'If You're Going to Trust the Machine, Then That Trust Has Got to be Based on Something': Validation and the Co-Constitution of Trust in Developing Artificial Intelligence (AI) for the Early Diagnosis of Pulmonary Hypertension (PH)

Peter Winter

School of Sociology, Politics and International Studies (SPAIS), University of Bristol, United Kingdom / peter.winter@bristol.ac.uk

Annamaria Carusi

Interchange Research; and Department of Science and Technology Studies, University College London (UCL), United Kingdom.

Abstract

The role of Artificial Intelligence (AI) in clinical decision-making raises issues of trust. One issue concerns the conditions of trusting the AI which tends to be based on validation. However, little attention has been given to how validation is formed, how comparisons come to be accepted, and how AI algorithms are trusted in decision-making. Drawing on interviews with collaborative researchers developing three AI technologies for the early diagnosis of pulmonary hypertension (PH), we show how validation of the AI is jointly produced so that trust in the algorithm is built up through the negotiation of criteria and terms of comparison during interactions. These processes build up interpretability and interrogation, and co-constitute trust in the technology. As they do so, it becomes difficult to sustain a strict distinction between artificial and human/social intelligence.

Keywords: Artificial Intelligence, technology development, early diagnosis, trust, collaboration, validation

Introduction

In this article, we consider the central question of trust in Artificial Intelligence (AI) technologies for medical diagnosis. As AI becomes increasingly integrated into existing workflows and implemented to support diagnosis and treatment, clinical experts will find it difficult to understand how AI algorithms have been validated: this is where the problem of trust arises (Scheek et al., 2021). For many clinical and technical experts (such as computer and data scientists), trust is a matter of explainability and transparency of the algorithm, or the justification of the outputs of an algorithmic model (Tonekaboni et al., 2019; Barda, 2019; Cutillo et al., 2020). One way to broach these issues of trust is through the development of guidance that aims to foster responsible and trustworthy applications of AI (Bærøe, 2020). Examples include AI4People (Floridi et al., 2018), Asilomar AI principles and the Independent High Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission (2019). Altogether, guidance and initiatives associated with developing trustworthy AI have in common ethical frameworks (principles and guidelines) to improve morally good outcomes. In particular, the AI HLEG argue that AI should be designed and developed in ways that build in interpretability from the start through assessment lists – a work process which assumes that trust can be accomplished through a rigorous application of pre-identified evaluation criteria.

Yet, despite these efforts, levels of acceptance of healthcare AI remain low: several studies have come to the conclusion that there is a lack of trust among clinical experts towards these kinds of technologies, which as a consequence, has led to low acceptance and use (Topol, 2019; Strohm, 2019; Cabitza et al., 2020; Sreedharan et al., 2020; Nagendran et al., 2020). Topol (2019) shows that the lack of data and proof is eminently to blame – indeed, he argues that there is a lack of research investigating the validation and readiness of Machine Learning (ML) models in clinical settings, prompting distrust on the assumptions underpinning many validation tests that have been assessed in the laboratory. Taking this idea of the validity of ML models one step further in the context of AI development in biology and medicine, Littmann et al., (2020) states that it is collaboration itself which leads to AI research that is more scientifically valid, in that it is more correct and reproducible. One could similarly compare such thought on collaboration with the work of Elish and Watkins (2020: 6) in Science and Technology Studies (STS) who take stock of the 'sociotechnical' engagements between different human experts and their algorithms and the work of building trust in new technology. We claim that an important source of trust is the collaboration between AI developers and clinical experts, and we aim to show how forms of collaboration support the collective construction of validation and interpretability, which ultimately grounds trust in the technology.

This article aims to give a detailed account of how collaboration informs the co-emergence of trust and validation in a setting where three AI algorithms are being developed for use in real-world clinical settings. In addition, we show how validation that looks towards real-world settings is not something that occurs at the endpoint of the development process. Instead, it occurs *throughout* the development process and is built into the application. This occurs primarily through collaboration with clinical experts from the initial stages of development and concrete practices of repurposing healthcare data. While validation may often be viewed as something that comes at the endpoint of algorithm development, the grounds upon which validation is based starts at the outset of collaboration and continues through the development process across contexts, practices and technologies. This approach is particularly relevant to our discussion on the practical efforts and meaningful selection of criteria for comparison in AI development. For example, as will be explored later, clinical experts who participate in the practice of selecting, testing, and refining criteria (e.g. labels, codes, or variables) for comparison are the ones who are able to interrogate the outputs of validation, whereas a clinician who has not been involved in that process may not comprehend or interpret the outputs in the same way and may open up the potential for "blind trust" in the technology (Gaube et al., 2021: 1).

The subject of trust has a wider relevance for social scientists interested in collaboration and development of new (AI) technologies, and will provide critical insight in an area imbued with high claims, promise and technological expectations (see Rajpurkar et al., 2017, 2018; Perry, 2017; Ming, 2018). This article will draw from the overlapping fields of STS and Computer Supported Cooperative Work (CSCW) on collaboration in the context of technology development, to treat collaboration as a set of work practices that are invoked at particular times for building trust towards algorithms. The first part of the article considers the relationship between trust and collaboration in general. The next part of the article deals with the basic notions related to validation in the technical and social science literatures. The focus then shifts to the background of the study and its associated methods are described. The central section of the article opens by showing how the first three dimensions of AI development (*Querying Datasets*, *Building the Software,* and *Training the Model)* play their role in the validation process. In the following sections, we take this further to consider the different kinds of collaborative practices for building trust in the validation process, specifically reflecting on the practical efforts and meaningful selection of criteria for comparison. After this salient presentation of data, we move on to discuss how validation is a collaborative endeavour, foregrounding our position that validation starts at the outset of collaboration and continues through the development process across contexts, practices and technologies. The final section of the article concludes with the notion that AI requires constant monitoring and refinement which are a far cry from providing a 'technological fix' for problems in society and healthcare in particular.

Trust and collaboration

The topic of trust has received a great deal of attention in research into how technologies are deployed to support tasks and decisions. How to trust the outputs of technologies is particularly acute when their development and use crosses across different expertises and disciplines. In these contexts, trust emerges through particular collaborative tasks between people with different expertise, as seen in multidisciplinary teams (MDTs) who jointly make diagnosis or treatment decisions (Van Baalen et al., 2017; Van Baalen and Carusi, 2019). A similar line of thought is followed by Elish (2018: 369) who argues that trust in AI technologies can also be built by including or "looping" in stakeholders (such as clinicians) from the very beginning and throughout the development process. Such collaborations are mediated by a 'local champion', a clinical expert involved in the development of the technology who does "vital trust-building work" throughout the hospital and the wider clinical community (Strohm, 2019: 58). The field of CSCW has developed a substantial and highly relevant body of work that explores trust in various contexts, and frequently focuses on the role of trust as a key aspect of collaboration between people, but also in relation to processes and technologies which directly impact how expert judgements are made (Fitzpatrick and Ellingsen, 2013). Here, trust is commonly conceptualised as linked to features of interpersonal relationships between people and often remains implicit with familiarity/lack of familiarity being a basis for trust/mistrust in human-human interactions (e.g., Jirotka et al., 2005; Carusi, 2009). Trust may also be conceptualised as generated 'in action', built up in some form of situated or contextual practical engagement of a work routine, often in contexts when people have a responsibility to build trust in new technology (e.g., Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Kuutti and Bannon, 2014; Papangelis et al., 2019).

