

# Trust and AI in Clinical Decision Support

Ben Wilson

1915417

Submitted to Swansea University in partial fulfilment  
of the requirements for the Degree of Master of Science



**Swansea University**  
**Prifysgol Abertawe**


Department of Computer Science  
Swansea University

30th September 2020



## Declaration


This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed  (candidate)

Date 30-Sep-2020

## Statement 1


This work is the result of my own independent study/investigations, except where otherwise stated. Other sources are clearly acknowledged by giving explicit references. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure of this work and the degree examination as a whole.

Signed  (candidate)

Date 30-Sep-2020

## Statement 2

I hereby give my consent for my work, if accepted, to be archived and available for reference use, and for the title and summary to be made available to outside organisations.

Signed  (candidate)

Date 30-Sep-2020



# Abstract

Machine learning (ML) and artificial intelligence (AI) techniques are increasingly visible in the domain of clinical decision support. Significant results have been obtained in the interpretation of diagnostic imaging using computer vision techniques - notably in the specialties of ophthalmology, dermatology, cellular pathology, cardiology, oncology and respiratory medicine. Risk prediction and prognosis tools are also beginning to emerge. However, there is limited research on how well and how far these techniques can be integrated into real clinical workflows. This project set out to study one of the key success factors in translating research into effective practice - that of the trust clinical decision-makers place in the intelligent systems they are able to access.

The project made use of initial clinical user engagement through workshops, interaction studies and interviews to develop a broad online user study that measured how a user's trust varies with different system characteristics. The intensive workshops and interaction work were used to inform design elements for the main study. In the main online study, clinical decision-makers were asked to evaluate seven different hypothetical systems in three different clinical contexts.

By holding the clinical decision context constant for a set of system characteristics in the online study, and by randomising the other presentational variables, the experimental methodology provides assurance that the results are indeed able to indicate a real signal if it exists.

In the course of preparing the online study, we demonstrated a successful approach to clinical engagement that constitutes a framework for further development of participatory design work.

We provide a scalable web-based assessment tool with a proven ability to capture sustained clinical input.

We co-created a clinically-informed subset of AI system characteristics that can be used to explore key components of the complex trust relationship users will experience when engaging with intelligent systems in clinical decision contexts.

We show some initial results that show promise for future work where more data can be collected.

This work suggests that there is value and opportunity in further exploring the characteristics in AI systems that engender trust in clinical decision-makers.

# Acknowledgements

I would like to acknowledge the support, advice and encouragement of Matt Roach, Alma Rahat, David Rawlinson, Fiona Goldman, Hamish Laing, Anne Marie Cunningham, Rhidian Bramley, Dafydd Loughran and Sabrina Zulfikar.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Overview . . . . .	1
<b>2</b>	<b>Literature review</b>	<b>3</b>
2.1	Trust as a social relation . . . . .	3
2.2	The determinants of trust . . . . .	5
2.3	Related work . . . . .	14
<b>3</b>	<b>Design</b>	<b>17</b>
3.1	Participatory design - what are we exploring? . . . . .	17
3.2	Participatory design - how do we explore it? . . . . .	25
<b>4</b>	<b>Implementation</b>	<b>33</b>
4.1	Advance recruitment . . . . .	33
4.2	Platform - practicalities under covid restrictions . . . . .	33
4.3	Technology stack . . . . .	33
4.4	Web framework . . . . .	34
4.5	Page content combining patient stories and AI systems . . . . .	34
4.6	Randomising the UX . . . . .	35
4.7	Testing . . . . .	35
4.8	A late change to participation . . . . .	36
4.9	Deployment . . . . .	38
<b>5</b>	<b>Results and analysis</b>	<b>39</b>
5.1	Nature of data collected . . . . .	39

5.2	Summary of data collected . . . . .	40
5.3	User predisposition . . . . .	43
5.4	Evaluation of the AI system characteristics . . . . .	45
5.5	Clinician vs non-clinician evaluations . . . . .	56
5.6	Qualitative responses . . . . .	57
5.7	Observations on the range of clinical participants . . . . .	58
<b>6</b>	<b>Discussion, Conclusions and Future Work</b>	<b>59</b>
6.1	Discussion . . . . .	59
6.2	Conclusions . . . . .	64
6.3	Contributions . . . . .	65
6.4	Future Work . . . . .	65
	<b>Bibliography</b>	<b>67</b>

# Chapter 1

## Introduction

This thesis describes an online user study conducted during September 2020. During the conception, design and execution of the study, the world of clinical decision-making (and with it the technology to support clinical decision-making) has been heavily impacted by the global COVID-19 crisis.

### 1.1 Motivation

Machine learning and artificial intelligence are set to bring about significant change in the work of clinical decision-makers. The importance of understanding the factors that allow appropriate trust to be built and sustained is significant.

In this document we present the context, work and results of preparing for and carrying out an online user study to explore factors affecting trust in relation to AI in clinical decision support. We outline the significance of the results both of the preparation and the study proper. And we suggest further work that is indicated by what has been found so far.

### 1.2 Overview

The remainder of Chapter 1 serves to orientate the reader by outlining the document structure - which is as follows.

Chapter 2 contains a review of the relevant literature. Sections within this chapter start with an exploration of the nature of trust and its character as a social relation, move on to review the factors that determine trust, and conclude with a review of closely related work.

Chapter 3 describes how the design of the experiment, the stimulus material and the website evolved through participation, discussion, iteration and testing. This considers two distinct parts of the design work.

First, the shaping of the design space - the patient stories that provide a decision context, the different artificial intelligence (AI) system characteristics that should be tested, and the presentation of these elements to the user. This co-creation of the user-experience in the experiment is a significant part of the work as it engaged a number of experienced clinicians in a key part of the design phase.

In the second part, the chapter looks at options to utilise the elements selected in the co-creation process. Specifically, it considers how those elements should be combined to produce useful measures in an online study - that is, the most suitable methodology for using these elements to explore trust. Usability is considered again, this time in terms of ensuring consistency and reliability in user-response.

Chapter 4 describes how the online experiment was implemented. This begins with how recruitment preceded launch. It goes on to provide details of the technical components used, the requirements met by the website and work of realising the user-experience elements. A description of the testing and deployment conclude the chapter.

Chapter 5 outlines the data that were collected during the online user study - summarising and profiling the overall dataset. It then describes the substantive evaluation data. And it includes some observations on participation.

Lastly, in Chapter 6 we discuss the issues arising from and identifiable in the process and results. We summarise the lessons and draw conclusions. A note is made of the contributions represented by this work. And then we discuss what should be done as future work in light of this study.

## Chapter 2

# Literature review

This chapter begins with a perspective on trust as a social relationship. We relate this to society's view of science, technology, computer science and health technology in general. We then go on to consider the relatively new entry of AI onto this landscape. If trust is a relation between humans (even if mediated by technology), how do we establish a social relation with an AI system? How do we bridge the divide between human considerations and artificial considerations?

### 2.1 Trust as a social relation

What is involved when we decide whether or not to place trust in anything? The distinguished sociologist and peer, Anthony Giddens defines trust as confidence in the reliability of something [1]. He argues that, despite appearances, it is just as much a social relation when we trust complex systems, machines, processes or organisations as when we trust a fellow human. This is because our thinking has "removed social relations from the immediacies of context. In other words, we experience the social relation in a mediated form. When we trust a complex process, we do so because we trust the social mechanisms and ultimately the people that lie behind it.

In a sequence of writings, Professor and Oxford Fellow, Diego Gambetta focuses more practically on the mutuality of humans trusting other humans [2, 3]. And in so far as studies of machines (agent models) replicate social networks, they also foreground the symmetries of trust and mutuality [4]. On the other hand, the *asymmetries* of trust that characterise relations between individuals on the one hand and the *organisations* with

which they work or interact on the other, has given rise to a significant literature that emphasises culture and reputation [5, 6]. One might ask, when it comes to AI systems, are they more akin to individuals or to representatives of their 'parent' *organisations*?

In a series of contributions on human-AI relations, Professor Joanna Bryson takes issue with the very idea of trusting an AI system [7, 8] "Trust is a relationship between peers", she argues. And so humans never can, and never should, trust an AI system. Rather, "we need to know we can hold the human beings behind that system to account" so there is no need to trust the AI system itself.

What is common to both Giddens and Bryson is their stipulation that, for trust to be placed, a prime condition is lack of full information. That is to say, complete transparency in a process obviates any need for trust [1, 7]. What distinguishes them is that Professor Bryson's argument makes a case for stripping away the mediating forms and placing the social relations at the centre of legal accountability. While Lord Giddens' approach places emphasis on the experience of the person placing the trust (or not). In our view, both are correct - for different purposes. Our particular task requires that we follow Giddens - we explore and assess the confidence users might place in the reliability of a system and focus on what is of practical help in the clinic at decision time.

It's worth noting that digital technology in health, even without the involvement of AI systems, has a less than glowing reputation with clinicians. Issues of usability [9, 10, 11, 12], data quality [13, 14], complexity and safety [15, 16] and interoperability [17, 18] are widely reported. And in a recognition of the need for better regulation, a recent major government-commissioned report in the UK calls for a legal and ethical governance framework for digital health products [19].

Despite concerted efforts to encourage favourable public opinions of science [20, 21, 22, 23, 24], public trust in science remains highly variable. Few surveys find concerns over the use of electric power in kettles or electric blankets or in the promotion of what amounts to lethal levels of power in consumer goods (cars, power tools). And in many fields, the 'most advanced' technology is frequently the most prized and trusted. But some aspects of food safety, public health and computer technology - for example, genetic modification, vaccine development and 5G communications respectively - among others, tap into fears about science overreaching itself with Frankenstein-like hubris [25].

Machine learning and AI are technologies that are frequently cited in connection with fear and dystopian outlooks. Concerns range from mass redundancy [26, 27] to loss of

human autonomy [28, 29]. And there are likely to be few anxieties that will have been quelled by a British Prime Minister this summer complaining that a previously ‘robust’ educational algorithm [30] was suddenly revealed as ‘mutant’ [31].

The reverse of the sense of awe can also be a problem for science. Some argue that scientific consensus amounts to ‘just another opinion’ and give it no more weight than any other. Science denial and pseudoskepticism [32] are set to continue playing a large part in public debates over epidemiology, climate science and sexuality in the next decade [33].

Trust in this space is complex. Researchers and practitioners are wrangling with delicate social interdependencies. Questions such as food safety and sustainability vvy with climate change, human rights and ethnic, cultural and religious freedoms in contests over how the world should move forward. Marina Jirotká’s team produced a 2019 HCI paper on human-centered responsible innovation in computing that describes ‘anticipatory governance’ [34]. A reflection of the fact that work is often required to ‘rebuild’ trust in this space.

And, at least according to Deng and Varzi, computer science itself has some work to do to improve its methodology in this arena. In a 2019 paper they make the point that there is little critical evaluation of the assumptions of computer science [35]. Dodge et al make a related call for improved reporting of results in machine learning [36]. Trust receives a specific focus in the context of human-AI interaction and relations [37, 38, 39]. And addressing trust via methodological assurances as much as by raw measures of performance is a key feature of any analysis of how AI applications will be extended [40, 41, 42].

To return to the relationship, we focus on the user experience. The reason we think the aim of this study has value is because it is an attempt to explore the ways in which the relationship is *experienced* by the users of AI systems. Whether we accept that our relationship is primarily with the system itself or with the humans behind it is not the direct point of interest here. What we try and explore is the *mediation* of that relationship through users’ actual experience. How do users think about and respond to the characteristics they experience during use?

## 2.2 The determinants of trust

Several writers consider the determinants of trust. Siau and Wang 2018 [43] discuss a tri-factorial model in which human, environment and technology characteristics combine to determine trust in technology. They go on to suggest that explainability helps establish

initial trust in AI systems - but that distinct trust-nurturing characteristics must be employed to develop continuous trust. These include usability, security and sociability.

Our approach has been to begin with the broad sweep of characteristics and allow our research and our clinical collaborators to sift out those they thought most pertinent to clinical decision support.

In this section we review six areas in which AI systems can demonstrate their reassuring qualities. As determinants of trust, they are by no means exhaustive. But they have been included for review following our workshop discussions with clinical participants. The same six characteristics are more or less recognisable in the different hypothetical AI systems featured in our online study described in the rest of this report. Although there they appear alongside an 'uncharacterised' (*vanilla*) version of an AI system. Here, rather than justify their inclusion, we explore their significance, strengths and challenges.

### 2.2.1 Explainability

Many writers argue that explainability is necessary for trust to be garnered [44, 45, 46, 47, 48]. There are other motivations for explainability - epistemic causality and ethical responsibility being two important examples [49]. In fact, since explainability is seen as essential for much more than trust, there is often a critique of the simple idea that trust and explainability are directly related. But the critique is one-sided. While it is asserted that explainability is necessary for trust, rarely is there a question as to whether explainability is sufficient for trust.

The recent growth in scholarly publications related to explainable AI is testament to the importance placed upon it by the research community [45]. In clinical settings studies have included the use of counterfactual treatment outcomes [50], independent conditional expectation (ICE) plots [51], and feature importance plots [52]. But many of the most prominent works make the assertion that trust and explainability are directly linked without citing research to justify the claim. Ribeiro's excellent work on local interpretability [53], to take one prominent example, describes two levels of trust (in an individual result and in a model) and then asserts that both are 'directly impacted' by understanding. It is assumed that the truth of the assertion is self-evident. And this is not an unreasonable starting point. But trust is a complex entity and its importance means that such foundational statements warrant a greater degree of attention.



Phil Blunsom's Oxford team reviewed a series of model explainers [54] focusing on the need for robust explanations. And in a paper essentially about meta-trust ('can we trust the trustworthiness feature itself?' would be a good paraphrase of the title) they assert the need for explanation as a pre-requisite. Similarly, in an important paper that calls for interpretable (inherently understandable) methods rather than explainable (post-hoc rationalisation) methods, Cynthia Rudin argues that explainers suffer from the jeopardy of more or less guaranteed error even if in a minority of cases [55]. But in arguing for interpretability, the assumption is implicit that trust automatically follows. Other work on interpretability parallels this view. Doshi-Velez and Kim assert that interpretability can be used to confirm trust as a desired trait [56].

In another seminal work in this field, Scott Lundberg's SHAP paper [57] asserts that interpretation is 'extremely important' because it engenders trust. But nothing further is explored in the relationship. The assertion is a starting point that is taken as read. In contrast, in a very interesting study this year, New Zealand surgeons Diprose, Buist et al [51] establish a significant relationship between physician understanding and trust using a single patient story presented with four different explainable AI decision support techniques and a control. However, no other potentially trust-inducing factor is varied in the experiment to compare with explanation. And they report that no difference was found in the level of trust engendered by the four techniques.

It could easily be argued that it is an unfair criticism to level at explainable AI researchers that they assume the connection between explanation and trust. After all, the papers cited here are ones that have had significant impact owing to their technical innovation and scholarly insight into the challenges and opportunities of machine learning and AI. But the reason for making this observation is to highlight how strong the assumption is. And to suggest that there is room for additional work in exploring what underlies it.

As an emphasis on this point, we turn to a very useful paper by Ehsan Toreini et al at Newcastle University that explores the relationship between AI technologies and trust. In this comprehensive and well-referenced survey of multiple determinants of trust, the authors cite only one source (Lipton's Mythos paper) to support the claim that explainability increases trust. And while Lipton indeed does argue that this is the case, he nowhere provides a source where the idea has been tested against other possible trust-inducing factors [44].

In a highly significant paper by Riccardo Guidotti et al at the University of Pisa, the relationship between trust and explainability in deep learning is also left unsubstantiated [58]. This is a compelling 40-page survey of explainability methods. But its only support for the claim that explainability increases trust comes from a series of papers on rule-based techniques.

A final note of caution on explainability comes from evidence in two 2019 papers that the effects of explanation are complex and potentially problematic. Where domain knowledge is high, the explanations available from intelligent systems are insufficient to influence trust. Worse still, if domain knowledge is weak, the provision of an explanation increases confirmation bias [59]. So it obscures the supposed benefit of an explanation in allowing a user to spot and reject unhelpful decision support, even when the machine error is sizeable [60].

### 2.2.2 Performance

In contrast to the challenges of quantification and definition we face with explainability, there is a reassuring sense of objectivity associated with performance. Precision, recall, F1 and support values can be quoted and directly compared between systems. They are well-defined, widely accepted and reproducible. But there are, of course, questions to be raised about what exactly is being measured [35, 36]. An impressive lab performance score that fails to translate into a real-world application is inevitably seen as an example of something over-promised and under-delivered. The effect on trust can be catastrophic.

