ethics, Trust Challenges & Future Research Directions, 2023 105894, https://doi.org/10.1016/j.engappai.2023.105894.
[30] M.C. Lewis, M. Maiya, N. Sampathila, A novel method for the conversion of scanned electrocardiogram (ECG) image to digital signal. International Conference on Intelligent Computing and Applications: ICICA 2016, 2018.
[31] Z.I. Attia, D.M. Harmon, E.R. Behr, P.A. Friedman, Application of artificial intelligence to the electrocardiogram, Eur. Heart J. 42 (46) (2021) 4717–4730, https://doi.org/10.1093/eurheartj/ehab649.
[32] S. Lebovitz, Diagnostic Doubt and Artificial Intelligence: an Inductive Field Study of Radiology Work, ICIS, 2019.
[33] S.K. Mun, K.H. Wong, S.-C.B. Lo, Y. Li, S. Bayarsaikhan, Artificial intelligence for the future radiology diagnostic service, Front. Mol. Biosci. 7 (2021) 614258, https://doi.org/10.3389/fmolb.2020.614258.
[34] K.G. van Leeuwen, S. Schalekamp, M.J. Rutten, B. van Ginneken, M. de Rooij, Artificial intelligence in radiology: 100 commercially available products and their scientific evidence, Eur. Radiol. 31 (6) (2021) 3797–3804, https://doi.org/10.1007/s00330-021-07892-z.
[35] K. Freeman, J. Geppert, C. Stinton, D. Todkill, S. Johnson, A. Clarke, S. Taylor-Phillips, Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy, BMJ 374 (2021) n1872, https://doi.org/10.1136/bmj.n1872.
[36] R.L. Pierce, W. Van Biesen, D. Van Cauwenberge, J. Decruyenaere, S. Sterckx, Explainability in medicine in an era of AI-based clinical decision support systems, Front. Genet. 13 (2022) 903600, https://doi.org/10.3389/fgene.2022.903600.
[37] A.S. Albahri, A.M. Duhaim, M.A. Fadhel, A. Alnoor, N.S. Baqer, L. Alzubaidi, O. S. Albahri, A.H. Alamoodi, J. Bai, A. Salhi, J. Santamaría, C. Ouyang, A. Gupta, Y. Gu, M. Deveci, A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion, Inf. Fusion 96 (2023) 156–191, https://doi.org/10.1016/j.inffus.2023.03.008.
[38] Kannelønning, M. S., Grisot, M., & Williams, R. (forthcoming). Towards experimental implementations: moving emerging AI technologies into real-world clinical settings. In P. Giardullo & F. Miele (Eds.), Algorithmic Care: STS Perspectives on Automation of Care. Palgrave Macmillan..
[39] Y. Kumar, A. Koul, R. Singla, M.F. Ijaz, Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda, J. Ambient Intell. Hum. Comput. (2022) 1–28, https://doi.org/10.1007/s12652-021-03612-z.
[40] S. Gilbert, S. Anderson, M. Daumer, P. Li, T. Melvin, R. Williams, Learning from experience and finding the right balance in the governance of artificial intelligence and digital health technologies, J. Med. Internet Res. 25 (2023) e43682, https://doi.org/10.2196/43682.
[41] G. Ellingsen, L. Silsand, G.-H. Severinsen, L.H. Linstad, Scaling AI projects for radiology–causes and consequences, in: B. Séroussi, P. Weber, F. Dhombres, C. Grouin, J.-D. Liebe, S. Pelayo, A. Pinna, B. Rance, L. Sacchi, A. Ugon, A. Benis, P. Gallos (Eds.), *Challenges of Trustable AI and Added-Value on Health* MIE, 2022. Nice.
[42] T.C.O. Hashiguchi, J. Oderkirk, L. Slawomirski, Fulfilling the promise of artificial intelligence in the health sector: let's get real, Value Health 25 (3) (2022) 368–373, https://doi.org/10.1016/j.jval.2021.11.1369.
[43] R. Savjani, P. Singh, How to successfully build and run AI competitions for medical imaging: insights from the PANDA challenge, Radiology: Imaging Cancer 4 (3) (2022) e229010, https://doi.org/10.1148/rycan.229010.
[44] D. Higgins, V.I. Madai, From bit to bedside: a practical framework for artificial intelligence product development in healthcare, Advanced intelligent systems 2 (10) (2020) 2000052, https://doi.org/10.1002/aisy.202000052.
[45] F. Gilbert, S. Smye, C.-B. Schönlieb, Artificial intelligence in clinical imaging: a health system approach, Clin. Radiol. 75 (1) (2020) 3–6, https://doi.org/10.1016/j.crad.2019.09.122.
[46] K.G. van Leeuwen, M. de Rooij, S. Schalekamp, B. van Ginneken, M.J. Rutten, How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatr. Radiol. 1–7 (2021) https://doi.org/10.1007/s00247-021-05114-8.