When interpersonal and practical trust-building becomes a mediator for the development of new technologies (i.e., algorithms), people become deeply embedded in technical and non-technical processes, and other temporalities. Here, technology development is characterised as a complex and active form of sociotechnical production with experts being influenced by a variety of parameters, pressures, and politics that make up the social construction of complex technologies (Mackenzie, 1990; Laurent and Thoreau, 2019). Mackenzie (1990), in particular, demonstrated how the accuracy of a technology can be constructed and shaped by both technical engagement and the perspectives of social actors involved in its process of development. In contexts of collaboration, these interactions may be seen as the often 'invisible work' that goes into technology development - although the people who perform such interactions are quite visible, the work they do is relegated to the background (Star and Strauss, 1999: 20). According to Star and Strauss (1999: 10), one important form of work which is often invisible in making technologies work is the concept of 'articulation work' – a type of work that happens after breakdowns or unanticipated contingencies as it is "work that gets things 'back on track' in the face of the unexpected". Pallesen and Jacobsen (2018: 173) suggest articulation work can also be understood as the work of coordinating between different sites of the experiment in collaborative research, in addition to being a salient concept for situated problem-solving. In other words, experts can bridge social worlds and thus 'mesh together' these very different social worlds to get work 'done'. Taking this approach, articulation work could also be seen (and needed) to support a type of 'sociotechnical infrastructure' that scaffolds medical and organisational work (Star, 1999). Star and Strauss' (1999) notion of 'invisible work' has also started to become an important analytical tool for understanding data work in healthcare (Bonde et al., 2019; Bossen et al., 2019; Bossen and Piras, 2020). In this context, invisibility may refer to the invisible nature of collaborative work performed by actors around practices of data; a process which plays a key role in ensuring the truthfulness and correctness of data to support clinical practice (Bjørnstad and Ellingsen, 2019).

Together, we might see these as two kinds of trust that complement each other: the interpersonal trust experts acquire when interacting with experts from different disciplinary backgrounds on the one side, and the practice-orientated trust experts acquire when they participate in developing the tool, technology or instruments. We seek to convey the idea that both types of trust work are forms of invisible work because they too often remain implicit and hidden in scientific accounts of validation. Taking this into account, the concept helps us to identify and surface the invisible work of trust, as well as also to become attentive to, how the mundane work of collaborative research and data practices are generative of validation.

Our research suggests that trust in healthcare AI is co-constituted by collaborators from throughout the development process, and that this underpins validation. This point about AI and the fact that trust, validation and the technical characteristics themselves are co-constructed is significant in a broader debate where AI tends to be seen as a 'technological fix' able to solve multiples issues, including the problem of trust in the ability of institutions to solve complex problems. According to Katzenbach (2019), AI is accepted in particular areas, like healthcare, transport, and social media, as a kind of technological fix for solving specific problems. For example, Katzenbach (2019) recognises that autonomous vehicles can help reduce traffic accidents, and sees the potential of using AI for detecting misinformation and hate speech online. Specifically, however, he argues that this talk about 'AI fixing things' is misleading because it obfuscates the importance of human labours and social relations that these technologies are built upon. For this reason, the objective of this article brings to light not only the technology's inherent technical properties, but also the role social processes such as collaboration play in the construction of trust in AI development.

With this article, we want to bring trust and validation together: we propose that collaboration plays a part in the generation and maintenance of trust relationships (between people and technologies) which directly impact how expert judgements are made and accepted. In the next section we suggest that the focus on validation as a technical solution to trust has left underappreciated the collaborative, social aspects of the validation process. These are the focus of the social science literature on validation, which proposes that validation is as much about people's social interactions with technology and each other as it is about any technical feature of the technology. As we will later show, the process of selecting and negotiating the criteria that go into evaluating the technologies, and considering it 'validated' are useful for building in trust in judgements made about the technology and its outputs.

Validation

Technical literature

In the technical literature, algorithms are required to pass some form of quality control in the form of a validating test (or set of tests or criteria) in the demand for trusted or trustworthy systems (Alpaydin, 2016; Tonekaboni et al., 2019; Barda, 2019; Cabitza et al., 2020). These tests are often based on a comparative performance of the technology, comparing its performance with other performances considered to be a 'gold standard', such as a human expert producing confirmed findings in a diagnostic report (e.g., Gulshan et al., 2016; Esteva et al., 2017; Rajpurkar et al., 2018; Annarumma et al., 2019). Accordingly, the performance of the algorithm against the gold standard is often expressed in statistical terms (e.g., 'accuracy', 'sensitivity', 'specificity') and by some kind of expert who is able to make a judgement about its performance, such as having high predictive accuracy, for example Rajpurkar et al's (2017, 2018) CheXNeXt algorithm being able to make accurate predictions at a level that "exceeds the average radiologist performance" (Rajpurkar et al., 2017: 2). However, such claims can prompt considerable scepticism and distrust across scientific and medical communities, as was the case with radiologist Oakden-Rayner (2017, 2018) who initiated a critique of the CheXNeXt model (along with the help of Rajpurkar and his team) to verify the accuracy of its predictions. The conclusion of that critique was that the images had not been labelled correctly, nor did the labels reflect clinical practice having the potential to produce meaningless predictions (Oakden-Rayner, 2018).

Oakden-Rayner's critique contains important epistemological questions that deserve consideration: questions about how comparisons can be made (especially between algorithms and human experts), and how data is labelled (who labels the data, who inspects the data and whether experts with relevant clinical experience are considered). Labels and codes or criteria for comparing performances come to matter greatly when it comes to validation because they are based on the so-called 'ground truth' of features that the algorithm has learned in the training data – the labels, annotations, or codes in this instance constitute the ground truth or ground for comparison (e.g., Gulshan et al., 2016; Esteva et al., 2017; Oakden-Rayner, 2018; Cabitza et al., 2020; Scheek et al. 2021).

In addition, such validation is commonly represented as consisting of two isolated approaches: internal and external (Topol, 2019; Cabitza and Zeitoun, 2019; Nagendran et al., 2020). The focus of most healthcare AI development is a form of internal validation, carried out within computer science laboratories and tested on retrospective datasets. External validation is usually referred to as the clinical validation of AI systems and tested on prospective datasets of entirely new data ('in the wild') (Cabitza and Zeitoun, 2019). As Nagendran et al., (2020) point out, most validation studies are tested on retrospective datasets only, with the number of prospective datasets tested in real-world clinical settings extremely low (only 6 out of 81). Cabitza and Zeitoun (2019: 161) also distinguish between 'statistical', 'relational', 'pragmatic' and 'ecological' validity. Statistical validity is claimed by them to be objective, 'intrinsic' and 'essential' to the system. However, relational, pragmatic and ecological validity consider the context of the algorithm in one or other way. For instance, either with respect to usability or pragmatic consequences (for example, how data is handled), or with 'ecological' consequences, (for example, with respect to work settings). Nonetheless, however technical these different forms of validation may seem to social scientists, they are important concepts in understanding how experts consider validation as a technical practice and something that comes at the endpoint of technology development.

Social science literature

Social science literature on model validation provides us with the capacity to investigate validation practices and trust practices *in the making*. This literature on validation in science has provided us with a sustained analysis of the confusions and uncertainties that accompany validation (Randall and Wielicki, 1997; Shackley et al., 1998; Küppers and Lenhard, 2005; Sundberg, 2006; Winsberg, 2010; Morrison, 2015). Science policy scholars have produced in-depth analyses of the validation of chemical or environmental models, showing the extent of uncertainties and disagreement on the model's validity, relevance and bias (Oreskes et al., 1994; Oreskes, 2004; White et al., 2010). A major reason for this would seem to lie in the nature of how evidence is subjected to different standards of 'proof' and different ways of thinking about proof in different sectors – a far cry from the supposed 'objectivity' of models or the quantitative nature of empirical data (Oreskes, 2004). For an STS view on this matter, see Mackenzie's (2001) work on *Mechanizing Proof* and how experts negotiate data to be worked with and construct 'proof' of the correctness of a program or software design. 'Proof' that the model or software is in absolute sense 'correct' or 'dependable' is very much a social process of iteration (e.g., doing testing, returning to the nature and use of data, redefining the test, repeating the test, finding the design fault, and so on) (Mackenzie, 2001: 43). At some point in this cycle, experts come to an understanding that their software is often reasonably reliable because of how humans interact with the technology and by testimony to a 'trustworthy agent' to whom they may turn (Mackenzie, 2001: 307). Other STS literature has also made the same points about the validation of models while exploring different factors affecting scientists' reasoning and choices (Sundberg, 2006; Böschen, 2009, 2013; Carusi et al., 2012; Carusi, 2014, 2016; Thoreau, 2016; Boullier et al., 2019; Laurent and Thoreau, 2019). Böschen (2009, 2013), in particular, has distinguished between what he calls four 'evidential cultures' and two of these are most relevant in this context. First that a 'restrictive evidential culture' rests primarily on experimental methods in controlled laboratory settings using models to establish causality, but often orient scientists to particular drawbacks of the phenomenon being tested (e.g., having limited available data on which to test the comparability of results). Second, that a 'holistic evidential culture' may be combined with other tests and different forms of knowledge to evaluate the phenomena. This time there is less interest in capturing causal phenomena and more of a move towards capturing complex elements of an ecosystem or the larger system of people's lives and cosmologies. This holistic culture chimes with the notions of pragmatic validity and ecological validity of other studies discussing validation (Cabitza and Zeitoun, 2019).