In an interesting study by Ming Yin et al last year [61], a psychological component of the trust relationship is observed. They used Amazon Mechanical Turk to conduct a series of experiments that find observed accuracy, if different from stated accuracy, does indeed affect users trust in a decision system. In particular, when significantly over-promising and under-delivering, there is an observed effect that involves users discounting machine decisions that would have been worth following. And as mentioned in Section 2.2.1, Siau emphasises the importance of continued performance for sustaining trust [43]. Yu finds that a system performance threshold (of 70%) appears to exist above which users more positively engage in constructing good collaborative decisions with intelligent systems [62].

In addition to this reflection of social relations, we should consider the impact of what is measured, what is measurable and whose threshold determines acceptability in terms of performance. The acts of selecting and measuring are not value-free, objective

activities [35]. We should ask whether existing performance measures adequately reflect the actual performance exhibited for different social sub-groups? Where inequalities in health outcomes are known, do we have the capability to re-design what we count as performance to begin to address the issues? Until recently, Buolamwini's critique of facial recognition technology [63] was largely uncommented in general news content. But a change in attitudes to race, surveillance and profiling during 2020 in the US led to three tech giants disengaging in development and supply programs [64, 65, 66]. The performance hasn't changed. But what counts as *acceptable* performance has.

In this context, the social relations and deployment landscape will make a notable difference to how we view the technology. In contrast to the US, an account of work in a resource-limited setting (a network of HIV clinics in Western Kenya), shows how facial recognition software may overcome the challenging lack of unique patient identifiers [67] and lead to better health outcomes.

In the wider clinical context, Challen [68] reports examples of unwitting introduction of bias into clinical datasets. This has implications for what machines learn from those datasets. An approach to responsible AI for healthcare is advocated by Wiens et al [69] to protect against harm from bias. But an AI approach could also serve to reduce bias in that it can be used to explore a greater field of alternatives for variable selection than would traditionally be explored in clinical studies [70].

And while it's important to place an obligation on AI and ML to be fair, it's easy to overlook the existing biases that permeate healthcare provision. In 2000, the NEJM reported that in Emergency Department presenting Heart Attacks, the risk of misdiagnosis was over four times higher for non-white patients and nearly seven times higher for women [71]. This is primarily owing to the historic study of heart attacks being modelled on white men and their data. In 2019, an obstetric journal reported that "Black women are three to four times more likely to die a pregnancy-related death as compared with white women." [72] So another key question for how performance relates to trust is whether high performance on existing data will serve to codify existing bias?

To return to the relationship, what is critical in the effectiveness of decision support systems is not the performance of the system itself, however that is measured, but the performance of the human-system collaboration. As noted in Section 2.2.1, trust in this context can be double-edged. Yunfeng Zhang's FAT\*/FAcT paper 2020 [73] makes an important distinction between *enhancing* trust and *calibrating* trust. Increasing misplaced

trust (*enhancing* trust in an erroneous suggestion) is certainly problematic. Knowing when to trust (having well-*calibrated* trust) should be the ambition.

### 2.2.3 State-of-the-art approaches

Intelligent systems are already shown to be effective in clinical contexts where there is high throughput of reproducible or standardisable decisions [74, 75, 76, 77, 78, 79]. Where there are large, structured datasets from imaging, sensors or laboratory analyses (such as dermatology, cardiology, ophthalmology and histopathology), AI systems have the potential to identify outlier values or patterns and triage large caseloads so as to maximise human attention on borderline instances.

Recent techniques for risk prediction are able to demonstrate effective improvement in clinical outcomes [80, 81, 82], albeit some of these improvements are modest owing to unaccounted factors [83]. And some provide intelligibility on individual cases while demonstrating high accuracy [52]. In a similar vein, discovery of predictor variables for breast cancer [84] and respiratory disease [85] has made progress.

Indeed, deep learning has made continual advances in the clinical domain. In a systematic review of clinical deep learning models, Liu et al find that they have higher pooled sensitivity and specificity than clinicians [86].

Going a step further, the development of automated machine learning (AutoML) aims to support the ambition of a domain expert to develop an entire machine learning pipeline without any prior machine learning expertise. Pickhardt (2020) reports good results in predicting cardiovascular events using this approach [87]. Related work by the van der Schaar Lab has had significant success with risk prediction (autoprognois) [88, 89, 90, 91, 92, 93, 94].

On the other hand, there are concerns about the security of AI deployed in clinical contexts. Given the significance of the AI behaviour, this is to be expected. Safety - with AI as with other software - is frequently compromised when sufficient attention is not paid to how human-computer interfaces need to work [95]. Mozaffari et al speak of poisoning attacks [96] while adversarial attacks on medical ML are reported in Finlayson 2019 [97] and Sittig reviews the landscape in 2020 for patient safety arising from health IT [98].

In each case, state-of-the-art approaches are shown to move the frontiers of capability. As with other contexts, reputation may be developed in one application and transferred to another with some success. But critical to all these is the method of translation. The

suitability of a particular machine learning architecture or ensemble may be estimated in advance. But there are reasons to be cautious about prospects for replicating the successes found in image-based diagnostics and risk estimation [99]. As a recent Lancet comment put it, the development of clinical AI must start with ‘the pull of unmet clinical needs, rather than the push of technology’ [100].

In this context, how we gauge and manage ‘unmet clinical needs’ is also relevant. Are needs only detectable through market mechanisms? Andrea Downey reports objections by commissioners and by GPs to the expansion of BabylonHealth (a primary health disruptor) in 2019 [101, 102] following reports of the adverse impact on its neighbours in the London health economy. Babylon Health make large-scale use of new technologies including AI-driven self-triage and mobile consultations. And they have exploited the government-led marketisation of healthcare provision and altered the landscape of primary care commissioning by breaking free of geographical constraints and attracting fee-paying young, relatively healthy clients. Another BMJ piece summarises the lessons of Babylon [103] as being around the implementation of policy rather than of AI.

There are plenty of writers forecasting how AI will change medicine [104, 105, 106, 107]. Many of these focus on organisational, administrative and workflow improvements. Improved physician ordering and review of tests, reduced errors, improved administration [108, 109, 110, 111]. But in the field of direct clinical decision support we note that once again it is the combination of the capability of the intelligent system and the capability of the clinical decision-maker that is the key to clinical effectiveness [112, 113].

A perspective piece in the influential *New England Journal of Medicine* on machine learning and medicine [104] points out that, ‘predictors are not causes’. An illustration of the need to deploy with great care is that observational data (the data available at quantities suited to training deep neural networks) are confounded by current care interventions. An asthmatic patient will be treated more intensively for respiratory disease than a non-asthmatic. As a consequence, their observed risk may appear lower. But an AI system that began with this raw data and no contextual insight might suggest the lower risk requires less intensive therapy.

The reason healthcare is a challenging space for intelligent systems is because modern medicine is a situated sociotechnical system [114, 115, 116, 117]. Indeed in the British Journal of General Practice, Nick Summerton questions whether AI can help GP diagnosis or just add to stress in clinic [118]. Effective human collaboration with the system requires

not just a disposition to trust the technology but also a framework and context in which that trust can be deployed, developed and continually calibrated alongside the capability of human actors and their collective work practices to evolve. For this work to progress, it is essential to recognise the value of human-centered evaluation [119].

### 2.2.4 Data

For clinicians, data impacts trust. The gold standard resource for clinical decision-making is the randomised controlled trial (RCT) feeding into a systematic review. And one of the crucial reasons for systematic reviews is that RCTs themselves are of variable quality. In common with RCTs, clinical machine learning needs to attend to reproducibility [120, 121]. And there is now discussion in journals (eg, JAMA) of how to conduct RCTs of clinical AI - to build trust [122]. Adaptive clinical trials are suggested by Atan [123]. And AI may play into the improvement of trial methodology. Marshall et al (2015) speak of automating the process of assessing risk of bias in trials [124] where humans are frequently attending to other issues and fail to notice weaknesses in the source data.

As mentioned in the comments on bias in Section 2.2.2, the quality of training data has the potential to affect reliability and hence trust. Quantities of available, reliable, labelled and preferably structured data are crucial for state-of-the-art techniques (and hence the trust that can be placed in them). Even with high-performing reinforcement learning techniques [125], there is a dependence on reliable accumulated decisions with judgements about whether they had desirable outcomes or not. So source data are critical. The UK Chief Medical Officer's report on global health security cites equity, sustainability and security as the three themes [126]. This emphasises the need for good data and evidenced trust.

In a call to recognise a watershed moment in healthcare, Melissa McCradden et al argue that there is a new opportunity to improve our use of data. 'Bias is not new', they say, 'however, machine learning has potential to reveal bias, motivate change, and support ethical analysis while bringing this crucial conversation to a new audience' [127].

### 2.2.5 Risk

As with data, evaluations of risk are critical in clinical decision-making. The situation is made more complex by the fact that different clinical contexts require different risk measures. While relative risk (RR) may be helpful when considering the effects of significant lifestyle changes on heart disease, if the condition is one with very low incidence, such as

omphalocele during pregnancy, then even high relative risk values may be less informative and an absolute value is often more useful [128, 129].

Further complication arises from the different concepts that can be estimated. As a statistical tool, an odds ratio (OR) may have an advantage over a RR value in providing a measure of diagnostic clinical risk. This is because an odds ratio can be calculated in case-control scenarios where absolute and relative risk figures are not known. On the other hand, since one of the main purposes of a risk estimate (especially prognostic risk) is to support shared decision making, most clinicians will prefer to work with a RR value - whose meaning they can convey more readily to a patient or relative. Attributable risk and number-needed-to-treat (NNT) statistics are further variations with their own characteristic usefulness [130]. And each is commonly used in its own context. So any AI system may need to be context-sensitive to be useful.

Non-AI risk prediction tools are used in various clinical specialties. They are most often based on multivariable regression models. And a common challenge is one of continual refinement so as to apply the most suitable model for the population at risk - as with the evolution of the POSSUM surgical risk system [131] or APACHE critical care risk score [132].

As distinct from the data available to an AI system, the production of a risk value involves a level of sophistication in what the system does with the data it has at its disposal. And with this sophistication comes a significant challenge wherever the tool might be applied beyond its capability. In attempting to address potential bias (a feature of existing health tools [133]), designers of AI systems need to try and tease out biologic from prejudicial social factors that influence clinical outcomes [134].

Despite the challenges, high-volume, data-rich specialties such as cardiology have been exploring AI-based risk prediction to aid decision-making for some years. Early attempts were useful lessons in moderating expectations [135]. While more recent work is showing signs of promise - albeit in 'bench tests' rather than live trials [136].

For AI-generated risk estimations to become effective, however, it seems likely that appropriate reliability and well-understood limits to applicability will be more important to clinical users than the ability to pinpoint causal relations. As Sendak et al argue, rather than become expert in all the technical and statistical nuances, many physicians would rather know that they understand the concepts and significances of risk values while being able to trust the detailed derivation to some other function in the healthcare system [117].

### 2.2.6 Approved processes

Guidelines for responsible AI have been produced by a wide range of organisations. Indeed, the importance of trustworthiness in innovations and deployments is reflected in the development of a multitude of codes and guidelines on AI [137, 138, 139, 140, 141]. Toreini lists 32 different principled AI frameworks published by various governmental, academic and industry bodies. But there is a call for more than guidelines.

Iacobucci says new regulations need to fill gaps and consider things like what should happen when diagnostic software misses symptoms [142]. The UK Academy of Medical Royal Colleges argue that the GMC, the CQC and NHS Digital will each need to play important parts in creating the appropriate approval landscape for AI systems [112]. Whether regulations, guidelines or professional body approval, there will need to be attention paid to the relationship between clinicians and any approval or constraint system. Approval that is not respected will rapidly lose any value.

In an ACM ethics paper, Larosa and Danks argue for a new regulatory framework in order to sustain trust - in particular, they focus attention on the way AI systems may alter the relations between patients and clinicians and other care-giving roles [143]. Paton and Kobayashi argue for an open approach to AI in healthcare in which clinical relevance and reproducibility are promoted [144].

Whether approved processes influence trust will depend largely on whether they are generated by trustworthy organisations with trustworthy processes - ultimately, this may come down to whether they are seen to influence the other determinants of trust.

## 2.3 Related work

A few writers have conducted experiments to gauge how well users trust different AI characteristics. Ashoori et al [145] used 362 participants recruited on Amazon Mechanical Turk and manipulated seven system characteristics in the context of non-clinical decision scenarios. The seven characteristics were: stakes, decider, trainer, interpretability, train and test data, social transparency and confidence. To vary the stakes of the decision, one scenario was on prison sentencing; another was on meal planning. The design of this experiment allowed the researchers to see whether participants placed different trust in systems with more or less explainability, more or less information about training and more or less information about performance. They were able to show that users felt



more able to trust an AI decision system making a high-stakes decision if it provided more information rather than less. Explainability showed this pattern more strongly than other characteristics.

Drozdal et al 2020 looked at trust and AI with autoML [146]. They conducted a think-aloud evaluation with four computer scientists plus a controlled experiment and parallel card-sorting exercise with 20 machine learning practitioners to find users' needs in respect of trust in AutoML. They found that the addition of information led to increased trust.

Fewer studies have been done on how clinicians view the output of AI-driven decision support systems. What is seen to instil confidence by clinicians is studied by Lahav et al [147]. In a two-part study, they asked 30 machine learning practitioners to try and predict what characteristics doctors would find engendered trust. Then they surveyed 14 doctors to see what they actually valued. The results showed that ML developers could not predict what doctors will prefer as characteristics that engender trust.

A slightly older study from the realm of Human Factors suggests that if you prime users with a negative mood before testing their response to an automated system, they trust the system less [148]. How much this is true of clinicians would be useful to know since many are exposed routinely to personal stories of loss and pain.



## Chapter 3

# Design

This chapter is divided into two sections. The first describes how a clinical advisory group was formed and used a substantial workshop to review a lot of material that was more or less familiar (AI systems, clinical decision contexts, scenarios of greater or lesser acuity). This produced the framework for individual collaborative sessions between an advisory clinician and a researcher. The purpose of these two kinds of work session in the first section was to build up the contextual detail of what our online study would explore. How would we use different clinical decision points to examine the relationship between clinicians and different AI systems? What characteristics should those systems display in order to provide an effective study?

The second section focuses on the methods to be used online in the substantive study to conduct the explorations determined in the first section. Here we describe the process of designing the study experience so as to generate useful data. This involved the clinical advisors as potential users of the study materials.

### **3.1 Participatory design - what are we exploring?**

#### **3.1.1 Clinical participation**

As mentioned, a significant feature of the work for this study is the involvement of a number of clinicians in the design process. We were able to make contact with a senior clinical academic in Swansea in addition to existing clinical contacts to work up initial ideas. An outline proposal was presented to the academic, a professor and former medical director, using visuals over zoom in mid July. And following this, he agreed to support

our aspiration of creating a small advisory group for the project. The group members were recruited over the next week, so that we had six clinicians who were willing to participate in the design and development of the study itself. In addition to the professor (a surgeon) the advisory group comprised an anaesthetist, a GP, a consultant radiologist, a chest physician and a doctor working in health technology. Two of the group were female.

### 3.1.2 Workshop preparation

A half day zoom workshop was convened with this group in late July during which a substantial part of the literature and design space were described. Preparation for the workshop involved the creation of numerous visuals in the form of mind-maps to communicate the related concepts and linkages around trust, decisions and clinical contexts. Mindmup2 software was used for this preparation [149] as it allowed rapid development and online sharing. An example is in fig Fig. 3.1.