[47] R.J. Chen, J.J. Wang, D.F.K. Williamson, T.Y. Chen, J. Lipkova, M.Y. Lu, S. Sahai, F. Mahmood, Algorithmic fairness in artificial intelligence for medicine and healthcare, Nat. Biomed. Eng. 7 (6) (2023) 719–742, https://doi.org/10.1038/s41551-023-01056-8.

[48] I. Kulkov, Next-generation business models for artificial intelligence start-ups in the healthcare industry, Int. J. Entrepreneurial Behav. Res. 29 (4) (2023) 860–885, https://doi.org/10.1108/IJEBR-04-2021-0304.
[49] H. Alami, P. Lehoux, Y. Auclair, M.d. Guise, M.-P. Gagnon, J. Shaw, D. Roy, R. Fleet, M.A.A. Ahmed, J.-P. Fortin, M. de Guise, M.A. Ag Ahmed, Artificial intelligence and health technology assessment: anticipating a new level of complexity, J. Med. Internet Res. 22 (7) (2020) 1–12, https://doi.org/10.2196/17707.
[50] A. Kumar, B. Finley, T. Braud, S. Tarkoma, P. Hui, Sketching an ai marketplace: tech, economic, and regulatory aspects, IEEE Access 9 (2021) 13761–13774, https://doi.org/10.1109/ACCESS.2021.3050929.
[51] S. Kapoor, A. Narayanan, Leakage and the Reproducibility Crisis in ML-Based Science, 2022 *arXiv preprint arXiv:2207.07048*.
[52] N. Mehta, A. Pandit, S. Shukla, Transforming healthcare with big data analytics and artificial intelligence: a systematic mapping study, J. Biomed. Inf. 100 (2019) 103311, https://doi.org/10.1016/j.jbi.2019.103311.
[53] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and Vascular Neurology 2 (4) (2017) 230, https://doi.org/10.1136/svn-2017-000101.
[54] L.P. Golding, G.N. Nicola, A business case for artificial intelligence tools: the currency of improved quality and reduced cost, J. Am. Coll. Radiol. 16 (9) (2019) 1357–1361, https://doi.org/10.1016/j.jacr.2019.05.004.
[55] R. Baumgartner, et al., Fair and Equitable AI in Biomedical Research and Healthcare: Social Science Perspectives, Artificial Intelligence in Medicine, 2023, https://doi.org/10.1016/j.artmed.2023.102658.
[56] D. Neyland, Bearing accountable witness to the ethical algorithmic system, Sci. Technol. Hum. Val. 41 (1) (2016) 50–76, https://doi.org/10.1177/0162243915598056.
[57] E. Jussupow, K. Spohrer, A. Heinzl, Radiologists' usage of diagnostic AI systems, Business & Information Systems Engineering 64 (3) (2022) 293–309, https://doi.org/10.1007/s12599-022-00750-2.
[58] F.M. Calisto, N. Nunes, J.C. Nascimento, Modeling adoption of intelligent agents in medical imaging, Int. J. Hum. Comput. Stud. 168 (2022) 102922, https://doi.org/10.1016/j.ijhcs.2022.102922.
[59] T.O. Andersen, F. Nunes, L. Wilcox, E. Coiera, Y. Rogers, Introduction to the special issue on human-centred AI in healthcare, Challenges Appearing in the Wild 30 (2) (2023) 1–11, https://doi.org/10.1145/3589961.
[60] R. Williams, N. Pollock, Moving beyond the single site implementation study: how (and why) we should study the biography of packaged enterprise solutions, Inf. Syst. Res. 23 (1) (2012) 1–21. https://www.jstor.org/stable/23207869.
[61] M.-T. Ho, N.-T.B. Le, P. Mantello, M.-T. Ho, N. Ghotbi, Understanding the acceptance of emotional artificial intelligence in Japanese healthcare system: a cross-sectional survey of clinic visitors' attitude, Technol. Soc. 72 (2023) 102166, https://doi.org/10.1016/j.techsoc.2022.102166.
[62] S. Hyysalo, N. Pollock, R. Williams, Method matters in the social study of technology: investigating the biographies of artifacts and practices, Sci. Technol. Stud. 32 (3) (2019) 2–25, https://doi.org/10.23987/sts.65532.
[63] K. Cresswell, M. Rigby, F. Magrabi, P. Scott, J. Brender, C.K. Craven, Z.S.-Y. Wong, P. Kukhareva, E. Ammenwerth, A. Georgiou, S. Medlock, N.F. De Keizer, P. Nykänen, M. Prgomet, R. Williams, The need to strengthen the evaluation of the impact of Artificial Intelligence-based decision support systems on healthcare provision, Health Pol. 136 (2023) 104889, https://doi.org/10.1016/j.healthpol.2023.104889.