However, the most important contributions of STS researchers in the analysis of validation for this article derives from research on the implementation of an AI algorithm for the early detection of sepsis ('Sepsis Watch') (Elish, 2018; Elish and Watkins, 2020; Sendak et al., 2020). Concerned with the validation of Sepsis Watch, these authors present validation as an integral component for establishing the trust of clinicians and point out how existing epidemiological or 'gold standard' definitions of sepsis were found to be inadequate at predicting the risk of sepsis in real-time cases in the clinical setting (Elish and Watkins, 2020: 18). What they found in the clinical setting was a negotiation and refinement of criteria and variables where trust had been manifested in the process. Trust of the sepsis algorithm was by no means dependant on some technical neutrality of the model, but a series of key activities that brought clinicians and statisticians together, promoting a potent combination of empirical observation, refinement and repair. The emphasis on real-time validation and the ongoing collaborative work of clinicians and statisticians shows that the algorithm came to be trusted through technical demonstrations of efficacy rooted within social relationships.

The central argument of these articles is that validation is as much about people's social interactions with technology as it is about any technical feature of the technology; it is inextricably socio-technical. The technology is not an inert thing passively being acted upon until it reaches a point where it is deemed 'validated'. Rather, it actively mediates interactions and fosters interpersonal trust and practice-orientated work, and through these, the creation of scientific knowledge and technical results, such as its accuracy (Mackenzie, 1990) or proof (Mackenzie, 2001; Laurent and Thoreau, 2019). The criteria according to which validation will be assessed are not pre-defined, but emerge during the process (Carusi, 2014). This makes for a technology that is more likely to be accepted by potential users, and actually embedded in their real-world context.

Taken together, these studies recognise the importance of validation on clinical experts' trust of models. However, there is still much work to be done in investigating the process of validation. This is especially the case when validation is associated with AI models in healthcare, which iteratively involves contesting and selecting criteria or classifications for comparison. This article does just this by paying deeper attention to the voices involved in the process of validation and making explicit the conditions under which their reasoning operates. It extends the previous STS literature by showing how the collaborations that give rise to AI co-produce the criteria that act as grounds for comparison which underlie validation practices.

Our study: AI in the clinic

Our study explored the development process of three AI technologies for the early diagnosis of pulmonary hypertension (PH). PH is a rare, progressive and life shortening lung disease that is often diagnosed at an advanced stage. Diagnosis for PH is assisted by a myriad of testing technologies (such as right heart catheterisation, blood tests and medical imaging). However, such technologies are often deployed too late in the disease process, and therefore may yield a late diagnosis with limited treatment outcomes or poor markers for prognosis (Kiely et al., 2013). Because of this problem of late diagnosis, clinicians and researchers around the world are looking to AI as a route to an earlier diagnosis for PH in order to bring about better life expectancy and quality of life for patients (e.g., Kiely et al., 2019; Swift et al., 2020)

The first AI being developed is a 'screening' algorithm to detect patients 'at risk' of PH trained on Hospital Episode Statistics (HES) data. The second algorithm being developed is an 'imaging' algorithm which uses Magnetic Resonance Images (MRI) of the heart in order to detect disease features of PH. The third algorithm being developed is the 'biomarker' algorithm to detect signs or signatures of PH in blood samples trained on biomarker data related to PH, to be included in the screening algorithm. At the time of our study, all three algorithms were at the proof-of-concept stage with the intention of being deployed and used in the context of a UK PH Referral Centre at a major NHS Teaching Hospital. Thus, we are studying three proof-of-concept projects in the early development phase, highlighting the invisible work of the sociotechnical infrastructure (Star, 1999), ideally for organising, supporting, and elevating the next steps of each project in order to facilitate their route into clinical trials.

Methods

This article is based on qualitative interviews with seven participants involved in developing three proof-of-concept AI algorithms for the early diagnosis of PH. Participants included: two PH clinicians, one consultant PH nurse, one radiologist, one computer scientist, one data scientist, and one biomedical scientist to fully take into account the sea of discourses, ideas, scientific criteria, and concepts that shape validation and trust in AI development. In total there were six face-to-face interviews in workplace offices. One of these interviews was a joint interview conducted with the computer scientist and radiologist both working together to develop the imaging algorithm. Data was collected between 17/05/2019 - 22/10/2019. Recordings were transcribed and uploaded to NVIVO 12 to help manage, code and analyse themes that emerged from the transcripts. Taking an inductive approach to thematic analysis (Braun and Clarke, 2006), the theme of validation explicitly emerged across the group of research participants with decisions involving validation understood to be inherently tied to trust: interpersonal interactions and various computer supported practices involved in validation demanded consideration.

Our fieldwork was conducted on three developing algorithms (mentioned above). These algorithms were small scale pilot projects or proof-of-concept projects being developed to show the viability of AI to tackle challenges of early diagnosis, projecting hopes of a 'technological fix' (Katzenbach, 2019). As such, the projects involved small numbers of people, and often just two or three people were the main developers and sometimes one person would be working on two, or even all three of the algorithms. Accordingly, our interview numbers are not high. This will affect the generalisability of our findings. We might say that the proof-of-concept nature of the projects we studied and our own study are limited in similar ways. Despite the relative intimacy of our research domain, our research produced some important insights concerning how these collaborations operated to establish trust and to set criteria for validating the performance of the AI applications.

Results

Laying the ground for validation: querying datasets, building software, and training the model

Validation is often represented as the final stage of technology development (Alpaydin, 2016). However, a significant amount of interpersonal and practice-orientated trust work, and a large proportion of training/testing activities occur earlier in the development process. These opportunities to build trust in the technology are crucial for technology development, but often remain 'invisible' and go unnoticed and unaccounted for, relegated to the background (Strauss and Star, 1999; Oudshoorn, 2008). Here, one needs to think of the previous three work activities (*Querying Datasets*, *Building Software*, and *Training the Model)* that take place before a formal validation phase, a perspective that shows how each activity lays the ground for validation. Whilst we have argued elsewhere how these three activities help to demystify the algorithm and 'de-trouble' transparency issues (Winter and Carusi, *forthcoming)*, we argue in this article how each activity can also be said to present interpersonal and pragmatic opportunities for building trust towards the final validation experiment. These activities lay the foundation for how trust and validation co-emerge in the sociotechnical infrastructures of diagnostic work through their negotiation and refinement of criteria and are explained in the following.