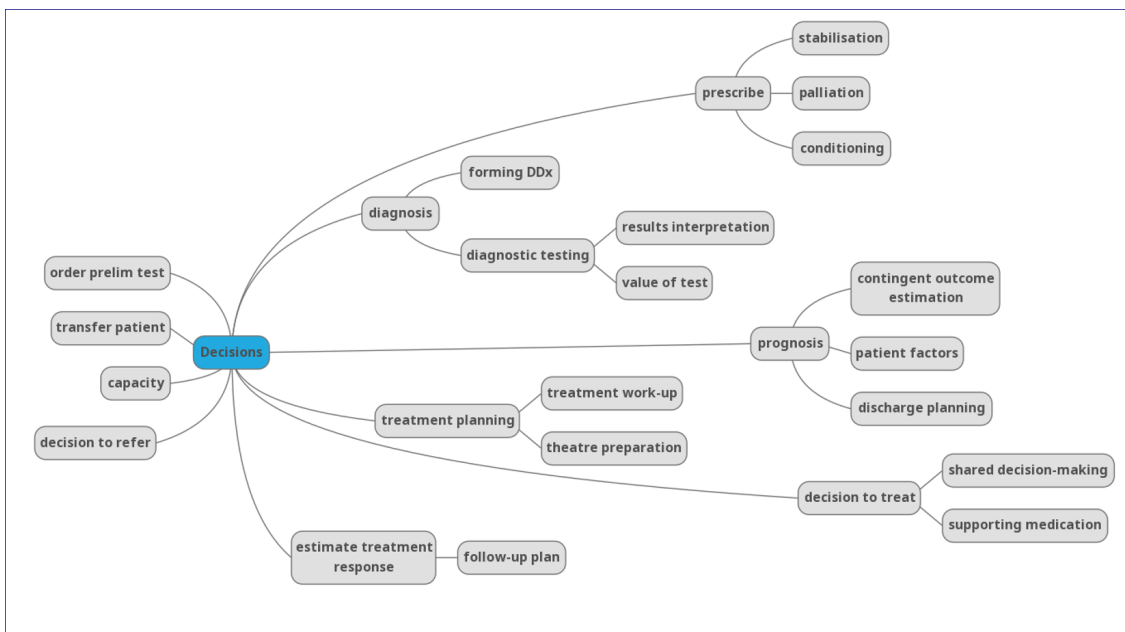


Figure 3.1: Mind map of clinical decision points. A highly simplified diagram of patient pathways in a district general hospital indicating all the points where clinical decisions would typically be made.

We wanted to use prepared activities to provide a shared stimulus so we could ask the clinical advisors to help shape the questions that we would eventually ask the study participants. The aim was to consider which type of clinical decisions would be suitable

for the widest possible clinical audience and at the same time which type would be best for drawing attention to how, and how well, an AI system might support the deliberation. All of this shaping of the user experience for participants was aimed at eliciting the greater or lesser sense of trust that might be created in the mind of the decision-maker.

### 3.1.3 AI system characteristics

To ensure we had a common focus and shared a degree of insight into the problem, a reasonably in depth presentation was prepared on the nature of clinical machine learning and AI. Because our prime focus was always the question of trust, we needed to ensure a common perspective on the real (as opposed to imagined) strengths and weaknesses of intelligent systems. Included in this presentation was an account of the architecture, performance and explainability features of existing AI technologies. This account included current examples of clinical applications showing the potential advantages of using AI techniques in clinical decisions. A workshop slide (Fig. 3.2) showing an example of the latter, provides a measure of computer vision performance in dermatology published by Esteva et al [74].

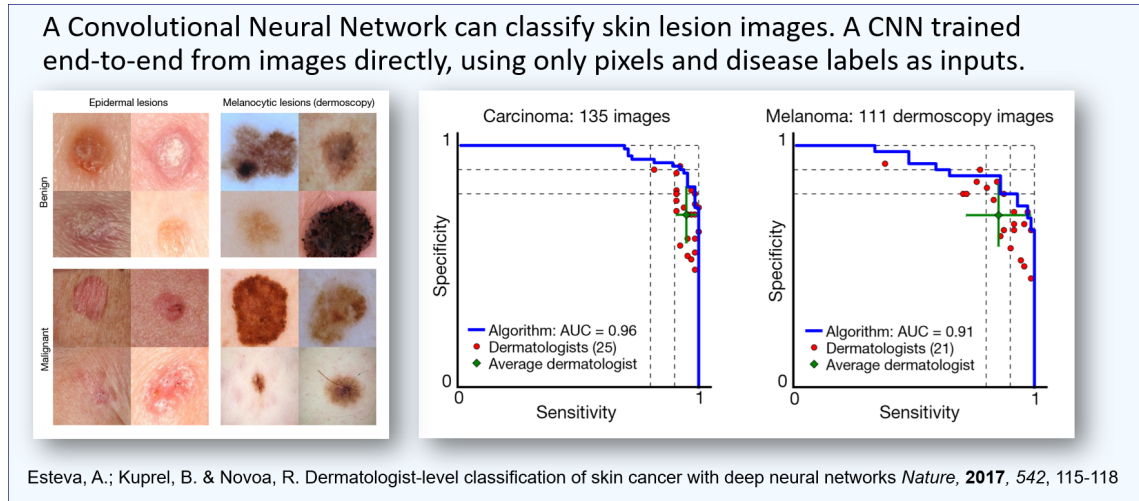


Figure 3.2: Machine learning achievement in dermatology. Esteva et al demonstrated a CNN could learn from images of histologically confirmed lesions and could then produce results in tests that compare favourably with trained dermatologists.

We also detailed some of the challenges faced by advanced techniques such as adversarial attacks on deep learning computer vision where unexpected results have alarmed developers and prospective users alike. Another workshop slide (Fig. 3.3) shows

### 3. Design

how an adversarial attack had been tuned to produce alarmingly poor results from a deep neural network.

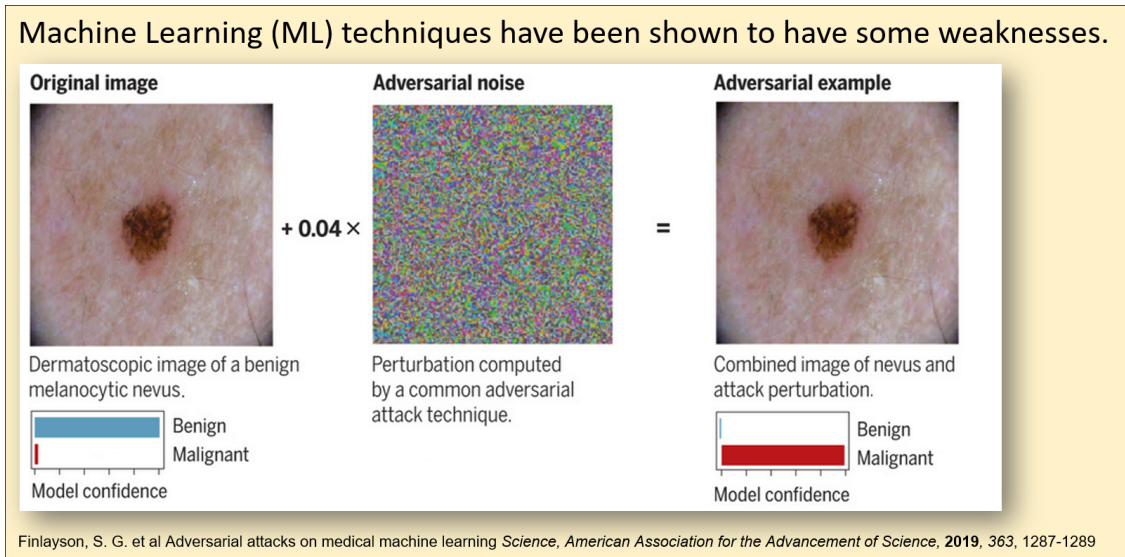


Figure 3.3: Machine learning challenges in dermatology. The addition of a small fraction of apparent noise makes a barely detectable difference to human eyes, but yet fools the AI system into confidently mis-classifying the lesion.

#### 3.1.4 The context of clinical decision-making

The decision-space in medicine is enormous. Quite apart from medicine being a complex and extensive subject in itself, there are many different clinical roles and responsibilities and each differs further by organisation. Then each decision-maker faces a different presentation in a different part of the patient pathway Fig. 3.1. And of course each case is different owing to the different nature, needs, expectations and comorbidities of human patients. Part of the preparation for the workshop was to try and find a way to navigate this space meaningfully and provide useful questions to make clear what was needed. The resultant workshop visuals and questions specifically guided participants to consider both the *context* of trust and what we termed the *ingredients* of trust .

#### 3.1.5 Shaping the focus with patient stories

The workshop aimed to involve all participants in narrowing down the possible space to the most effective context and variables. This was achieved in an iterative process with a sequence of break out discussions in smaller groups tasked with choosing from the

stimulus material or from their own ideas and completing a prepared proforma, then returning to summarise in a plenary session before repeating the process. The subject areas covered were as shown in Fig. 3.4

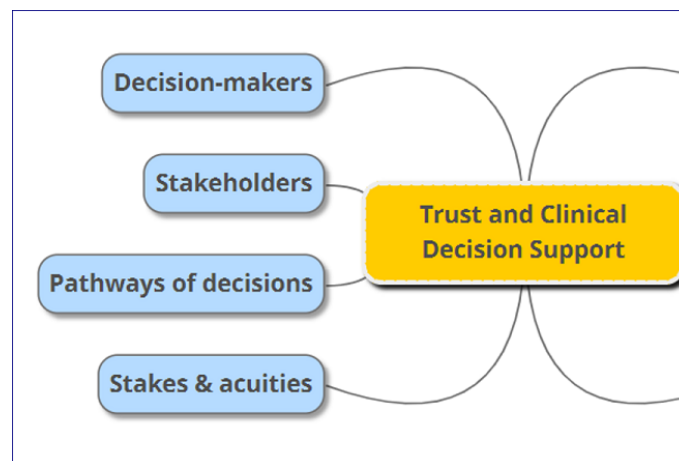


Figure 3.4: Our *contexts* of trust. The elements shown on the left here were the high-level stimulus points for the detailed workshop discussions about which patient stories should be presented to study participants.

### 3.1.6 Shaping the focus with system characteristics

The four system characteristics that were presented to the workshop (the *ingredients* of trust in Fig. 3.5) had been derived from a review of the literature connecting AI systems and clinical decision support. They were constituted to support exploration and discussion of the technical details lying behind any intelligent decision system.

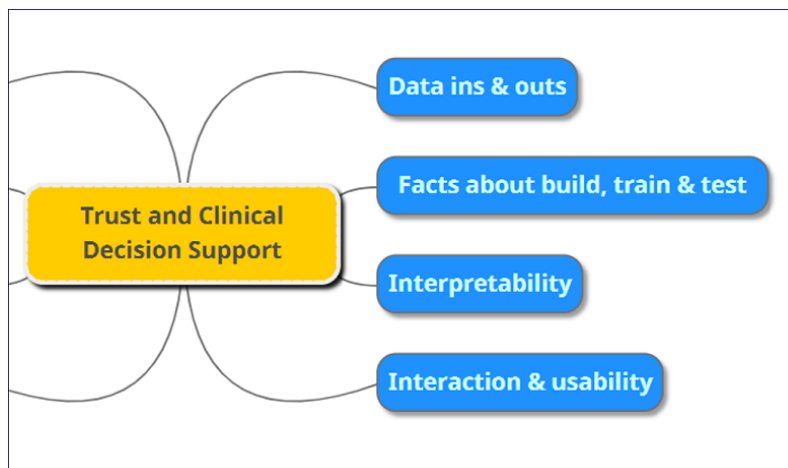


Figure 3.5: Our *ingredients* of trust. The elements shown on the right here were the high-level stimulus points for the detailed workshop discussions about which AI system characteristics should be presented to study participants.

During the workshop itself, these were re-shaped and augmented to produce six characteristics. Two were added - a confidence level and an approved development process. These characteristics were thought by the clinicians to be highly desirable in the case of the confidence level and an interesting discriminator in the case of the approved process. And one was separated into two components - splitting the development process that had been labelled as 'Facts about build, train & test' into two separate components - architecture and test performance. Note that the size and quality of the train and test datasets were already included as a separate characteristic.

Training data	Blue	– provides information on the ground truth data used in training the ML tool
Performance	Green	– provides information on the on the performance achieved in testing the ML
Approved	Indigo	– provides information about approval by a clinically competent body
Result only	Orange	– provides no information about the ML tool other than the result itself
Technical detail	Red	– provides information on the technical development of the ML tool
Explanation	Violet	– provides an explanation for the result obtained by the ML tool
Quantified with CIs	Yellow	– provides quantified results with confidence intervals

Figure 3.6: The seven AI system characteristics emerging from the post-workshop discussions. These are near their final form. The colour names would be randomly assigned per user to reduce the effect of bias.



At the end of three hours, the discussions still held plenty of interest for the participants and we left with significant notes and a recording of all the discussions. These notes and recordings were written up and circulated with further visuals to promote critical review and aid further refinement. The visuals are shown in Fig. 3.7 and Fig. 3.8.

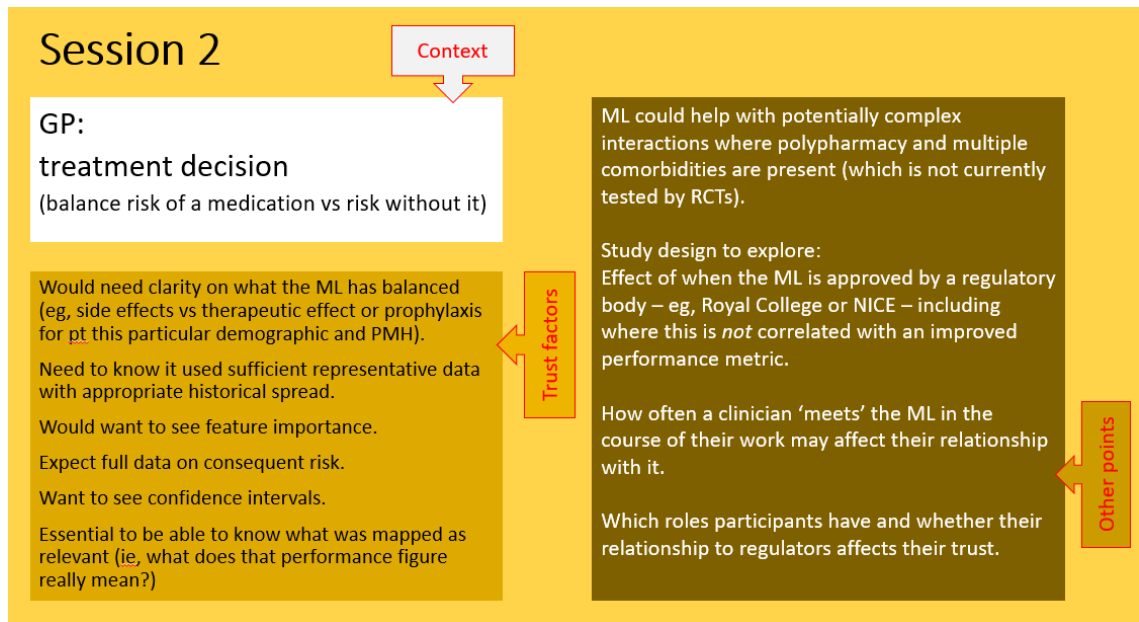


Figure 3.7: Workshop results - direct output. Participants completed a proforma with the results of the structured activities in order to shape the elements of context, trust factors and additional points.

At this point we had three embryonic patient stories, each with a proposed decision and some detail on the associated decision factors and trust-related characteristics. And we had several other (fragmentary) patient stories in addition.

This was a long way from having a fully designed online user study. But critically, the process so far had carved from the huge theoretical space of clinical decision-making a subset of decision contexts that were both meaningful and congruent for a diverse group of practising clinicians.

The next challenge would be to maintain the collaborative momentum while getting into the detail of each context with the appropriate scene-setting description and the right balance of clarity and uncertainty for a meaningful decision-space.

## Example study component – anticoag in AF

*Participants' initial level of trust in an ML tool to support decision on anticoagulation treatment in AF will be assessed along with their level of trust in CHADS-VASc/HAS-BLED which the tool extends.*

*Participants will be exposed to multiple design patterns in which characteristics will be varied. Possible variations could be accuracy score, size and relevance of additional training data, inclusion of feature importance, or presence/absence of a regulatory approval mark.*

*Participants' post-hoc level of trust in the ML tool will be assessed along with their level of trust in CHADS-VASc/HAS-BLED.*

Figure 3.8: Workshop results - derived output. From the completed proformas, a few sample study components were drafted. The anticoagulation in atrial fibrillation (AF) patient story is shown here.

### 3.1.7 Individual collaborative sessions

Beyond the workshop, we continued the collaborative design process by means of telephone, email and zoom discussions.

It soon became clear that there was a convenient match between certain members of the group and certain of the patient stories we had developed. And the other group members had more general strengths that were complementary in terms of task breakdown. This made for a simple arrangement for continued collaboration. Each session with a clinical group member would be a *'think-aloud'* exploration of either one of the stories or one of the general design features. In this way, we were able to make use of all group members without convening the group into one common session.

An early change during these small *'think-aloud'* sessions with individual clinicians was the development of an acutely unwell diagnosis scenario from a workshop fragment. And soon after this came the abandonment of a patient story involving an acute infection. Challenges with the way this story would work in practice and how it would be influenced by local policies rather than an independently trained AI system meant that it was deemed less straightforward to work up as a scenario.