Querying datasets is concerned with how external or internal datasets are curated. It brings into play questions around how the datasets have been labelled or coded and by whom and whether they have sufficiently included clinical experts, which may lead to imprecise datasets and to inaccurate tests. As Oakden-Rayner (2017, 2018) reminds us, dataset quality is crucial in relation to the way in which criteria such as labels on medical images lay the ground for validation, namely how the labels are used to validate its performance. In our study, a radiologist developing the imaging algorithm echoed this concern by highlighting the difference in quality between datasets that have been collected prospectively and retrospectively:

> When evaluating very large cohorts with thousands of patients, people will question, unless it's a prospective study, 'how do they know that person actually had that condition?' And if it's from a clinical database, how was that really done? If all patients went through a multidisciplinary team meeting with recorded outcomes, that's very robust. But when data is collected retrospectively without an MDT diagnosis or similar assessment this can leave uncertainty as to the validity of the data.
> (Participant 4, Radiologist)

The quote expresses the radiologist's concern about the quality of prospective datasets and retrospective datasets. For the radiologist, if a label can be traced through to a prospective study in which the radiologist is either directly involved in the labelling of data, or is familiar with the experts who have participated in its labelling, the dataset is considered "very robust". However, if labels have come from a retrospective study where the labelling is not first-hand, the labelling process is less certain, because the radiologists are not sure of the processes used by the experts in applying the labels, asking for example, "how do they know that person actually had that condition?", and "how was that really done?". The radiologist's trust is anchored in previous interactions with expert members of the MDT and serves as the basis for the radiologist's perception of the quality of the dataset, and in this sense, is a form of interpersonal trust (Jirotka et al., 2005; Carusi, 2009; Van Baalen and Carusi, 2019).

Despite the lack of certainty regarding how labels were applied in a retrospective dataset, these datasets are used for technology development. Retrospective datasets provide the raw material for reconstructing and interpreting diagnoses, as seen in the quote below:

> Retrospective data labelling has its limitations and it's going to require us to go back into it and look at the scans and make a retrospective diagnosis on some cases because it comes from a number of different acquisition methods, different radiographers, and in the case of derived measurements different software [...] So coding is very, very important […] it needs work for people to go back in and classify patients retrospectively sometimes.
> (Participant 4, Radiologist)

Consequently, our focus on practice calls attention to the lengthy struggles clinical experts may face with research materials to reconstruct them in a way that facilitates their diagnosis, for example through labelling or coding key features of interest and aligning them with their own clinical experience and local work practice. This treatment of retrospective datasets demonstrates how practical work of querying and relabelling features on images is required for the radiologist to trust the dataset. CSCW scholars will recognise this as one type of data work that takes place to elucidate the emerging requirements for management and work system design (Bossen et al., 2019). Indeed, the complexity of 'repurposing' data to serve secondary purposes beyond the practices of its initial use (Bonde et al., 2019; Bossen and Piras, 2020) challenges our radiologist to work with conflicting qualities or ambiguities of data and the activities needed to ensure the 'correctness' of data (Bjørnstad and Ellingsen, 2019). Importantly for this article, the radiologist's reconstruction of diagnoses through negotiation and refinement of diagnostic criteria reminds us of how trust can actually be engendered in a practical situation (Clarke et al., 2006a; 2006b; Oudshoorn, 2008; Papangelis et al., 2019) and moreover, calls attention to their 'articulation work' (a form of invisible work) that "gets things back on track" when unanticipated situations arise (Star and Strauss, 1999: 10).

Building the software means the building of a classification software. It is an activity that continues to lay the ground for validation because it draws on the experience of clinical experts who participate in the negotiation or refinement of appropriate criteria (e.g., diagnostic labels/codes or other variables) for software building. As part of this process, clinical experts start learning how the software arrives at its classifications, how the software is assessed, and how to participate in future refinements of criteria.

Training the model takes the last activity further by inviting clinical experts to assess the training outputs of the algorithm in an imagined clinical context. Clinical experts are included in the critical assessment of the software's outputs and participate in discussions about whether outputs are relevant or plausible, using their clinical experience to change or refine existing criteria included as features of the model.

However, as we will see in the second half of the article, this process of establishing what could count as criteria for comparison is never static or fixed (Carusi, 2014). Rather, it continues throughout the whole of the development process and sets up the algorithm for a formal validation test. The next section continues to look at this process, particularly focusing on the method of internal validation and the collaborative work involving AI developers and clinical experts in setting up the criteria for comparison between the algorithm's results on different or unused datasets. Building on the previous discussion about the negotiation and refinement of labels in 'Laying the Ground for Validation', we investigate how criteria and variables under retrospective conditions have to be retemporalised for clinical contexts accordingly by bridging or 'meshing' the nexus between external validation and internal validation.

Different forms of validation

As we have previously discussed, there are two main steps to validation: testing against retrospective datasets and testing prospectively (Topol, 2019; Cabitza and Zeitoun, 2019; Nagendran et al., 2020). The focus of most AI development is on retrospective datasets, which is a form of internal validation, carried out within (mostly) computer science laboratories in universities or industries. External validation is the testing of the AI application against entirely new data, 'in the wild', as it is not the data in the same dataset as the algorithm was trained on. In our study, AI developers invited clinicians into the laboratory to assess the performance of the algorithm on the retrospective datasets: work that bridged the gap between internal and external validation and allowed both AI developers and clinical experts to gain an understanding of *how* validation was carried out. The involvement of the clinical experts in bridging the gap between internal and external validation shows how knowledge can be co-produced and how the knowledge from the laboratory needs to be related to the real-world (Boullier et al., 2019). The bridging between two different settings for validation purposes continued the process of establishing appropriate criteria for comparison, showing how criteria continue to be negotiated and refined in ongoing iterations of tests (Carusi, 2014), and in the process, how trust and validation co-emerge. This bridging process begins with the clinical expert's first encounter with the results of the first internal validation test.

Internal Validation

Here we join the computer scientist and radiologist in an interview about their method of internal validation for the imaging algorithm in the laboratory. Internal validation here involves the computer scientist and radiologist pursuing the goal of setting up a meaningful comparison between the algorithms results on unused segments of the imaging dataset, and then later on refining the criteria for comparison between the algorithm and the radiologist**.** When asked how they went about validation for the imaging algorithm, the computer scientist replied:

> We use cross-validation. Basically, we partition the data set into ten parts, ten partitions, then we use nine of them for training and one for testing and then just rotate. So that is one of the classical methods in machine learning to validate a method when we have limited number of samples in a dataset. I think in the beginning it really gives us quite a good estimation of how much the algorithm can achieve compared to the current approach of manual segmentation.
> (Participant 5, Computer Scientist)

In their approach to validate the imaging algorithm, the computer scientist states that they are using the method of "cross-validation". The computer scientist explains how this specific validation process is informed and dominated by the separation of datasets into nine training sets ("we use nine of them for training") and one testing set ("one for testing") which are then rotated ("then just rotate"). The comparison that this approach relies on is with "manual segmentation". That is, it is with the diagnostic labels that have already been applied to the dataset and queried by clinical experts (as described above). When asked about how they arrived at this judgement of how good the validation was and who was involved, the computer scientist highlights the important part the labels play in providing 'ground truths':

> The data are all labelled with ground truths […] When we try to predict the label of the individual patient on that test set, we're doing the prediction pretending the label is not available. Then we use the ground truth labels to compare the predicted label and then we compute an error, so if this error is small then that's high accuracy.
> (Participant 5, Computer Scientist)

First, this quote shows how each label on a dataset of medical images constitutes a 'ground truth' – a process established earlier in the article by the radiologist's querying of datasets (e.g., the relabelling of features). Second, the performance of the imaging algorithm in arriving at the 'correct' detection of PH-related features is compared with the clinical labels embedded in the dataset. On the basis of this comparison, the size of the error between the computer's performance and the labelled dataset is computed. This becomes the metric of how well the algorithm performs ("so if this error is small then that's high accuracy"). This establishes the statistical validity of the algorithm (Cabitza and Zeitoun, 2019). Importantly for this article, this excerpt from the interview highlights how the objective of AI development is to build models that are accurate *enough* and highlights how accuracy is negotiated (Mackenzie, 1990) which for Laurent and Thoreau (2019: 165) is 'part and parcel' of technology development. Moreover, the picture of what can be deemed equivalent to what becomes clear in practice (Carusi, 2016): labels/codes become essential criteria and underpin judgements about the accuracy of validation tests (Scheek et al., 2021). Importantly, for this article, internal validation tests provide further opportunities to mediate practice-orientated trust between collaborators (Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Kuutti and Bannon, 2014; Papangelis et al., 2019). The next section will illustrate how this trust building deepens, paying particular attention to how clinical experts generate meaning with respect to the labels/codes or variables in the model – a process which is particularly useful when it comes to the 'interpretability' of outputs and continues the bridging between internal and external validation.

Interpretability

In the previous sections, we showed how clinical experts query the quality of datasets. We argued how clinical experts play a crucial role in establishing the quality of its curation: this helps them better understand the criteria that they are dealing with (e.g., labels/codes), builds practice-orientated trust work, and lays the ground for validation tests (e.g., cross validation). We also argued in the previous section that clinical experts bridge between internal and external validation. The next section will illustrate in detail the action of this bridging, highlighting the role clinical experts play in interpreting the performance of the imaging and screening algorithms in the validation tests. In fact, this is a process which shows how internal validation is inscribed with a view from clinical experience, however implicit that view might be.

Designing an AI system with interpretability in mind from the start opens up opportunities for not only practical interpretation and interrogation (and questions around what the output is, or how the output is made to matter in different situations), but also for building trust. The quote below from the computer scientist developing the imaging algorithm highlights why this practical interpretation matters:

> I actually have an end user over there to ask me questions […] like Participant 4 to give some suggestions on how to visualise the features and so on. I think that's something fresh to me and that also inspires me to write like interpretable machine learning […]. I think those kinds of challenges are real only when you start to interact with the community. So only when I interact with a domain expert, with an end user, then the question will come in.
> (Participant 5, Computer Scientist)

The computer scientist highlights how their collaboration with the radiologist brings in a variety of benefits: 1) questions that the computer scientist may not have thought of; 2) interpretability, that is, a kind of translation between the performance of the algorithm and the context of the domain expert; and 3) reality in terms of the uses to which it could be put in the radiologist's world. Working with the radiologist is a chance for considering the outputs of the algorithm in a clinical context and thereby highlights the radiologist's potential for bridging between internal and external validation – thus continuing to highlight the articulation work of clinical experts who 'mesh together' otherwise divided tasks, users and different systems (i.e., internal vs. external) and remains invisible because of its implicit nature (Star and Strauss, 1999; Pallesen and Jacobsen, 2018: 173). Nevertheless, interpreting algorithmic outputs is essential for the ongoing validation of the software, as iterative querying and questioning by clinical experts anchors the performance of the algorithm to their real-world context of use. In turn, this connection between meaning and use lays the ground for comparison for validation tests and engenders trust.

We observed similar processes of establishing the interpretation of algorithm outcomes for real-world contexts through querying and interrogation in the development of the screening algorithm. The main collaboration here was between a bioinformatics company and clinician. Again, we see the importance of the algorithm's outcomes being something 'real' in the clinicians' world ("From Participant 1's point of view, they're like 'this is something that I can relate to, I can relate to that number": Participant 2, Data Scientist). According to the data scientist in this case, the process of selecting the most appropriate ICD-10 codes was for "making sure that your comparative group are somehow relevant", which "is really important" and that without this clinical insight into how patients are diagnosed in the real-world clinic means that "the model at the end is just so trivial". Together, the consequences of this interpretation work in the development of the imaging and screening algorithms iteratively feed into the *Training of the Model*. This is because the results of any validation test feeds into further refinement of the model of the domain enacted in the algorithm - a further example of the ongoing 'articulation work' of the software (Star and Strauss, 1999). It is a process which occurs in the ongoing cycle of iterations for testing models (Carusi, 2014) and an integral aspect of all software development (Mackenzie, 2001). It also highlights the role of clinical experts engaging in linking or 'meshing' otherwise divided social worlds of the laboratory and the clinic, and how an understanding of criteria (labels/codes/variables) are negotiated within these laboratory settings with a view to their clinical application. Again, this practice-centred approach adds to the formation of a context of trust where the broader context is taken into account (Kuutti and Bannon, 2014).

Trusting questions

Trust, as we have seen in the sections above, is threaded implicitly throughout the whole of the development process and consists of a set of interpersonal interactions (Jirotka et al., 2005; Carusi, 2009; Van Baalen and Carusi, 2019) and practical engagement of the technology in question (Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Kuutti and Bannon, 2014; Pallesen and Jacobsen, 2018; Papangelis et al., 2019). However, trust is spoken about explicitly when it comes to some final validation test or method. From our interviews, clinical experts considered trust and validation as closely associated. Clinicians, in particular, considered validation a proxy for trust and to be on the terms of those whose trust is required for acceptance:

> Do you trust the information that you've been given and how much validation do you require? And I think that's the important thing. That element of trust. […] If you're going to trust the machine, then that trust has got to be based on something. So, it can be blind faith. So maybe some people are fairly evangelical about things, you've got blind faith that actually that machine is really good, so I'll just go with that.
> (Participant 1, Clinician)

As the quotation from the clinician reveals, trust is evidently directed at validation. Ultimately what it means if something is 'validated' is that it is trustworthy. For this clinician, validation is open-ended, since they are aware that the demands of validation could vary ("how much validation do you require?"). However, it is still possible to distinguish between requiring some form of validation and "blind faith", which they also associate with being "evangelical" about machine learning. The clinician then goes on to talk about an attitude of curiosity which comes into play in understanding what is meant by validation:

> A lot of people have got a certain degree of curiosity about 'do I really believe that?', 'is that really true?' and it's like that when you see a patient. You can take everything a patient tells you at face value or you can try and interrogate that information to see whether or not it's right. And you need to recognise sometimes that you are not very good at extracting information. Sometimes the patient is not very good at giving you a clear story so there's always those sorts of balances, checks in the system.
> (Participant 1, Clinician)

The clinician highlights the key role of "curiosity" and draws an interesting analogy between themselves as a clinician working to understand what a patient tells them, and trying to understand what the algorithm's outputs are telling them. The clinician does not necessarily take the patient's descriptions or statements at face value - not because they do not believe or actively mistrust the patient - but because there are many reasons why there may be lack of clarity in a patient's account. For example, there may be many confusing factors in a patient's experience of the condition ("sometimes the patient is not very good at giving you a clear story"). A general sense of being curious about the patient's presentation of a condition is an essential component of diagnosis. The diagnostic puzzle brings out the non-judgemental but epistemically driven attribute of curiosity - though we might also see in this a kind of constructive questioning or scepticism.

For this clinician, curiosity is also about how to make sense of an algorithm. Among the questions they ask themselves are: "do I really believe that?", "is that really true?". In this way, the clinician extends their professional attitude towards patients to the outputs of machine learning: that is, the clinician does not simply and straightforwardly believe it. Much like the patient, the clinician is unlikely to take the algorithm's output at "face value"; but is instead likely to "interrogate that information". This clinician also recognises that just as there is sometimes a lack of clarity in the accounts of patients, there may be a lack of clarity in the outputs of the machine. Crucially it involves both an interpretation of what the patient/algorithm is 'saying', *and* a questioning of its truth, a potential withholding of belief. Here too the collective and collaborative aspect of clinical practice is at play, and the clinician refers to how the checks and balances of other colleagues often work in these situations to raise questions so that diagnosis can be revised and rectified ("there's always those sorts of balances, checks in the system"). The clinician's suspicion towards model outputs on the one hand whilst also acknowledging the different skills required for interpretation chimes with the findings of other studies on computational models and validation (Randall and Wielicki, 1997; Sundberg, 2006).