Prior to each *'think-aloud'* session, additional research work had been carried out on the clinical topics used for the patient stories. The presenting cases were drawn up using this research and the associated AI system responses crafted to match. This was an important preparation step as the clinical advisors had strictly limited time in which to provide ongoing input. During this phase, the clinicians emphasised that while the scenarios were clearly artificial, there was a need for the clinical content to be sufficiently realistic to not be a distraction from the study questions. And the decision itself had to be finely balanced in order for recourse to decision support to be motivated. At the same time, we needed to keep the clinical detail sufficiently general so that the clinical decision factors would be accessible to the broadest range of participants.

While they were close, there were several points of detail in the stories, as prepared, that were identified as points of improvement. For example, the atrial fibrillation (AF) patient story needed to have a small detail inserted into the patient's social history in order to add sufficient doubt about the recent episodes being a natural evolution into persistent AF (in which case the decision would be too obvious and would not be a natural candidate for any AI support).

With these two kinds of work-sessions, we had been able to consolidate and extend what could be achieved with a workshop alone. We had, through a process of clinical co-creation, resolved the details of three useful and well-formed patient stories along with the ways in which we would like our selected hypothetical AI characteristics to show their different forms of decision support.

## **3.2 Participatory design - how do we explore it?**

### **3.2.1 Study design**

Parallel to the work on crafting patient stories and selecting appropriate AI system characteristics, we explored the detailed aspects of the study design. How were we to gather the information on trust that participants could provide? Each user must be exposed to a range of hypothetical AI systems that provided decision support along with their distinctive system characteristics. And this exposure needed to be in a suitable decision-making context (the patient story). We of course needed participants to evaluate the experience of being given the decision support by each distinct AI system. This process

had to be repeated in such a way as to keep users engaged while data were captured across a range of AI system characteristics to compare the evaluations.

We initially considered a before-and-after measure of trust. But it soon became clear that any attempt to evaluate trust without context of a patient story (a before-exposure measure) would be challenging. There is no useful sense of trust in the abstract that we could assess that could be directly compared to an assessment of trust within a patient story context and some concrete AI system characteristics. So instead we included among the AI systems an unadorned system that provided no information supporting its recommendation - just the recommendation itself. This served to provide a control against those systems with characteristics we wanted to test. We called this control the *vanilla* system. And rather than place this *vanilla* control as the first system exposed to participants, we decided to include it in a randomised sequence of exposures.

We recognised that, as long as users were engaged, using online forms to capture users' responses would permit efficient data recording across all participants. But having unsupervised users would present distinct challenges. It would inevitably place a premium on the need for an intuitive workflow, a relatively friction-free navigation and an engaging experience. Testing a clinician's patience with a jarring or confusing workflow would likely result in a high proportion of incomplete responses and a high drop-out rate.

#### 3.2.2 Likert scale questions

To capture quantitative data on the evaluations, we used Likert response sets that could be easily rendered on an interactive web page. The prompts for the questions are shown in Table 5.3. These question prompts were adapted from those used by Ashoori et al [145] - which are in turn adaptations of Madsen's & Gregor's scales for measuring human-computer trust [150]. Further following Robert Hoffman's approach [151], we paid attention to *face-validity* and *construct-validity* for each of these. The combination of questions forming an evaluation set was chosen to ensure a balance between sufficiency and richness of data and a reasonable experience for the participant.

**Study questions – thinking about the red system, please answer the following**

1. Trustworthiness \*  
 This decision-making process is trustworthy  
 Strongly disagree   Disagree   Agree   Strongly agree

2. Change the process \*  
 I would change one or more aspects of this decision-making process to make it trustworthy  
 Strongly disagree   Disagree   Agree   Strongly agree

3. Contextual use \*  
 The use of machine learning is appropriate in this scenario  
 Strongly disagree   Disagree   Agree   Strongly agree

4. Technical implementation \*  
 I trust that the technical implementation of the machine learning model is correct  
 Strongly disagree   Disagree   Agree   Strongly agree

5. Personally confident \*  
 I am confident in this decision-making process. I feel that it works well.  
 Strongly disagree   Disagree   Agree   Strongly agree

6. Personally wary \*  
 I am wary of this decision-making process  
 Strongly disagree   Disagree   Agree   Strongly agree

7. Personally content \*  
 I like this decision-making process  
 Strongly disagree   Disagree   Agree   Strongly agree

**Red system**   **Finish**

Figure 3.9: Wireframe design - Likert scale questions. Each AI system with a different characteristic (and different colour) would be followed up with a set of evaluation questions.

In anticipation of the data analysis (Chapter 5), we needed to ensure the instrument was measuring a variable that is intrinsically continuous. That is, while we provided four discrete response options, there is no discrete character to the notion of *'degree of trust'* itself. And we established that the interval character of the separation between Likert anchors was uniform. That is, the difference between *'Agree'* and *'Disagree'* is an equivalent difference to that between *'Agree'* and *'Strongly agree'*. Our theoretical attachment to this coding treatment is based on the exclusion of a neutral option (*'Neither agree nor disagree'*). While having no neutral option risks user frustration, it removes the risk of users *'hedging'* and requires them to indicate a non-zero value. A proportion of the responses thus recorded must therefore be considered as being close to zero.

These design decisions allow us to consider the coded responses as interval data and use rank tests to compare the evaluations of different AI system characteristics [152]. We should note here that there are several conflicting schools of thought in the literature on the mathematical analysis of Likert scale data [153, 154, 155, 156, 157]. Our approach is to assert the consistency of the intrinsically continuous variable with the mathematical tests employed on the discretised data.

To complement (and further cross-validate) the structured responses, we added an optional general free text question to each evaluation.

In addition to the individual system evaluations, we wanted an overall indication of a user's pre-disposition towards or away from using AI support. This pre-existing value-set held by a user, if strong and variable across the participants, might be influential on the results. So we introduced a preliminary set of ten questions to capture this.

In recognition of the significance of priming [158], we needed to ensure that, for each user, the preliminary questions were answered before any substantive evaluations were carried out. This ensured that there was an equivalence to the priming experienced by each participant and reduced the possibility of skew in the results for these ten questions.

#### 3.2.3 Wireframe and iterations

Alongside gathering further input from the advisory group in the form of the individual *think-aloud* sessions, we were able to draw up a wireframe of the user interaction. We used presentation software to make prototypes with high visual fidelity. Because these necessarily had low functional fidelity, we also made a video walk-through to ensure the intended user-experience was communicated and to facilitate orientation. This extra step was necessary because organising a second workshop would have been too demanding against clinical commitments. The video presentation was cloud hosted and linked to a copy of the underlying slide deck so all members of the advisory group could also explore the content for themselves.

This part of the design refinement was aimed at checking the bulk of the end-to-end user-experience - with attention paid to the shifts required in user focus as well as the sustained attention and levels of abstraction expected of users.

To allow study participants to navigate the elements of the experimental space efficiently, we wanted to anchor the different AI system characteristics using colour and graphics. This was felt important since the number of navigation steps for a participant was quite large and we wanted to ease the cognitive burden on users by leveraging pre-attentive perception of colour for categorical discrimination [159]. Any navigational step had to provide strong non-textual cues as to where the user was in their mental map of the process [160]. In addition to navigational ease-of-use, we were obviously keen to ensure that each evaluation was actually completed on the correct AI system experience.

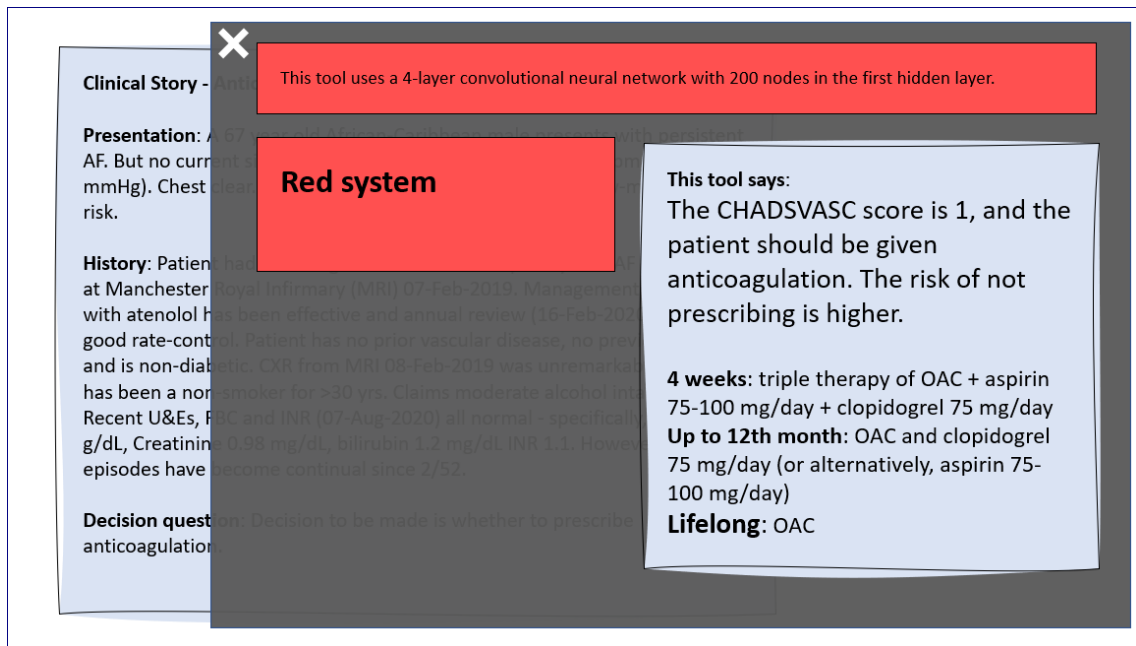


Figure 3.10: Wireframe design - Hypothetical AI system response. Each AI system (colour) would provide a response to the patient story that constituted decision support. The user would then be asked to evaluate that decision support.

The colour-anchoring allowed a separate web form for evaluation to be on a linked but visually-related page without weakening the association. We didn't need to provide a similar process for the different patient stories for two reasons. First, our target audience (clinicians) were familiar with navigating around, and returning to, an electronic patient record. There was no corresponding familiarity with moving between different AI systems. So it was the AI system anchoring we prioritised. Second, we had decided to promote with participants the idea that a 'session' would involve completing a full complement of evaluations (all AI system types) on a given patient story (a single decision context). So most users would only experience a change of patient with a subsequent 'session', meaning that, unlike the distinction between AI systems, the distinction between different patients did not have to be emphasised.

In the *think-aloud* walk through of the wireframe prototype, we were able to pick up on additional design improvements - such as the need to easily return to both the patient story detail and the AI system characteristic detail.

### 3.2.4 Designing for unbiased and uniform data

The use of colour to anchor the different AI system types was useful, but it needed to be bold to be effective. With this in mind, we felt there was a risk it might bias the results owing to users evaluation responses being influenced by the colour more strongly than the characteristic described. For this reason, we decided we would want to randomise the association of colours and AI systems for each user in the final design. As long as they were logged in, a particular user would see the system with high performance as a green system, for example. And this would be consistent across multiple sessions for that user. But for a different user, the high-performing system might be violet and so on.

The same consideration was given to the order of presentation, both of the patient stories and of the AI system types. This randomisation of the patient stories and AI system types would ensure that our response space would fill with a uniform distribution of the various combinations of stories and AI types. This was an important consideration when thinking ahead to the analysis.

**Clinical Story - Anticoagulation in AF**

**Presentation:** A 67 year old African-Caribbean male presents with persistent AF. But no current signs of haemodynamic instability (HR 78bpm, SBP 122 mmHg). Chest clear. No hyperthyroidism. CHADS-VASc 1, “low-moderate” risk.

**History:** Patient had ECG-diagnosed non-valvular paroxysmal AF documented at Manchester Royal Infirmary (MRI) 07-Feb-2019. Management since then with atenolol has been effective and annual review (16-Feb-2020) showed good rate-control. Patient has no prior vascular disease, no previous stroke and is non-diabetic. CXR from MRI 08-Feb-2019 was unremarkable. Patient has been a non-smoker for >30 yrs. Claims moderate alcohol intake. Recent U&Es, FBC and INR (07-Aug-2020) all normal - specifically, Hb 15.5 g/dL, Creatinine 0.98 mg/dL, bilirubin 1.2 mg/dL INR 1.1. However, AF episodes have become continual since 2/52.

**Decision question:** Decision to be made is whether to prescribe anticoagulation.

Put yourself in the position of reviewing the patient story shown here.

There are seven different machine learning tools to review. Each will provide a recommended decision response when you open it.

You can choose whether to make your own judgement before reviewing that of the tool or afterwards.

Red	Orange
Yellow	Green
Blue	Indigo
Violet	Rank

Figure 3.11: Wireframe design - Patient story. Each clinical decision point would be represented with a patient story. The user then chose an AI system (a colour) to evaluate in the context of that story.

So the effect of these design ideas was that for one user, the first patient story may have been with atrial fibrillation and the first AI system a state-of-the-art convolutional neural



network presented in yellow. But for another user, the first patient story could have been with a pre-diabetic patient and their first system one approved by a responsible body and presented in blue. We knew this would present its own difficulties in testing and in discussion between project advisors. But the benefit for the experiment would make it essential.

### **3.2.5 Detailed design considerations**

Visual considerations were judged to be important for this study given a) the quantity of information we were asking participants to process (with a large amount being repeated in later sessions) - and b) the different levels of attention we were asking participants to give. While we knew the clinical advisors were becoming comfortable with the proposed study experience, we were also aware that users experiencing the material for the first time would need to be supported by a carefully crafted user experience.

We needed participants to pay attention to patient stories for context at one level, to the characteristics of AI systems at another level - and to complete substantive evaluations that required thinking about trust - a further level of abstraction). For this reason, we took care to ensure that the different levels of information were indicated with consistent visual cues. In addition to the bold colours, we added graphical elements to communicate and reinforce the distinct AI system characteristics.

Another visual support step we added was to include some image elements in the AI system descriptions where this helped both describe the characteristic and anchor the particular system in the user's memory as well as further distinguish it visually from the others.

With these features incorporated, we knew we had increased the prospects of engaging the right elements of clinical users' thought processes - focusing on the variable of interest when it came to the actual user study.



## Chapter 4

# Implementation

### 4.1 Advance recruitment

Anticipating the need to move fast once the user study was live, we began recruitment ahead of deployment. We asked contacts and networks to promote sign up for the study during late July and early August. We had created a set of static pages hosted on the computer science department network which allowed embedded deployment of a MS forms function. This meant the branding and introductory material was available at sign-up and consistent with the final deployed site.

### 4.2 Platform - practicalities under covid restrictions

Given the restrictions on travel and face-to-face meetings during development and data collection (Jun-Sep 2020), the technology had to allow for unsupervised remote sessions by participants. So it was essential the platform included user registration with the capture of some participant information and allow return sessions with the minimum of friction for participants. The study design meant that we had to be able to modify the presentation per user and store the data they submitted. Hence we needed a modern web development framework to ensure this functionality was provisioned safely and securely.

### 4.3 Technology stack

We aimed to be on the university network, so we created the skeletal site (using php) in June with development carried out on a local LAMP stack. Realisation that user

authentication, data capture and storage required a virtual server prompted a switch to developing a remote virtual machine (and a lot of new learning in a sys admin role for the researcher). We tried various OS servers including CentOS 8 and Ubuntu 20 owing to the expected need for service desk support (favouring CentOS) but also familiarity, community resources and ease of use (favouring Ubuntu). Much work went into build a local development workstation to allow development on linux and avoid the perils of deploying to a different OS (the researcher's personal workstation is windows 10). A deal of research and new learning was experienced with firewalls, server setup, key-based authentication and remote command line server management. A VPS was procured on Digital Ocean (DO) for use as a UAT environment as the department network server wasn't available in time. Eventually we deployed the production version to DO and opened that to users when the department network VM couldn't be made to work.