This questioning and interrogation also leads to a refinement of the whole validation process. This is clear in the quotation below from the data scientist's collaboration with the same clinician, describing their processes of checking the performance of the algorithm:

> What can I solve myself by looking at the data and then what can I raise to them to say 'this looks kind of strange?' I think that's what's hugely valuable […] if you can have a clinical expert to be part of the development procedure I found that to be just priceless because they and the team they saw all of the things that we did, they saw when we were worried, they saw when we were like 'no this actually looks okay now' and I think you can't put a price on the value of that in growing the trust.
> (Participant 2, Data Scientist)

Here again we have an indication of how important the collaboration is: for mutual understandability linked to mutual agreement regarding how it should be tested, for joint 'ownership' of the AI application, and for establishing trust as practice-orientated (Clarke et al., 2006a, 2006b; Oudshoorn, 2008; Papangelis et al., 2019). Nevertheless, for the data scientist, making changes and refinements of the algorithm's variables (after questioning and interrogation) resembles the beginning of a journey through which the clinician acquires the understanding that will eventually allow them to 'get it'. As the data scientist, stated:

> You need your key clinical champions to be part of it and to say 'I've been on this journey with this development and I get it, and I've contributed and I can see where it's going', I think it's so important.
> (Participant 2, Data Scientist)

As the quotation from the data scientist reveals, clinicians act as "clinical champions", thereby opening doors to the broader community. One example is participant 1 whose act as a champion for the screening algorithm and PH community makes sure that the bioinformatics team who are helping them develop the algorithm have access to the clinician's PH networks of clients and partner hospitals they need. The clinician's role as clinical champion is articulated in the quote below from the data scientist:

> Participant 1 invited us to an advisory board where we had about 8 of the different specialists from the 14 centres all across the UK. We presented the algorithm to them, we said 'this is what we're doing'. We invited their comments, we invited a lot of criticism to be honest and it was a very productive discussion and at the end we said: 'we're excited about this, but we need more information and evidence to be sure about it. Would you like to be involved as a collaborator?' and they said 'yes'. So, they've signed a letter for us, which we then gave to NHS digital.
> (Participant 2, Data Scientist)

In the words of Strohm (2019: 35) the clinician as a 'local champion' acts as a mediator of trust and forms a bridge between the computer/data scientists, the AI application in development, and the broader community. There would be very low prospects for external validation without this.

External validation with unseen data in the real-world

The whole process of development is geared towards external validation. These validations require an "independent cohort" (Participant 3, Biomedical Scientist) or "virgin population" (Participant 1, Clinician), that is, the AI algorithm is required to be tested in real-world clinical conditions on prospective data (Topol, 2019; Cabitza and Zeitoun, 2019; Nagendran et al., 2020). For the team that we interviewed developing the screening algorithm, external validation is a "really important" step towards identifying those patients who could be asked to come into the clinic for further tests, in the hope of arriving at earlier diagnosis. However, this process is highly challenging because it involves real people: not only data points in a data set, but people whose data has not been definitively classified and labelled in a clean dataset of 'ground truths', and also who may have a deadly disease:

> What we haven't done yet is a prospective validation which I think is really important. And I would say of all the patients today 'who do we flag as the most high risk?' and then follow them up, so wait a bit of time, so wait for six months, wait for twelve months and see 'did they actually get

> diagnosed or referred?' 'Are these people actually being managed?'
> (Participant 2, Data Scientist)

In the quote above, the data scientist highlights the limitations of data collection in clinical contexts for developing analyses and validating the screening algorithm. That is to say, the data scientist is hard pressed to tell us how they will actually steer this course:

> I think it's a slightly trickier validation because they could still be people who would be PH patients but just don't get referred in that period of time. But I think it's still useful. One of the things that we discussed which we haven't really ironed out yet is we could actually invite them to a clinic and some of the specialists could say 'oh I'd be really happy to bring them into a clinic if you flagged them'. But I think again we would need to think very carefully about what we would need to do in order to operationalise that and also again the risks and all the ethics in all that.
> (Participant 2, Data Scientist)

The data scientist's desire for prospective validation on the one hand, whilst also fearing what this challenge might indicate on the other, is perhaps not unusual and chimes with the desire for prospective validation in clinical contexts (Cabitza and Zeitoun, 2019). The data scientist, in particular, argues that the data collection corpus is still reliant on patients being referred to specialty centres, as initial referral patterns are constantly changing and have different patterns in different regions. Furthermore, in some instances referrals provide only general ICD-10 codes or basic patient information that almost inevitably fail to capture the holistic understandings that can be found in a MDT diagnosis that are critical for establishing ground truth labels and dataset credibility. For all the researchers we interviewed, it was critical to move onto prospective validation, so that ultimately a much broader range of patients could be screened for PH. This process would need to be constantly re-anchored into real-time outputs and closely examined and refined by diagnoses from actual clinical practice. For Elish and Watkins (2020: 50) this validation process is a type of 'feedback loop' which combines clinical expertise and machine learning prediction and, in effect, gives us an idea of how validation will occur in clinical practice as an accomplishment of sociotechnical work.

Discussion

The close collaboration between AI developers and clinical experts throughout the development process brings the AI application out of the laboratory into the 'real world' of its clinical users. This bridging between the laboratory and the clinic brings meaning to AI applications, making their key features interpretable in their context of use. This bridging also affects the way in which the performance of the algorithm is assessed and validated. In internal validation this occurs through comparing the outcomes of runs of the software against different segments of the dataset queried and labelled by clinical experts. Comparisons can be carried out in a number of different ways, and according to multiple different criteria; there are different grounds for comparison for different domains, uses and practices. Identifying which criteria should be used and how depends on *who* is making the assessments and for which purposes, and crucially on how the outcomes are found to be relevant or not, given meaning or not in the context of use (recall that clinician 1 when interpreting the performance of the algorithm in a validation test remarked: "this is something that I can relate to, I can relate to that number"), and how the algorithm is questioned and interrogated by clinical experts, according to their expertise and experience. Finding the grounds for comparison goes hand in hand with the interpretability of the algorithm's outcomes in that context. Like Elish and Watkins (2020), our analysis aligns with the concept of 'articulation work' as a form of invisible work which is necessary during innovation (Star and Strauss, 1999). In our analysis, this process of articulation begins by coordinating and embedding clinical experts into the work of AI developers, and once embedded, participate in iterative activities to get things "back on track", such as the querying of datasets and bridging between internal and external validation (Star and Strauss, 1999: 10).

Bridging performs a social and an epistemic function. We saw that the involvement of both AI developers and clinical experts results in the algorithm gaining meaning, interpretability and comparability in the real-world context of use. We then saw how trust and bridging between laboratory and clinic are in a somewhat circular relationship, which is however not a vicious circularity. One of our clinician participants put the central trust question thus: "if you're going to trust the machine, then that trust has got to be based on something". The 'something' it is based on is built up through collaborative practices in every stage of the development process, and even draws on close interpersonal interactions as part of their every day clinical practice. This makes the validation of the AI application's software jointly produced by everyone in the collaboration, a co-production that establishes the criteria for assessment of the performance of the AI algorithm in a way that is epistemically accessible for all involved, if not absolutely, in ways that are relevant to each expertise and use. Besides this crucial epistemic role, co-production plays a crucial social role in establishing sufficient acceptance for the validation process to proceed to the next stage, with the user-collaborators becoming 'local champions' of the application for the broader community (Strohm, 2019).

Criteria for assessment of the algorithm or criteria for judging its outputs have come to occupy a particularly significant position within the validation context. This has been related by many to be the 'evidential culture' that is required for making credible decisions, where criteria are not defined by a standard of proof or regulatory organisation but emerge in the social dynamics of co-production. It is a collection of human judgements about similarities between objects of interest where people use their experiences of a phenomena in the real-world and anticipate whether it is comparable, or sufficiently similar to tests such as computational models that predict risks (Böschen, 2009, 2013). By focusing on the process of how collaborators establish grounds for comparison which is the basis for validation and for trust, this article offers a novel contribution to this existing focus. Although the Sepsis Watch research allows us to understand how the development and interpretation of criteria for comparison can take place in clinical contexts (Elish, 2018; Sendak et al., 2020; Elish and Watkins, 2020), our research yields an understanding of the crucial role of co-producing the grounds for comparison on which assessments are based, which precede the AI system reaching clinical settings. This article does so by considering the process by which this is achieved where details and nuances matter and remain underexplored.

Even though the grounds for comparison are often expressed statistically, which metrics and which variables go into the statistics are determined in the context of collaboration, depending on their relevance, usefulness, etc. In addition, statistical validity is dependent on a number of important further trust practices in the domain: the trust of the clinicians who query the dataset, in the diagnostic practices of other clinicians; the trust of the computer scientist in the clinical experts querying the data set (a kind of trust by proxy); the trust that each of the collaborators have in the abilities and expertise of the other. Given the social and epistemic complexity of trust practices, it is clear that statistical validity is never standalone, but rests on the shifting sands of these practices. It is hard to find any feature of AI that is an intrinsic, 'objective' feature of an application, as even the statistical assessments are highly relational (cf. Cabitza and Zeitoun, 2019). Far from AI being a technical fix for problems faced in healthcare settings, an AI system that works in these contexts is produced through a complex interplay of social, epistemological and technological factors, that require sustained attention to bring to the surface invisible work and sociotechnical infrastructures underpinning the development process. Research into developing healthcare AI needs to broaden its focus to encompass the clinician's situated participation in sociotechnical work environments when it comes to processes of trust building. Doing this will allow us to reach a better understanding of the details in how trust is engendered, and indeed to assess the extent to which these trust practices are robust, given the real-world tasks that these intelligent systems but perform.

Conclusion

Following the process of developing AI applications for supporting diagnosis in clinical settings shows validation to be neither purely technical, nor simply the final step of the development process in a formal validation phase. Establishing what counts as validation occurs through an iterative and piecemeal process, that brings together people with multiple different expertises, and the real-world contexts in which those expertises are used to make complex decisions. The grounds for comparing the performance of the algorithm with other performances, so that it can be both meaningfully interpreted and evaluated by all those involved with it emerge simultaneously with the developing AI. In this way, epistemic accessibility is built into the algorithm, and traced into it. This allows trust to be built into the system and co-constituted by collaborators throughout the process, and not by some 'end point' realisation. This trust is multi-faceted, as it is engendered by interpersonal, multi-expert collaboration (e.g., computer scientist, data scientist, biomedical scientist) and practical interactions with the technology before it even gets to a formal validation phase. Rather than trust being produced by validation, trust supports meaningful validation. This is not a backward pipeline with the arrows simply going in the opposite direction: it is a form of trust which works in a complex system intertwining social, epistemic and technological aspects. AI development needs to get better at attending to this intertwinement.

Acknowledgements

We would like to thank the Wellcome Trust for the Seed Award that funded this research [Grant number: WT/213606 awarded to Dr. Annamaria Carusi]. We would also like to thank our collaborators who took part in this research, without whom this research would not be possible. Furthermore, we would like to thank the reviewers for their insightful comments on the manuscript, which helped us to strengthen our argumentation. Finally, we would like to thank all the participants who attended the AI in the Clinic Network Event (26/03/2021) whose helpful comments enhanced our thoughts for this article.

References

Alpaydin E (2016) *Machine Learning: The New AI*. Cambridge: MIT Press.

Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V and Montana G (2019) Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology* 00:1-7.

Asilomar AI Principles (2017). Principles Developed in Conjunction with the 2017 Asilomar Conference [Benevolent AI 2017]. Available at: https://futureoflife.org/ai-principles (accessed 07.01.2020)

Barda AJ (2019) *Design and Evaluation of User-Centered Explanations for Machine Learning Model Predictions in Healthcare*. PhD thesis, University of Pittsburgh, US.

Bærøe K, Miyata-Sturm A and Henden E (2020) How to Achieve Trustworthy Artificial Intelligence for Health-care. *Bulletin of the World Health Organisation* 1;98(4): 257-262.

Bjørnstad C and Ellingsen G (2019) Data Work: A Condition for Integrations in Health Care. *Health Informatics Journal* 25:526–535.

Bonde K, Danholt P and Bossen C (2019) Data-Work and Friction: Investigating the Practices of Repurposing Healthcare Data. *Health Informatics Journal* 25: 558–566.

Bossen C, Pine KH, Cabitza et al. (2019) Data-Work in Healthcare: An Introduction. *Health Informatics Journal* 25(3): 465-474.

Bossen C and Piras EM (2020) Introduction to the Special Issue on Information Infrastructures in Healthcare: Governance, Quality Improvement and Service and Service Efficiency. *Computer Supported Cooperative Work (CSCW)* 29: 381-386.

Böschen S (2009) Hybrid Regimes of Knowledge? Challenges for Constructing Scientific Evidence in the Context of the GMO-debate. *Environmental Science and Pollution Research* 16(5): 508–520.

Böschen S (2013) Modes of Constructing Evidence: Sustainable Development as Social Experimentation - The Cases of Chemical Regulations and Climate Change Politics. *Nature and Culture* 8(1): 74–96.

Boullier H, Demortain D and Zeeman M (2019) Inventing Prediction for Regulation: Modelling Structure-Activity Relationships at the US Environmental Protection Agency. *Science & Technology Studies* 34(2): 137-157.

Braun V and Clarke V (2006) Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3(2): 77-101.

Cabitza F and Zeitoun JD (2019) The Proof of the Pudding: In Praise of a Culture of Real-World Validation for Medical Artificial Intelligence. *Annals of Translational Medicine* 7(8):161.

Cabitza F, Campagner A and Balsano C (2020) Bridging the 'last mile' gap between AI implementation and operation: 'data awareness' that matters. *Annals of Translational Medicine* 8(7): 501.

Carusi A (2009) Implicit Trust in the Space of Reasons and Implications for Technology Design: A Response to Justine Pila. *Social Epistemology* 23(1): 25-43.

Carusi A (2014) Validation and Variability: Dual Challenges on the Path From Systems Biology to Systems Medicine. *Studies in History and Philosophy of Biological and Biomedical Sciences* 48:28-37

Carusi A (2016) In Silico Medicine: Social, Technological and Symbolic Mediation. *Humana-Mente Journal of Philosophical Studies* 30: 67-86.

Carusi A, Burrage K and Rordríguez B (2012) Bridging Experiments, Models and Simulations: An Integrative Approach to Validation in Computational Cardiac Electrophysiology. *American Journal of Physiology-Heart and Circulatory Physiology* 303: H144–H155.

Clarke K, Hardstone G, Hartswood M, Proctor R and Rouncefield M (2006a) Trust and organisational work. In: Clarke K, Hardstone G, Rouncefield M and Sommerville I (eds) *Trust in Technology: A Socio-Technical Perspective*. Dordrecht: Springer, pp. 1-20.

Clarke K, Hughes J, Martin D et al. (2006b) 'It's about time": Temporal Features of Dependability. In: Clarke K, Hardstone, G, Rouncefield M and Sommerville I (eds) *Trust in Technology: A Socio-Technical Perspective*. Dordrecht: Springer, pp. 105-121.

Cutillo CM, Sharma KR, Foschini L et al. (2020) Machine Intelligence in Healthcare – Perspectives on Trustworthiness, Explainability, Usability, and Transparency. *Npj Digital Medicine* 3(47): 1-5.

Elish MC (2018) The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care. *Ethnographic Praxis in Industry Conference Proceedings* 2018: 364–380.

Elish MC and Watkins EA (2020) Repairing Innovation: A Study of Integrating AI in Clinical Care. *Data & Society*. Available at: https://datasociety.net/library/repairing-innovation (accessed 06.01. 2020).

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks. *Nature* 542: 115–118.