### 4.4 Web framework

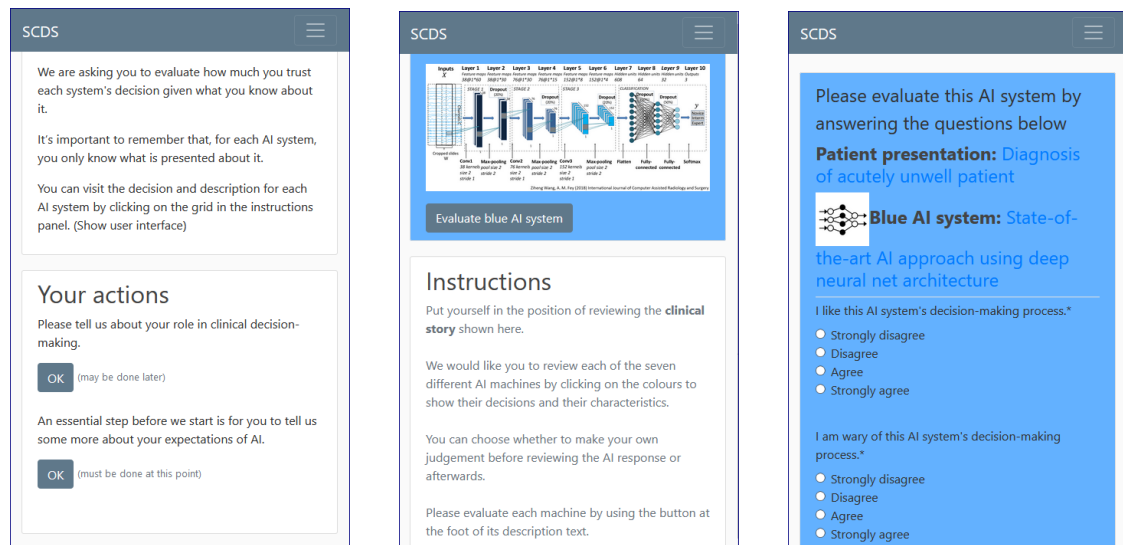
The web framework used was django 3.1.1. This is written in python and features a flexible object relational mapper (ORM) and a suite of security features out of the box. It also provides a site administration module. So a lot of 'boiler plate' web development is taken care of. As a framework, it also allows good separation of content, business logic and presentation. And it is an extremely popular framework, so there is a significant online community which is helpful when creating experimental websites with multiple custom features.

### 4.5 Page content combining patient stories and AI systems

At the time the site was being developed, the total quantity of pages wasn't fixed. We knew we needed to accommodate a number of patient stories (up to five) combined with AI systems (up to seven). This could mean 35 distinct pages with much repetition. So a suitable data structure was needed. We used a dedicated json-like structure that was accessible using django/python modules. We used the early page content to design the user workflow and ensured that it was extensible. Django allows use of parameters to be passed in url slugs and accessed as variables on the page template. So we were able to drive the user experience (UX) from a core data-driven page of user stories whose content was determined by the participant's user id.

## 4.6 Randomising the UX

As described in Chapter 3, we needed to randomise the patient stories and the presentation of AI systems. Interestingly this was a frequent point of feedback from discussions with trialling users. They were concerned (until they were reassured) that the colours used or the order of presentation could influence the responses. The randomisation was achieved by using the registered user id as a seed for pseudo-random functions. The simplest approach was to create a matrix of all possible combinations (patient stories and AI systems) and randomise the sequence (presentation order) per user. A simple query of the stored data could get the position in the user's overall progress by a count of steps completed. This allowed a session variable to store what the next story-system combination should be and pass this to the ORM when the user requested the next step. This simple design approach allows new stories and new systems to be added without re-coding.



a) Approaching preliminary questions.

b) Approaching evaluation.

c) Evaluation questions.

Figure 4.1: Screen shots of the mobile design showing the approach to preliminary questions, the approach to evaluating an AI system and the evaluation screen.

## 4.7 Testing

We asked ten users to trial the web application, including trials on android and apple mobile devices (Fig. 4.1). This led to improvements in the navigation, styling, consenting

and instruction elements. Testing also revealed that there were a lot of steps being asked of the user prior to the substantive trust responses being recorded. We looked at trying to reduce the click count so that completion of one AI system evaluation would land the user immediately on the next. But the technical work to implement this would have delayed deployment and was judged to be less urgent than getting to go-live. As mentioned in Chapter 3, we did decide that it was important for data consistency to require each user to answer the preliminary questions *before* beginning any evaluation of different AI systems. These preliminary questions are aimed at establishing any predisposition towards or against AI. We also altered the sequence of required steps in recognition of the need to make the induction as frictionless as possible (registration and consent were critical requirements, with nice-to-have data submissible by the user later).

### 4.8 A late change to participation

Participation criteria were touched on in the workshop but not formulated in detail. An early concept had been to recruit a broad range of clinicians with an expectation that the core users (as with any real deployment of AI decision support) would be those with clinical governance responsibility - ie, medics in most cases. However, there was a strong feeling among some of the clinical advisory group that it should not be confined to doctors alone. A wide range of advanced nursing staff and allied health professionals are required to make clinical decisions in their work and these would be useful to recruit. In light of this, we needed to record some personal detail about each participant's role. We kept this to a minimum in order to fulfil the requirement of making the experience of sign-up and induction as frictionless as possible.

On the day the online study element launched, there was a strong argument from a clinical advisor that some expert patients would find it appropriate to have access as participants. As a result, we allowed a late change to the participation criteria that included non-clinicians.

## 4.8. A late change to participation

The screenshot shows a web interface for 'SCDS' with navigation links 'Home Studies About' and a user profile 'avo Logout'. The main content is divided into two columns. The left column contains a 'Clinical story' titled 'diagnosis of the acutely unwell patient'. It includes a 'Presentation' section with a detailed medical history, a 'History' section, and a 'Decision question' section. Below this is an 'AI characteristic' section for the 'Blue AI system', which describes it as a state-of-the-art approach using deep neural net architecture. The right column contains 'Instructions' for the user, explaining that they should review the clinical story and evaluate seven different AI machines by clicking on their respective colored buttons. A vertical stack of seven colored buttons is shown: blue AI system, orange AI system, green AI system, violet AI system, yellow AI system, red AI system, and indigo AI system.

a) Approaching evaluation.

The screenshot shows the evaluation screen for the 'Blue AI system'. It features a blue background and a white text area. The title is 'Please evaluate this AI system by answering the questions below'. Below this is a 'Patient presentation' section with the text 'Diagnosis of acutely unwell patient' and a 'Blue AI system' section with the text 'State-of-the-art AI approach using deep neural net architecture'. There are five Likert scale questions, each with four radio button options: 'Strongly disagree', 'Disagree', 'Agree', and 'Strongly agree'. The questions are: 'I like this AI system's decision-making process.\*', 'I am wary of this AI system's decision-making process.\*', 'I am confident in this AI system's decision-making process. I feel that it works well.\*', 'I trust that the technical implementation of the machine learning model is correct in this AI system.\*', and 'The use of machine learning and AI in this scenario is appropriate.\*'. At the bottom, there is a text input field for 'Please add any comments on the subject of trust with this AI system.' and a 'submit' button.

b) Evaluation questions.

Figure 4.2: Screen shots of the desktop design showing the approach to evaluating an AI system and the evaluation screen.

## 4.9 Deployment

An introductory video for participants was created to ensure a smooth orientation and induction. The study went live on 09-Sep with a mail out to the previously logged subscribers and a series of promotions on clinical social networks.

Post-deployment on the launch day, we noticed an alteration to the style rendering that had arisen from the inclusion of images in the AI system descriptions. We also found two typographic errors in the text files that would have compromised the results. Both of these were fixed, tested and an update deployed before any data submissions had taken place.

The result of the implementation work was a mobile-friendly and visually consistent and easily navigable online experience that ensured unsupervised users would complete the preliminary steps before the substantive online study. The study involved them being exposed to a random selection of patient stories and different AI system characteristics as well as being asked to evaluate each exposure.

Screen shots of the steps near evaluation for one of the systems (*CNN* type rendered in blue under the acutely unwell patient story) are shown in Fig. 4.2.



## Chapter 5

# Results and analysis

This chapter reviews the nature of the data and a summary profile of the data collected in the substantive online user study. It considers the preliminary question responses, including a potentially interesting incidental feature in relation to user-type. The bulk of the chapter is devoted, naturally, to the evaluation by users of the AI system characteristics. This section looks at the core data collected by the online user study. A few additional observations on response by user-type, qualitative responses and the range of participants round out the chapter.

### 5.1 Nature of data collected

Aligned with the focus of our study - evaluating the trustworthiness associated with different AI system characteristics - the main evaluation questions for each AI system are captured on a sequence of five Likert-scales. This is followed by an optional free-text question allowing observations and commentary by the participant.

Hence the structured data generated by the user-submitted evaluations are Likert-scale values. The design of the Likert scales is detailed in Section 3.2.3. While storage of the raw data uses the string labels ('Disagree', 'Strongly agree' etc), these are readily coded into directed numeric values as shown in Fig. 5.1.

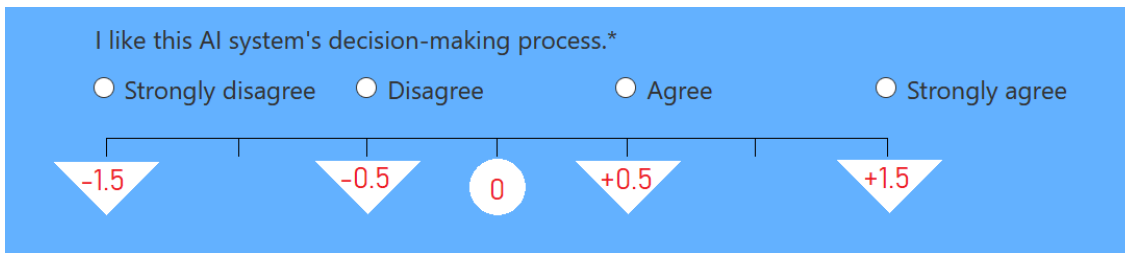


Figure 5.1: Likert coding showing directed scale with a zero neutral point between the inner-most options.

Each coded evaluation-set consists of five such values in response to five distinct trust-related questions on one presented AI system. Our design allowed for a session to consist of seven evaluation-sets - one for each of the seven AI system characteristics presented and all on the same clinical patient scenario (which we called a patient story). Each participant was able to submit up to three sessions (three patient story contexts). So a single participant could generate up to 105 (5 questions x 7 AI systems x 3 patient stories) structured values out of a response space of 420 possible values (4 Likert options x 105 questions)

For each user, we also recorded a preliminary assessment of their awareness of, and predisposition to trust, AI in general (outside of a clinical setting). The website design required this to be completed prior to the substantive study questions in order to avoid any bias resulting from differences in priming [158]. The preliminary questionnaire also used a combination of closed and open questions. The first ten of these preliminary questions probed users' attitudes to questions about how they judged the effect of AI on themselves and on society. The preliminary ten Likert-scale questions each have four response options. This creates a submission per-user of ten structured data values out of a response space of 40 possible values per user. A further optional free-text question allowed each user to elaborate on any or all of the points raised or ideas prompted.

## 5.2 Summary of data collected

The study captured 292 evaluations from 31 participants over the period 09 to 25-Sep-2020 (17 days). The level of participant engagement was extremely good. Seven individuals (over 22% of the participants) provided the maximum possible three full sessions (each submitting a total of 105 responses in 21 evaluations). The submission levels are summarised in Table 5.1

Table 5.1: Participation counts.

submissions each	number of users	total submissions
21	7	147
14	2	28
7	5	105
<7	7	12

The breakdown of participants by role is of relevance when considering the late inclusion of non-clinicians and exploring the match of design to execution. This is shown in Table 5.2. Of the 292 submissions, 234 (80%) are by clinicians (of which 176 or 60% of the total are by doctors). Meanwhile, There are 58 (20%) by non-clinicians.

Table 5.2: Participation counts by role.

role	number of users	submissions	% of total
clinicians	25	234	80
(of which) doctors	19	176	60
non-clinicians	6	58	20
total	31	292	100

For the different AI system characteristics, the number of evaluations is shown in the count column of Fig. 5.2

To describe the evaluations of the AI system characteristics, we need to consider the response space. To do this, we introduce the tabulation shown in Fig. 5.2 where each row contains data on an AI system characteristic and each column contains data on a Likert-scaled question.

AI type	count	Like	Wary <sup>1</sup>	Conf	Tech	Appr	Median
Appv	42	0.5	-0.5	0.0	0.5	0.5	0.5
CNN	43	0.5	-0.5	0.5	0.5	0.5	0.5
Data	42	0.5	0.5	0.5	0.5	0.5	0.5
Expl	41	0.5	0.5	0.5	0.5	0.5	0.5
Nila	42	-0.5	-0.5	-0.5	-0.5	0.5	-0.5
Perf	42	0.0	-0.5	0.0	0.5	0.5	0.0
Risk	40	0.5	0.5	0.5	0.5	0.5	0.5
	Median	0.5	-0.5	0.5	0.5	0.5	

1 - reverse coded

Figure 5.2: Median evaluation scores for each AI system characteristic and likert question.

## 5. Results and analysis

---

This figure summarises all data in the response space with a median value for each combination of AI characteristic and question. The data are further aggregated at the row and column ends with the median of the row or column respectively. The randomisation mechanism ensures that the distribution of counts per AI system characteristic is close to uniform and approximates uniformity more closely the more participants are involved. This uniformity can be seen in the individual counts (per AI type) being very close to each other. They are all within 3.6% of their mid-range value (41.5). The ordering of rows in this table is a default (alphabetical by AI characteristic label). This is useful when comparing different presentations of results later. But ordering the systems differently is also useful, as we shall see.

The label codes used in our response space tabulation Fig. 5.2 are explained in Table 5.3 and Table 5.4. The label codes used for the three patient stories are explained in Table 5.5.

Table 5.3: Labels used for questions in the response space.

Question label	Likert-scaled question
Like	I like this AI system's decision-making process.
Wary	I am wary of this AI system's decision-making process.
Conf	I am confident in this AI system's decision-making process. I feel that it works well.
Tech	I trust that the technical implementation of the machine learning model is correct in this AI system.
Appr	The use of machine learning and AI in this scenario is appropriate.

Table 5.4: Labels used for AI system characteristics in the response space.

AI label	AI system characteristic
Appv	Approved by Academy of Medical Royal Colleges
CNN	State-of-the-art AI approach using deep neural net architecture
Data	Large, representative, diverse and robustly labelled training dataset
Expl	Provides some explanation
Nila	An AI system
Perf	Demonstrates high performance in tests
Risk	Provides relative risk values

Table 5.5: Labels used for patient stories in the response space.

Story label	Patient story (clinical presentation)
AF	Atrial Fibrillation (whether to anticoagulate)
AU	Acutely unwell patient (diagnosis)
DM	(Diabetes Mellitus) Pre-diabetic patient (whether to prescribe metformin)

### 5.3 User predisposition

Before exploring the substantive evaluations of trust, we review the responses to preliminary questions that each user was asked to provide. These questions (Table 5.7) were designed to assess the prior disposition of each user towards AI systems in general. And although not the primary focus of our analysis, they do provide the context of the substantive data collection.

Table 5.6: Preliminary questions completed by all users prior to exposure to the study scenarios.

Question label	Likert-scaled question
no benefit	AI does not personally benefit me.
saves time	AI saves me time.
wk less int	AI will make some of my work less interesting.
less hum err	AI reduces mistakes made by humans.
wk more int	AI will make some of my work more interesting
imprv servs	AI improves the services I receive outside of my work life.
und hum aut	AI is likely to undermine human autonomy.
more prod	AI makes people more productive.
lives worse	AI technologies are making human lives worse.
lives better	AI technologies are making human lives better.

Table 5.7: Evaluation questions completed for each AI system under a given patient story.

Likert-scaled question
<ul style="list-style-type: none"> <li>• I like this system’s decision-making process.</li> <li>• I am wary of this system’s decision-making process.</li> <li>• I am confident in this system’s decision-making process. I think it works well.</li> <li>• I trust that the technical implementation of the machine learning model is correct in this AI system.</li> <li>• The use of machine learning and AI in this scenario is appropriate.</li> </ul>

## 5. Results and analysis

---

From the plotted data (Fig. 5.3), we can see that none of the median values lie outside of the central pair. So no strong feelings on aggregate were recorded. The plots suggest that there is a tendency among participants to agree with the propositions: *'AI reduces mistakes made by humans.'* and *'AI improves the services I receive outside of my work life.'*

Meanwhile, there appears to be a tendency to disagree with the proposition: *'AI technologies are making human lives worse.'*

We note that this puts our users collectively (though not individually) in the camp of those not completely hostile to AI. We also note that there are instances of strong feelings in both directions on every question other than that AI reduces mistakes made by humans.