European Commission (2019) *Ethics Guidelines for Trustworthy AI: High-level Expert Group on Artificial Intelligence*. Directorate-General for Communications Networks, Content and Technology. Publications Office, European Commission. Available at: https://data.europa.eu/doi/10.2759/177365 (accessed 09.11.2019).

Fitzpatrick G and Ellingsen G (2013) A Review of 25 Years of CSCW Research in Healthcare: Contributions, Challenges and Future Agendas. *Computer Supported Cooperative Work* 22: 609-665.

Floridi L, Cowls J, Beltrametti M et al. (2018) AI4People – An Ethical Framework for a Good Artificial Intelligence Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28(4):689–707.

Gaube S, Suresh H, Raue M et al. (2021) Do As AI Say: Susceptibility in Deployment of Clinical Decision-Aids. *Npj Digital Medicine* 4(31): 1-8.

Gulshan V, Peng L, Coram M et al. (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316: 2402.

Jirotka M, Procter R, Hartswood M et al. (2005) Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study. *Computer Supported Cooperative Work* 14: 369-398.

Katzenbach C (2019) Busted: AI will fix it. In: *Alexander von Humboldt Digital Society Blog*, 29 October. Available at: https://www.hiig.de/en/busted-ai-will-fix-it/ (accessed 18.08. 2021).

Kiely DG, Elliot C, Sabroe I and Condliffe R (2013) Pulmonary hypertension: diagnosis and management. *British Medical Journal* 346(1): f2028.

Kiely DG, Doyle O, Drage E et al. (2019) Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension. *Pulmonary Circulation* 9(4): 1-9.

Kuutti K and Bannon LJ (2014) The Turn to Practice in HCI: Towards a Research Agenda. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA, pp. 3543–3552.

Küppers G and Lenhard J (2005) Validation of Simulation: Patterns in the Social and Natural Sciences. *Journal of Artificial Societies and Social Simulation* 8(4): 1-13

Laurent B and Thoreau F (2019) Situated Expert Judgement: QSAR Models and Transparency in the European Regulation of Chemicals. *Science & Technology Studies* 32(4): 158-174.

Littmann M, Selig K, Cohen-Lavi L et al. (2020) Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nature Machine Intelligence* (2): 18-24.

MacKenzie D (1990) *Inventing Accuracy. A Historical Sociology of Nuclear Missile Guidance*. Cambridge, Mas: MIT Press.

Mackenzie D (2001) *Mechanizing Proof: Computing, Risk, and Trust (Inside Technology)*. Cambridge, Mas: MIT Press.

Ming D (2018) This Algorithm Reads X-rays Better Than Doctors Do. In: *Vice News,* 13 December. Available at: https://www.vice.com/en_us/article/kzvkym/this- algorithm-reads-x-rays-better-than-doctors (accessed 16.12.2019).

Morrison M (2015) *Reconstructing Reality: Models, Mathematics and Simulations*. Oxford: Oxford University Press.

Nagendran M, Chen Y, Lovejoy CA, et al. (2020) Artificial Intelligence Versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies. *BMJ* 368: m689

Oakden-Rayner L (2017) Exploring the ChestXray14 Dataset: Problems. In *Luke Oakden-Rayner Blog*, 18 December. Available at: https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/ (accessed 16.12. 2020).

Oakden-Rayner L (2018) CheXNeXt: An In-Depth Review. In *Luke Oakden-Rayner Blog,* 24 January. Available at: https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/ (accessed 16.12. 2020).

Oreskes N, Shrader-Freshette K and Belitz K (1994) Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263: 641-646.

Oreskes N (2004) Science and Public Policy: What's Proof Got to Do With It? *Environmental Science & Policy* 7: 369-383.

Oudshoorn N (2008) Diagnosis at a Distance: The Invisible Work of Patients and Healthcare Professionals in Cardiac Telemonitoring Technology. *Sociology of Health and Illness* 30(2): 272-288.

Pallesen T and Jacobsen PH (2018) Articulation Work From the Middle – A Study of How Technicians Mediate Users and Technology. *New Technology, Work and Employment* 33(2): 171-186.

Papangelis K, Potena D, Smari WW et al. (2019) Advanced Technologies and Systems for Collaboration and Supported Cooperative Work. *Future Generation Computer Systems* 95:764-774.

Perry TS (2017) Stanford Algorithm Can Detect Pneumonia Better than Radiologists. In: *IEEE Spectrum Biomedical Blog,* 17 November. Available at: https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/stanford-algorithm-can-diagnose-pneumonia-better-than-radiologists (accessed 16.12.2020).

Rajpurkar P, Irvin J, Zhu K et al. (2017) CheXNet. Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Available at: http://arxiv.org/abs/1711.05225 (accessed 16.12. 2020).

Rajpurkar P, Irvin J, Ball RL et al. (2018) Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of CheXNeXt to Practicing Radiologists. *PLOS Medicine* 15: e1002686.

Randall DA and Wielicki BA (1997) Measurements, Models and Hypotheses in the Atmospheric Sciences. *Bulletin of American Meteorological Society* 78(3): 399–406.

Scheek D, Rezazade Mehrizi MH and Ranschaert E (2021) Radiologists in the Loop: The Roles of Radiologists in the Development of AI Applications. *European Radiology* 31: 7960-7968.

Sendak M, Elish MC, Gao M et al. (2020) The Human Body is a Black Box: Supporting Clinical Decision-Making with Deep Learning. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 27-30 January 2020, pp. 99–109.

Shackley S, Young P, Parkinson S and Wynne B (1998) Uncertainty, Complexity and Concepts of Good Science in Climate Change Modelling: Are GCM's the Best Tools? *Climatic Change* 38: 159-205.

Sreedharan S, Mian M, Robertson RA and Yang N (2020) The top 11 most cited articles in medical artificial intelligence: a bibliometric analysis. *Journal of Medical Artificial Intelligence* 3(3): 1-12.

Star SL (1999) The Ethnography of Infrastructure. *American Behavioral Scientist* 43(3): 377–391.

Star LS and Strauss A (1999) Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work* 8: 9–30.

Strohm L (2019) *The Augmented Radiologist: Challenges and Opportunities for Widescale Implementation of AI-based Applications in Dutch Radiology Departments*. Master's Thesis, Utrecht University, NL.

Swift AJ, Lu H, Uthoff J et al. (2020) A Machine Learning Cardiac Magnetic Resonance Approach to Extract Disease Features and Automate Pulmonary Arterial Hypertension Diagnosis. *European Heart Journal-Cardiovascular Imaging*. 0:1-10.

Sundberg M (2006) Credulous Modellers and Suspicious Experimentalists? Comparison of Model Output and Data in Meteorological Simulation Modelling. *Science Studies* 19(1): 52-68.

Thoreau F (2016) 'A mechanistic interpretation, if possible': How does predictive modelling causality affect the regulation of chemicals? *Big Data & Society* July-December: 1-11.

Tonekaboni S, Joshi S, McCradden MD and Goldenberg A (2019) What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of Machine Learning Research 1-21*. ArXiv 2019; published online May 13. https://arxiv.org/abs/1905.05134 (preprint)

Topol EJ (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine* 25(1):44-56.

Van Baalen S, Carusi A, Sabroe I and Kiely DG (2017) A social-technological epistemology of clinical decision-making as mediated by imaging. *Journal of Evaluation in Clinical Practice* 23(5): 949–958.

Van Baalen S and Carusi A (2019) Implicit trust in clinical decision-making by multidisciplinary teams. *Synthese* 196: 4469–4492.

White DD, Wutich A, Larson KL, Gober P, Lant T and Senneville C (2010) Credibility, salience, and legitimacy of boundary objects: water managers' assessment of a simulation model in an immersive decision theatre. *Science and Public Policy* 37(3): 219-232.

Winsberg E (2010) *Science in the Age of Computer Simulations*. Chicago: Chicago University Press.

Winter P and Carusi A (*Forthcoming*) (De)Troubling Transparency: Artificial Intelligence for Clinical Applications. *Medical Humanities.*