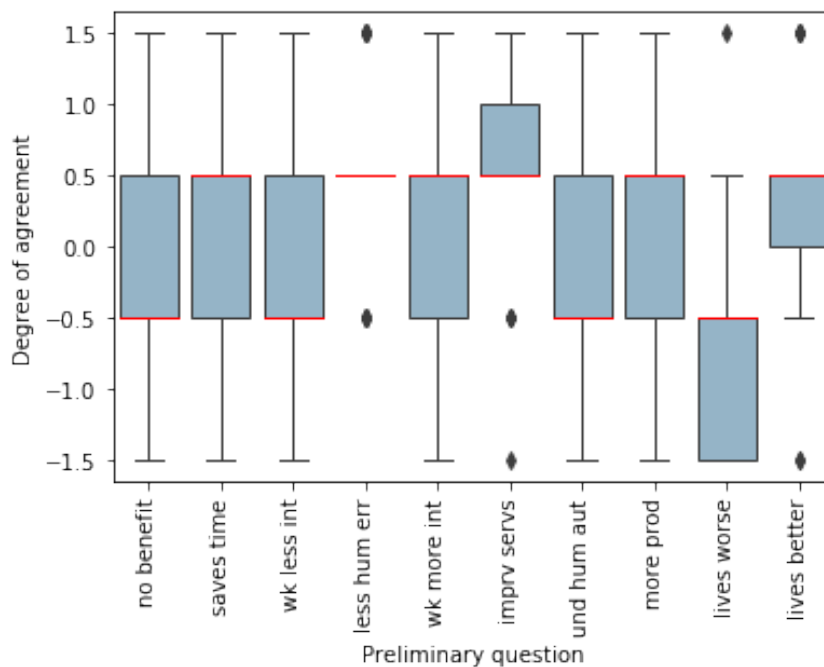


Figure 5.3: Box plots of preliminary question responses. All users.

Although not an explicit design feature, the ability to split participant responses between those who are trained clinicians and those who aren't may produce some potentially interesting results later. So we take the opportunity here to quickly review the preliminary questions by role (user-type) shown in Fig. 5.4.

Here it can be seen that the medians of the responses to the first proposition (*'AI does not personally benefit me'*) lie at different values. The difference is not large, however. We

also note that some of the inter-quartile ranges (IQRs) do not lie adjacent to each other. Overall, the plots for the two user groups align quite closely.

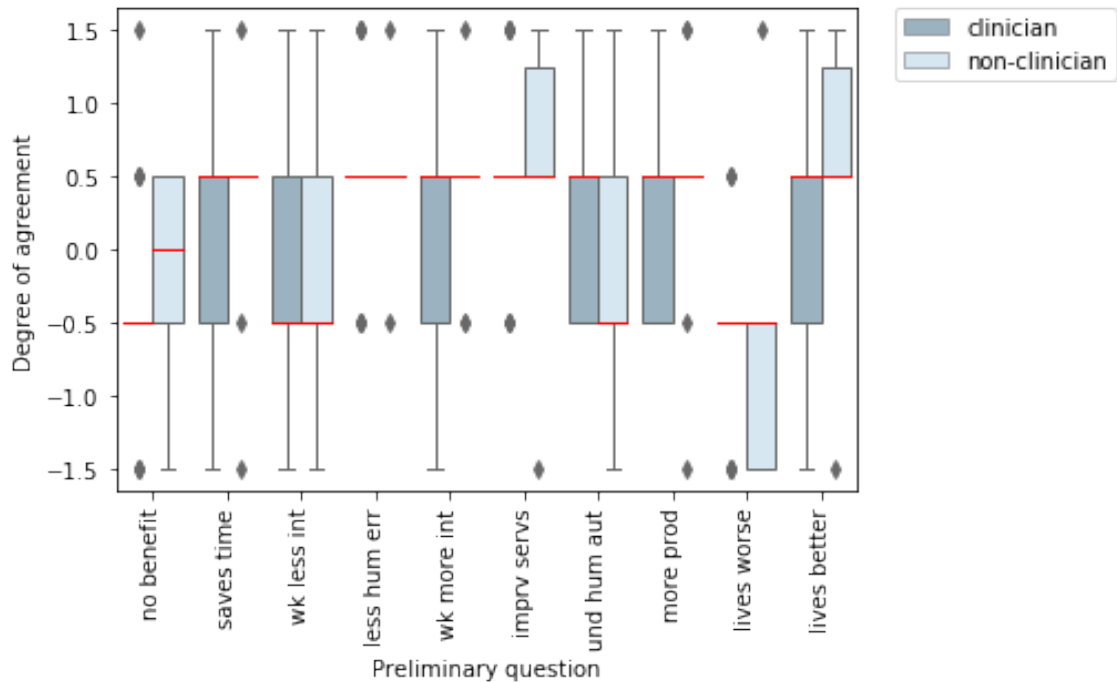


Figure 5.4: Box plots of preliminary question responses by user-type.

## 5.4 Evaluation of the AI system characteristics

The response space introduced in Section 5.2 necessarily hides a lot of detail. As described in Section 3.2.2, the nature of the data requires that we use the median as an average for a range of values. But we can explore the detail behind each of the 35 medians presented in the table body.

Before going into that detail, it is useful to note the information presented in the summary provided by the overall medians. By the coding used (see Section 3.2.2), a value of 0.0 could be considered neutral. Note that the data for 'Wary' (*I am wary of this AI system's decision-making process*) is reverse coded so as to align its direction with our quantified construct of trust and hence make it readily comparable to the other question results.

In Fig. 5.2, negative values indicate disagreement with the prompt question (except for 'Wary', as just explained). Thus, high (green) indicates agreement in the direction of trust. Low (orange) indicates less agreement.

## 5. Results and analysis

Despite being a ‘blunt’ instrument [154], there is a value in reviewing a sorted version of this table. Fig. 5.5 shows the same data as Fig. 5.2 but with the sort order being provided by a mean calculation. The mean is used as a convenience since it is more sensitive to the full data in each row and in this case does not violate the need for non-parametric techniques with the likert data.

AI type	count	Like	Wary <sup>1</sup>	Conf	Tech	Appr	Median	Mean <sup>2</sup>
Data	42	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Expl	41	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Risk	40	0.5	0.5	0.5	0.5	0.5	0.5	0.5
CNN	43	0.5	-0.5	0.5	0.5	0.5	0.5	0.3
Appv	42	0.5	-0.5	0.0	0.5	0.5	0.5	0.2
Perf	42	0.0	-0.5	0.0	0.5	0.5	0.0	0.1
Nila	42	-0.5	-0.5	-0.5	-0.5	0.5	-0.5	-0.3
	Median	0.5	-0.5	0.5	0.5	0.5		

1 - reverse coded

2 - although median is the most appropriate average for the data, mean-ordering reflects the detail

Figure 5.5: Median evaluation scores for each AI system characteristic and likert question ordered by overall average score.

We see that there is a possible indication that people don’t feel as comfortable with what we call the *vanilla AI* (rendered as ‘Nila’ in the table) as they do with those having some expressed, human-readable characteristic. And they may feel most comfortable (and least wary of) the machines characterised by good training data, providing an explanation or reporting a relative risk value. The ‘high performing’ AI system appears, on the face of it, to make users more wary and to get lower confidence and preference scores. As we discuss later (Section 6.1.3.7), this might be owing to the particular performance figure quoted as the characteristic. Approval by the medical royal colleges and a state-of-the-art technical design also appear to produce weaker expectations of confidence and greater wariness. Now it is time to look in more detail at the data behind these medians.

### 5.4.1 Plots on the trust dimension

How the 292 responses to each question vary across the different AI system characteristics is summarised in a series of box plots in Fig. 5.6.

In the red lines of Fig. 5.6 we can see the five lots of seven different median values that appeared in the tabulation in Fig. 5.5. And we likewise see what appears to be a



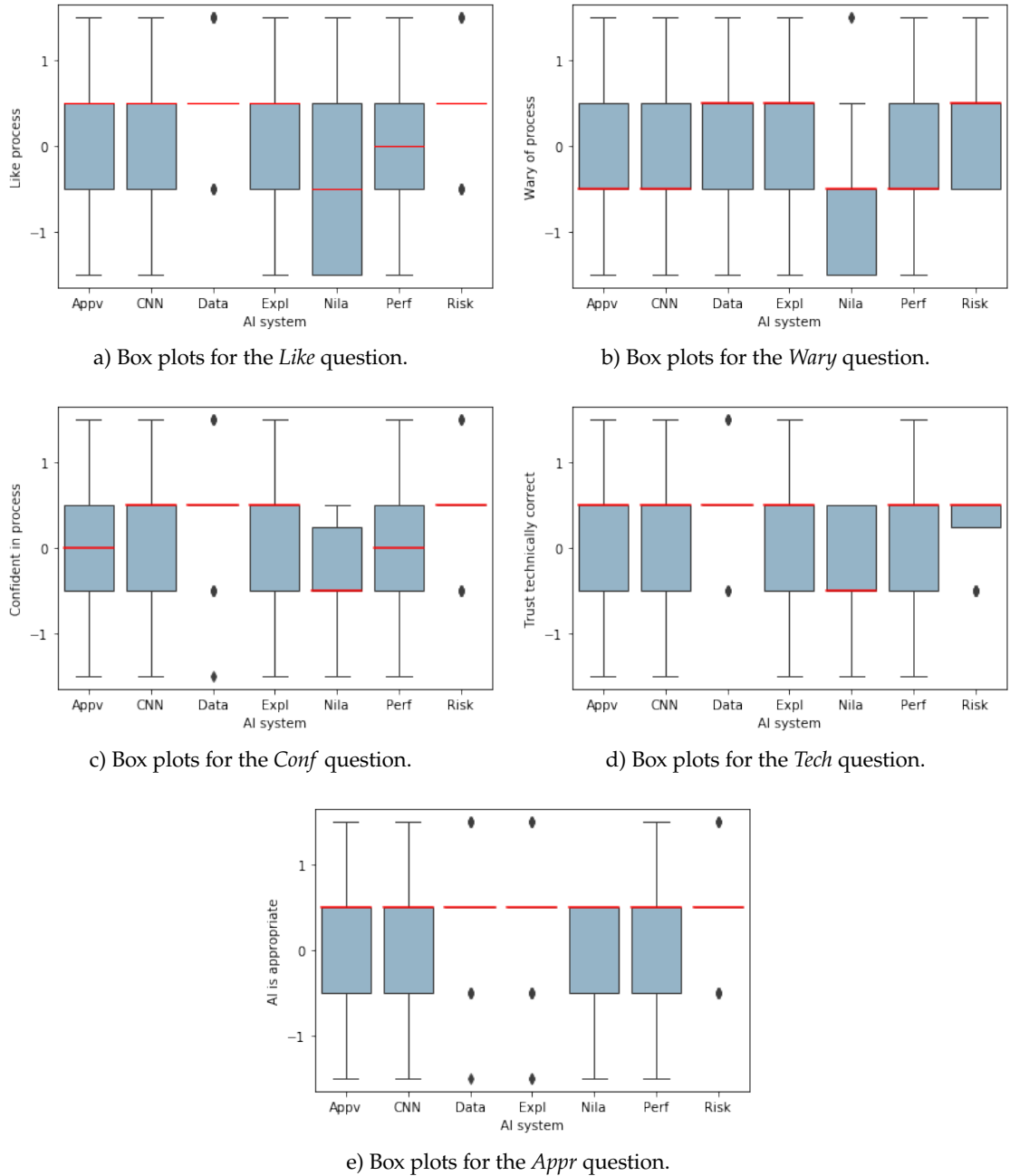


Figure 5.6: Box plots for five evaluation questions across seven AI system characteristics. See Table 5.3 and Table 5.4 for the label explanations.

separation of the *vanilla* result from the others in the *Wary* question in Fig. 5.6b). That is, the IQRs have no overlap.

The *Conf* question chart suggests differences in user confidence between the *vanilla* system and both *Data* and *Risk* systems. This apparent difference in distribution is shown in the last two plots, too, to different degrees. But we also see that there is a good deal of overlap between most of the IQRs. While there may be a real difference in the evaluation of the *vanilla* system and the others presented here, with their different medians, we do not have evidence to assert this from box plots alone. In most cases, contiguous IQRs will lead us to view the distinct medians as a poor indicator. On the other hand, there may be some value in looking more deeply at the differences between the *vanilla* system and the *Data* and *Risk* systems. In most cases, both their medians and their IQRs are distinct.

### 5.4.2 Considering the Wilcoxon signed-rank test (WSR test)

To test whether there is any statistical difference detectable in the responses to different characteristics, we need a non-parametric test of the null hypothesis that the distributions are identical. Although there are over 40 evaluations for each AI system characteristic, many of these are submitted as a set of three by the same participant (under different patient stories). As a result, we cannot assume independence of these submissions.

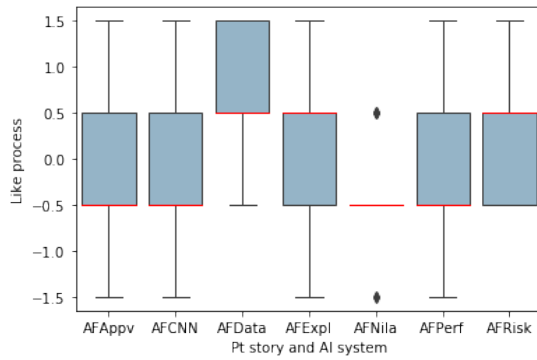
But if we split the dataset into the three distinct patient stories (as listed in Table 5.5), then we ensure that we can create independent pairs of evaluations submitted by the same user on different AI systems. This allows us to fulfil the criteria for a Wilcoxon signed-rank (WSR) test. There is a minimum requirement of 5 pairs of observations (with non-zero differences) to make the WSR test. So the WSR test looks like a suitable candidate for our exploration. What we will be testing for is the null hypothesis that the median of the differences between ranked pairs of evaluations is zero. We will conduct a series of tests across the AI system characteristics - each test on all the responses to a single question under a single patient story.

To ensure a focus for our exploration and therefore reduce the total number of tests, we will first prepare the data and generate box plots for the individual evaluations.

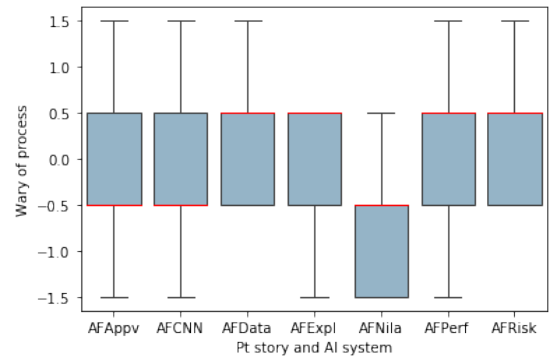
### 5.4.3 Preparing data for the Wilcoxon signed-rank test (WSR test)

As mentioned in Section 5.4.2, we are not free to use all the submitted data in a Wilcoxon signed-rank test (WSR test).

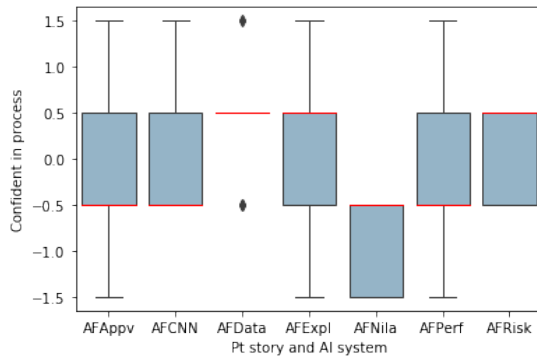
#### 5.4. Evaluation of the AI system characteristics



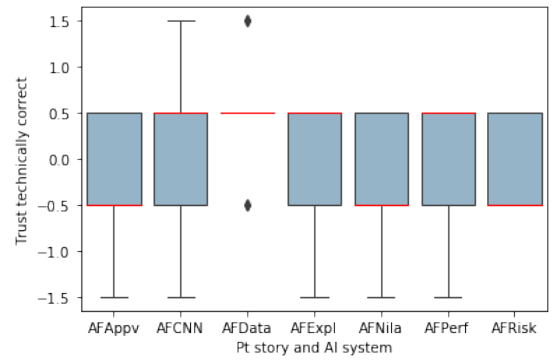
a) Box plots for the *Like* question. AF patient story.



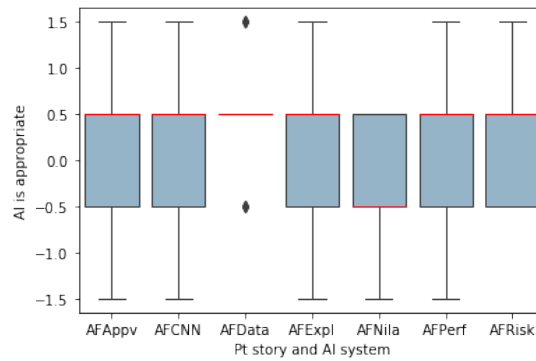
b) Box plots for the *Wary* question. AF patient story.



c) Box plots for the *Conf* question. AF patient story.



d) Box plots for the *Tech* question. AF patient story.



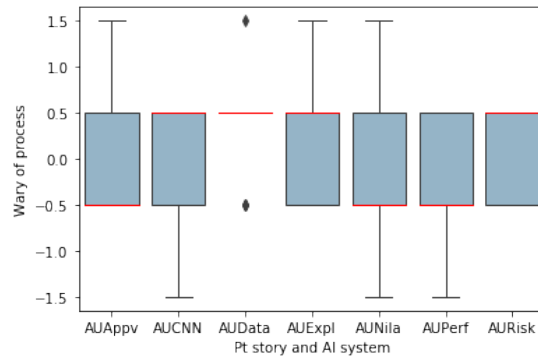
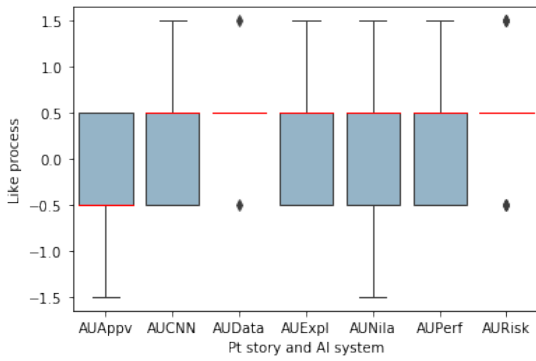
e) Box plots for the *Appr* question. AF patient story.

Figure 5.7: AF patient story. Box plots for five questions across seven AI system characteristics. See Table 5.3 and Table 5.4 for the label explanations.

Those users who did not submit a full ‘session’ of seven evaluations on a given patient story did not provide us with an intact group for the WSR test. Sub-setting the data to prepare for the WSR test leaves us with 13 participants on each patient story, amounting

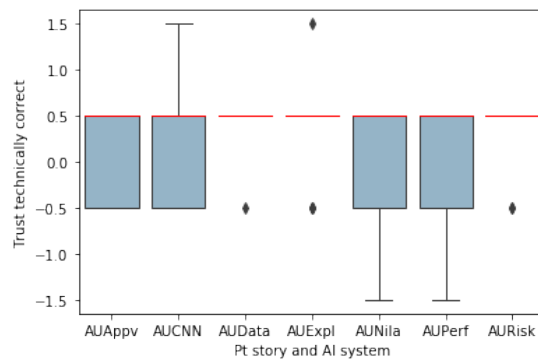
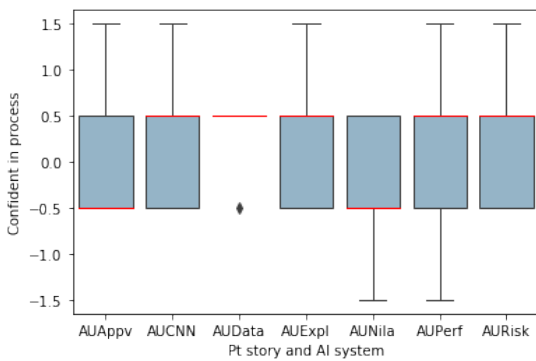
5. Results and analysis

to 273 submissions in total. That is, for this analysis, we are forced to dispense with 19 submissions out of the original 292.



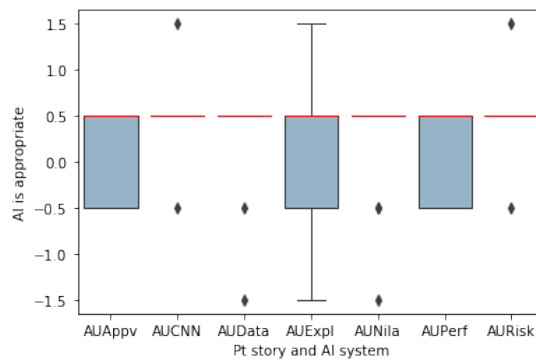
a) Box plots for the *Like* question. AU patient story.

b) Box plots for the *Wary* question. AU patient story.



c) Box plots for the *Conf* question. AU patient story.

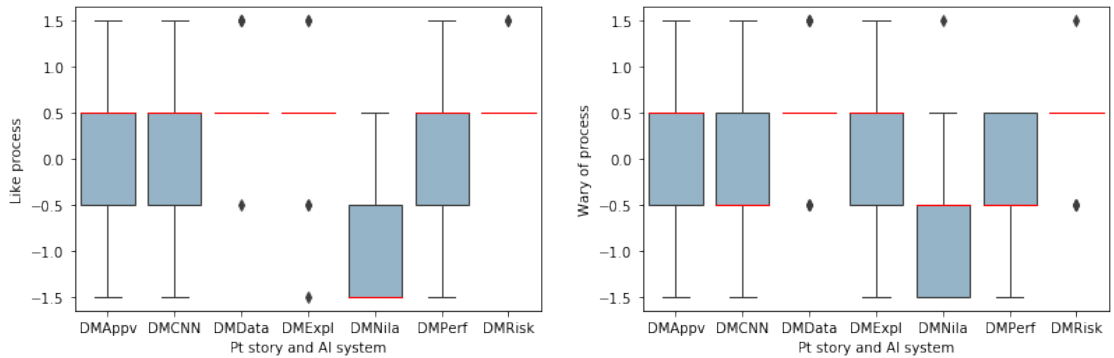
d) Box plots for the *Tech* question. AU patient story.



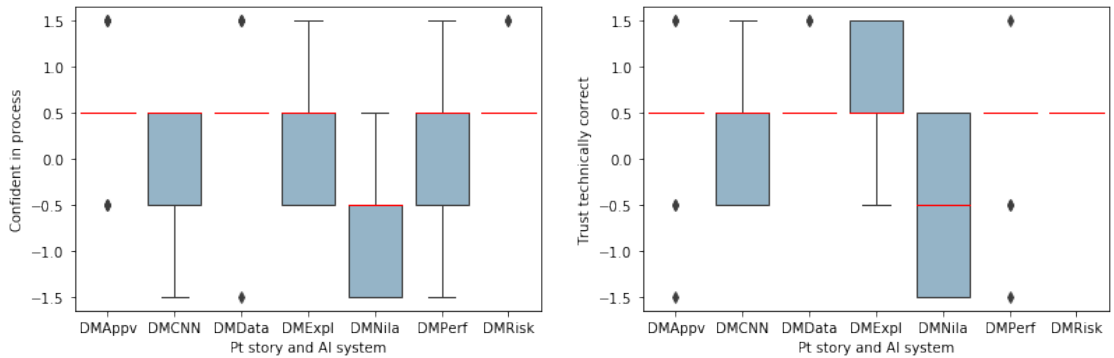
e) Box plots for the *Appr* question. AU patient story.

Figure 5.8: AU patient story. Box plots for five questions across seven AI system characteristics. See Table 5.3 and Table 5.4 for the label explanations.

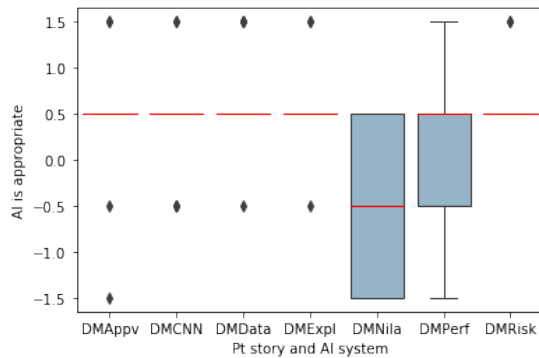
We then repeat the box-plotting exercise but separately for each patient story. Despite meaning that we will review all seven AI system characteristics on all five questions on each of three patient stories, this extensive approach is justified.



a) Box plots for the *Like* question. DM patient story. b) Box plots for the *Wary* question. DM patient story.



c) Box plots for the *Conf* question. DM patient story. d) Box plots for the *Tech* question. DM patient story.



e) Box plots for the *Appr* question. DM patient story.

Figure 5.9: DM patient story. Box plots for five questions across seven AI system characteristics. See Table 5.3 and Table 5.4 for the label explanations.

The justification arises from the fact that the review of visuals is more efficient than the full set of 315 pair-wise WSR tests. By reviewing the box plots at a patient story and question level, we are able to discriminate about which tests to conduct. In addition, we are tacitly acknowledging that the patient story (the decision context) could be a factor in influencing whether a given AI characteristic might promote trust, which is consistent with previous work [145].

One important consequence of this exploratory approach is that our multiple comparisons risk finding an apparently significant result by chance alone. So a Bonferroni correction must be applied before any consideration of statistical significance. We return to this in due course.

Looking at the plots for the AF (Atrial Fibrillation) patient story in Fig. 5.7, the plots indicate there may be some value in exploring the comparison of two AI system characteristics across most of the five questions. The *Data* characteristic appears to contrast with the *vanilla* AI system in all five questions. The *Risk* and *Expl* characteristics do so in all but the *Tech* question. The *Perf* characteristic shows some sign of distinguishing itself from the *vanilla* AI system in the case of the *Conf*, *Tech* and *Appr* questions. The *CNN* and *Appv* characteristics appear worth investigating under the *Appr* question only.

Turning to the plots for the AU (Acutely Unwell) patient story. We see in Fig. 5.8 that the *Data* and *Risk* characteristics look worth exploring in their comparison to the *vanilla* AI system for the *Wary*, *Conf* and *Tech* questions. In addition, the response to the *Expl* characteristic looks distinct from that of the *vanilla* AI system under the *Conf* and *Tech* questions. And the *Appv*, *CNN* and *Perf* systems also show signs of distinguishing themselves from *vanilla* under the *Tech* question. Interestingly, in a departure from contrasts with *vanilla*, we see three pair-wise comparisons under the *Like* question that might be worth exploring.

Now reviewing the plots for the DM (pre-Diabetic) patient story, we can see in Fig. 5.9 that most characteristics look like they contrast with the *vanilla* AI system for all five questions. Only under the *Wary* question are there a few pair-wise comparisons that look less likely to yield a significant difference. This gives us a total of 28 tests for the DM patient story where differences are worth testing.

#### 5.4.4 Wilcoxon signed-rank test (WSR test) results

Given the overview and focus established in Section 5.4.3, we are in a position to conduct a sequence of tests directed at a subset of the entire pair-wise dataset.

For this series of tests, we make use of the `wilcoxon` method of the `statistics` module of scientific python package `scipy` (v 1.5.2). And taking the number of comparisons to be 315, we note that instead of a direct 5% significance threshold, our Bonferroni-corrected (equivalent to 5% significance) requires an actual threshold of 0.016%.

For the AF (Atrial Fibrillation) patient story, as shown in Table 5.8, the majority of our focused tests fail to produce a significance level for the difference that is below 0.00016. Indeed, only one pair of evaluations appears to show a significant difference according to these figures. This is for *Data v Nila* under the *Conf* ('*I am confident...*') question. And this is close enough to our threshold to be of questionable interest in any event.

Table 5.8: Wilcoxon signed-rank test results for selected AI characteristic pairs - AF patient story.

ps	question	test	w	p	sig
AF	like	Data v Nila	74.0	0.002330	
AF	like	Risk v Nila	45.0	0.001946	
AF	like	Expl v Nila	30.5	0.033395	
AF	wary	Data v Nila	66.0	0.000883	
AF	wary	Risk v Nila	45.0	0.001946	
AF	wary	Expl v Nila	28.0	0.006937	
AF	conf	Data v Nila	91.0	0.000122	*
AF	conf	Risk v Nila	78.0	0.000395	
AF	conf	Expl v Nila	55.0	0.001755	
AF	conf	Perf v Nila	46.0	0.022790	
AF	tech	Data v Nila	36.0	0.004687	
AF	tech	Perf v Nila	15.0	0.158655	
AF	tech	CNN v Nila	15.0	0.158655	
AF	appr	Data v Nila	45.0	0.002875	
AF	appr	Risk v Nila	21.0	0.009815	
AF	appr	Expl v Nila	32.5	0.016736	
AF	appr	Perf v Nila	15.0	0.016947	
AF	appr	CNN v Nila	15.0	0.016947	
AF	appr	Appv v Nila	28.0	0.004075	

For the AU (Acutely Unwell) patient story, the computed results are shown in Table 5.9. The test results show no significance level below 0.00016 for the difference between characteristics of AI system in users' responses.

## 5. Results and analysis

---

At this point we also note that the tabulations display frequent repeats for values of  $w$  and  $p$  - a warning sign that our dataset is rather small for the task we might be asking it to perform.

Table 5.9: Wilcoxon signed-rank test results for selected AI characteristic pairs - AU patient story.

ps	question	test	w	p	sig
AU	like	Data v Appv	36.0	0.004150	
AU	like	Risk v Appv	36.0	0.004150	
AU	like	Data v Risk	10.5	0.500000	
AU	wary	Data v Nila	29.5	0.049121	
AU	wary	Risk v Nila	28.0	0.065834	
AU	conf	Data v Nila	21.0	0.009815	
AU	conf	Risk v Nila	37.0	0.035091	
AU	conf	Expl v Nila	12.0	0.089856	
AU	tech	Data v Nila	21.0	0.009815	
AU	tech	Risk v Nila	15.0	0.016947	
AU	tech	Expl v Nila	31.5	0.016947	
AU	tech	Appv v Nila	31.5	0.016947	
AU	tech	CNN v Nila	31.5	0.016947	
AU	tech	Perf v Nila	31.5	0.016947	

A challenge here is that any effect size (if it exists) may be small. And we have no guidance in this experimental area as to what would constitute a meaningful effect size in practice. In other words, we know very little about this abstract and elusive notion ‘*degree of trust*’. And, specifically, we don’t know how much extra of it might be needed to help bring worthwhile benefit to decision-making. This limits our ability to determine a meaningful power calculation. But we accept this as a limitation, given the exploratory nature of the study.

Another aside worth noting here is that our conscientious confinement to non-parametric techniques will inevitably make it harder to establish a confident result on a modestly-sized dataset.

But all this uncertainty is to be expected. We are yet exploring the indistinct outlines of the space of trust for clinicians and AI systems. We will return to this in Section 6.1.3.7. For completeness, we finally look at the WSR tests for the pre-diabetic patient story (the Diabetes Mellitus or DM story).



When responding to the pre-diabetic patient story (DM) Table 5.10, the WSR test calculations again suggest there are no significant differences in participants' evaluations of the various AI system characteristics.

Table 5.10: Wilcoxon signed-rank test results for selected AI characteristic pairs - DM patient story.

ps	question	test	w	p	sig
DM	like	Data v Nila	66.0	0.001441	
DM	like	Expl v Nila	45.0	0.003229	
DM	like	Risk v Nila	55.0	0.002033	
DM	like	Appv v Nila	43.5	0.005080	
DM	like	CNN v Nila	34.0	0.011307	
DM	like	Perf v Nila	45.0	0.002875	
DM	wary	Data v Nila	82.0	0.004028	
DM	wary	Expl v Nila	44.5	0.037667	
DM	wary	Risk v Nila	61.5	0.004053	
DM	wary	Appv v Nila	33.5	0.013599	
DM	conf	Data v Nila	55.0	0.002008	
DM	conf	Expl v Nila	51.5	0.006024	
DM	conf	Risk v Nila	55.0	0.002123	
DM	conf	Appv v Nila	42.5	0.008171	
DM	conf	CNN v Nila	25.0	0.026603	
DM	conf	Perf v Nila	25.0	0.026603	
DM	tech	Data v Nila	45.0	0.003229	
DM	tech	Expl v Nila	52.0	0.005561	
DM	tech	Risk v Nila	45.0	0.003085	
DM	tech	Appv v Nila	42.0	0.009128	
DM	tech	CNN v Nila	41.5	0.009847	
DM	tech	Perf v Nila	45.0	0.002479	
DM	appr	Data v Nila	36.0	0.005160	
DM	appr	Expl v Nila	36.0	0.005160	
DM	appr	Risk v Nila	36.0	0.004150	
DM	appr	Appv v Nila	28.0	0.007882	
DM	appr	CNN v Nila	33.5	0.013599	
DM	appr	Perf v Nila	28.0	0.005706	

While these tests have yielded results showing no statistical significance in the differences that were apparent in the box plots, there are reasons to suppose this is not the end of the story.

As mentioned, the small number of pairs obtained in this study (and available for this kind of comparison) make it unlikely we would detect anything but the very largest of differences, should they exist. And the large probability of zero-valued differences

(owing to the short Likert scale) adds to the challenge here. While the box plots may indicate areas of interest for further research, we cannot draw a conclusion here either that the apparently significant difference (*Data* v *Nila* under the *Conf* question in AF) or the the apparently non-significant differences (*Risk* or *Expl* v *Nila*) are reliable. We return to this in the discussion in Section 6.1.3.7.

### 5.5 Clinician vs non-clinician evaluations

Although not an explicit design feature, the ability to split participant responses between those who are trained clinicians and those who aren't produces some potentially interesting indications for further study, as already mentioned.

The alignment between clinicians and non-clinicians is close for many of the AI systems. But notably diverges on the AI system that provides some explanation. Fig. 5.10 shows a set of boxplots that indicate non-clinicians gave a trust rating that was distinctly higher than that provided by clinicians to such a system under all but one question. Only under the *Wary* question is this not visible. The plots where the difference is apparent are shown.

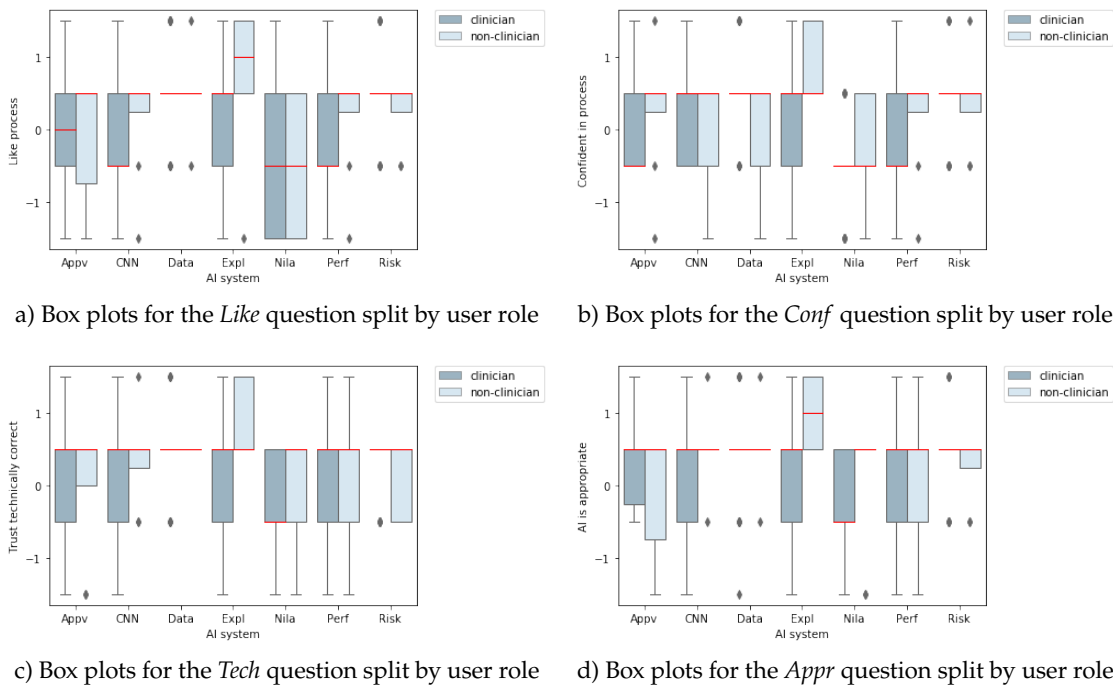


Figure 5.10: Box plots for four questions across seven AI system characteristics split by user role. See Table 5.3 and Table 5.4 for the label explanations.

Since the stimulus material (the patient stories) were not designed with non-clinicians in mind, the experiment is weak in its ability to detect such a difference. So testing for significance is not considered appropriate here. Rather, the plots merely suggest another area for possible future exploration.

## 5.6 Qualitative responses

The free text responses were completed in 40% of submissions. Non-clinical participants were more likely to complete free text responses. Among clinical users, the completion rate for the optional free text sections fell to 30%.

The most common point made was that there was insufficient information to make an informed evaluation. This was frequently phrased in terms of what was lacking in a system (that was present in one of the others), so that what was conveyed in the narrative text was a dissatisfaction with the narrowness of the given AI system characteristics. Where performance figures or a relative risk were given, the user might complain that there was no information about the test and training data used. Where an explanation of the result was provided or the training data were elaborated, the complaint might be the dearth of performance data and so on. This is completely understandable. And it is useful to have these reminders of user needs. An important observation is that complaints are often associated with relatively high trust scores. That is, the critique is a constructive criticism, indicating what would make the system even better.

Interestingly, among medics, approval from the medical royal colleges was a characteristic received with highly variable support. Some found it very reassuring - others were frostily dismissive.

Non-medically trained participants were definitely left behind by some of the information provided in relation to data-grounded systems and risk-based assessments. This serves as a warning not to place too much reliance on apparent differences between different user-types.

Finally, it should be emphasised that the majority of users didn't provide free text narrative material. And among clinicians at least, this may have been seen as a time trade-off. Those clinicians who added free text completed, on average, only half as many evaluations. Meanwhile, among non-clinicians, those who added free text completed a comparable number of evaluations to those who didn't.

## **5.7 Observations on the range of clinical participants**

The breakdown of participants by role is of potential interest. Clinicians are notoriously difficult to recruit to occasional studies, so the good response from clinicians suggests both that the study is of interest to them and that the user experience was not overly off-putting.

## Chapter 6

# Discussion, Conclusions and Future Work

### 6.1 Discussion

Trust and AI in clinical decision support is not a single measurable concept. It brings together three wide-ranging and heterogeneous spaces of human experience, thought and action: medicine, intelligent systems and the concept of trust itself. Each on its own is too extensive to assess completely within the scope of a single research project. Combined, they present a huge space for exploration. The importance of exploring nevertheless provides the motivation for a small attempt.

#### 6.1.1 A framework for clinical engagement in participatory design

We have engaged with a core group of clinical advisers to co-create a focused means of examining six hypothetical AI systems, each with a distinctive characteristic that we believe should be tested in respect of its tendency to induce trust when employed in clinical decision support.

The framework of clinical engagement for this participatory design project has proven to successfully bring together a small group of busy, working clinicians and have each of them contribute to a complex project. By means of well-prepared, appropriately-stimulated discussions, we have been able to draw out from them the domain-specific insight and detailed knowledge that allowed us to produce a carefully crafted user-study.

We believe the value of this work lies in its having demonstrated how much clinically-sourced data can be accumulated in what has been a very condensed phase of the project. The fact that such a high proportion of study participants were prepared to complete the maximum three sessions is an indication that the design was well-suited to the target group and, as a result, experience was not too burdensome.

### **6.1.2 A scalable web-based study tool for the exploration of trust and AI**

We have created a robust web application that allows user-authentication, good participant experience, tailored interaction and secure storage of submitted data.

The web application design has proven capable of serving a mobile-friendly data-capture process with a code base that is secure and scalable. The modular, object-oriented approach allows efficient updating of stimulus material and questionnaire design so that extension, re-use or refactor are all equally straightforward to accomplish.

The value of this work, we believe, is that all the code is to a modern commercial production-level standard and is re-usable in further research - whether in the clinical domain or outside it.

### **6.1.3 Discussion of the participant evaluation data**

#### **6.1.3.1 Preliminary questions**

The responses to the preliminary questions do show a range of opinions. A minority of participants have recorded strong agreement and strong disagreement with the general (non-clinical) AI-related propositions. While the box plots are informative, parallel co-ordinate plots would show better whether we have greater variation between users or between questions. In any event, the responses here are not a prime focus of the study, but were captured to provide some confidence that there is not a consistent skew either in favour or against AI technology among the participants.

#### **6.1.3.2 Participation**

As mentioned in Section 6.1.1, the recruitment and retention of numbers of clinical users demonstrates a strength in the quality and suitability of the user-experience for the target audience. There are, however, both strengths and weaknesses to the late inclusion of non-clinical participants.

On the one hand, given that clinical users are a distinct sub-set of the population and are likely to have distinct responses in relation to the trust they will place in an AI system, it is useful to understand how well (or how poorly) other studies that explore responses in the general population can serve as a proxy for clinician opinion. Our research discovered two comparable papers where clinicians were recruited to ascertain their propensity to trust an AI system. In the first [147], the opinions of 14 doctors were recorded against four general questions and against one question on each of four ‘patient stories’. Alongside this, the views of 30 non-clinicians were sought against the same questions, not to elicit their own views, but to see whether they could, as designers, anticipate clinical preferences. They could not. In the second [51], 170 doctors were surveyed and no non-clinicians. In neither case was there a comparable involvement of clinical and non-clinical participants.

On the other hand, despite the positive motive and apparent opportunity, the decision to include non-clinical participants does not provide an equivalent test of their inclination to trust various AI systems. At least one of the motivations to invite non-clinicians was that shared decision-making is an important and historically neglected aspect of clinical care. But a corollary to shared decision-making is the need to make clinically relevant information accessible to the patient. In this case, none of the stimulus material was provided with an appropriate ‘*patient-speak*’ treatment [161]. As a result, the apparent differences between clinicians and non-clinicians as recorded need to be treated with extreme caution.

### 6.1.3.3 Experimental design and the distribution of independent submissions

Of those participants who did whole sessions, half did fewer than the full three that were possible. Had these all been undertaken on one patient story rather than uniformly distributed over the three different stories, we would have had larger numbers on a single story. And all of these would have been independent submissions. The significance of any resulting differences would have been likely to be more robust as a result. But on the other hand, that experimental design (providing only one patient story) would have missed the opportunity afforded by the other half of those participants to broaden the context on which the trust questions were based. In this case, we have insufficient numbers both of the ‘*full-set*’ participants and of the ‘*part-set*’ participants to show whether significant differences exist in the underlying population for any or all patient stories. But the recruitment, design and capture process is appropriate for each of the two possible outcomes and we believe the decision to cater for both was therefore justified.

#### 6.1.3.4 Sampling design

The sample of participants used in the project results from a convenience sampling approach. No attempt was made to ensure the sample population accurately reflected a particular mix of clinical (or non-clinical) roles or specialties. The recruitment mechanism employed only partially targetted requests via a 'snowball' technique (where participants and other contacts are asked to promote the study to their own networks). The partial targetting was achieved through the specific clinical networks of the clinical advisory group plus some other key clinical contacts.

A larger scale study might legitimately use the same recruitment approach but might include a specific mechanism for ensuring a representative sample was achieved - for example by continuing to promote and recruit on some channels in order to increase sub-population representation.

#### 6.1.3.5 Randomisation mechanisms

While the technical mechanism for randomising the presentation is simple, the significance of its inclusion is considerable. Randomisation is used to set the order of patient stories and AI system characteristics along with the colour-association of each system on a per-user basis. This has significance for the creation of a uniform distribution of submissions across the response space on the dimensions of patient story and AI system type. It also mitigates any risk of the aggregate of responses to the different system characteristics being influenced by the order of presentation or colour treatment given by the website.

#### 6.1.3.6 Likert scales

Although the work to create suitable Likert scale response options is supported by reliable sources, the preparation work did not extend to any reliability testing. This was a conscious compromise that could have been addressed with more time. An improvement would be to check the question set passes the Cronbach's alpha or the Kappa test of intercorrelation and validity [153].

An alternative to multi-label likert scales (which are hard to make reliable) is to have multi-point scales that are labelled only at the extremities. These are not without their own problems for users. But they would likely help avoid the high probability of 'collisions' experienced with short scales.



Another consideration for Likert scale design is to form groups of closely related scale questions. The values for these can be summed to provide more sensitive results [153]. On the other hand, this approach requires users to complete what are often viewed as pointlessly similar questionnaire responses - with the risk that they lose motivation and submit less data overall.

In retrospect, a weakness of this study is that, in attempting to ensure a user-experience that is '*low friction*' and hence not off-putting to busy clinicians, the design choices have created a greater reliance on getting larger numbers of users recruited. An attempt to address this would, however, have to consider whether an approach such as the multi-point scales mentioned above would provide a more reliable and discriminating result set without a significant increase in participant numbers.

#### **6.1.3.7 How users trust AI decision support can depend on system characteristics**

The box plots on all data in Fig. 5.6 are useful as an overview of the response space. They reveal more detail than can be shown in a table, for example. Where there appears to be a lower level of trust in a performance score, we should consider the possibility that the response may be conditioned by expectation (a form of priming). So that the choice of performance figure presented in the scenario may be a sensitive factor for some users if they are less familiar with the specialty area and have an expectation of a higher performance figure.

Where the plots have been split by patient story (Fig. 5.7, Fig. 5.8 and Fig. 5.9), we begin to see more strongly the effect of the low numbers. In many of the plots, the AI system groups begin to look very similar to each other. Conversely, where they do show a difference, this does give a good indicator of where further work might provide good evidence.

The results of the Wilcoxon signed-rank (WSR) tests are also interesting in themselves (in a negative sense). They provide outputs that don't support our tentative interpretations of the data - and as such, they are an important part of the overall spread of research findings in this area. We inevitably see that the low number of data points means that any calculation of significance should be considered as less reliable. The effect of the short Likert scale is to exacerbate this reliability issue. The WSR test relies on the rank (ie, related to absolute value) and sign (ie, direction) of the differences between pairs across the two groups under scrutiny. Where there is a skew in the distribution of differences (eg, skewing to be larger in one direction), then the statistical test can measure significance on

the strength of this skew. But if too many pairs of values are equal, then the remaining differences cannot establish significance.

Our use of a four-point likert scale means that ‘collisions’ and hence zeroes occur with high probability. The solution to this problem would be to collect more data.

The separate sets of box plots for the three patient stories shown in Fig. 5.7, Fig. 5.8 and Fig. 5.9, show a large degree of consistency. One place that is an exception is in the responses to the *Nila* system type under the AU and DM patient stories. The difference can be seen in Fig. 5.8a) compared to Fig. 5.9a). Whereas under the AU story, the median response is agreeing with the *Like* sentiment, under the DM story, the median is strongly disagreeing.

Given that we cannot establish reliable measures of significance for these groups, it is not possible to draw a conclusion from this observation in the values. However, the suggestion that there may be a greater inclination to trust an AI system under one patient story than another raises two points for further study.

First, a method of calibrating the stakes of any given decision context would be a useful addition to the design armoury. Ashoori et al found that users were less likely to trust an AI system in the context of a higher stakes decision (a prison sentencing decision vs a meal planning decision) [145]. In this case, the mechanism identified by Ashoori et al is unlikely to account for the difference since the clinical risk of patient harm is much greater in the acutely unwell decision scenario than in the pre-diabetic decision scenario.

Second, it is reasonable to assume that there are multiple contributory processes playing into an inclination to trust an AI system in any given context. A highly competent acute physician (in common with most trained medics) will be little challenged by the urgent assessment and action sequence she expects to follow when faced with an acutely unwell patient. But clinicians in other roles may, at a certain point of acuity, be willing to find help in even a relatively unknown system, if it helps bridge the gap between their sense of control and what the situation demands. For this reason, future work should consider recording much more detail about the clinical role, specialty, experience and seniority of participants.

## 6.2 Conclusions

This study found some evidence that users respond differently in terms of trust to AI decision support systems with different characteristics. The type and direction of differences are unremarkable for the most part. There are suggestions that some differences

may be impactful for designers of clinical decision support systems. With more data, some interesting and reliable results are possible.

## 6.3 Contributions

The main contributions of this work can be summarised as follows:

- **An effective framework for clinical engagement in participatory design**

The framework proved itself invaluable in the preparation of the study material in this project. It is useful going forward because it demonstrates a successful approach that will allow further exploration of aspects of human-computer interaction with this important group.

- **A scalable web-based study tool**

Written in Django with a modular approach which means the priority given to clinical opinion recording in this instance can be leveraged for future work - or it can be replaced by material suited to another domain.

- **Results showing that there are areas of apparent difference between participants' responses to different AI system characteristics and that the relationship may be complex**

The experimental design has proved capable of drawing data of adequate quality together. With more data points, we would be able to reliably discover where differences do exist. We have shown that establishing the existence of differences in response, if they exist, will be straightforward.

## 6.4 Future Work

### 6.4.1 More data would support more robust statistical test results

Any extension of this work should consider ways to increase the count of independent submissions. This could be done by altering the randomisation to 'fill up' one of the patient stories to a level that would provide sufficient power before allowing subsequent submissions to switch over to the next patient story and so on. We considered this strategy during the design stages but discounted it for two reasons. First was the development overhead. More time spent developing translates to less time collecting data. The second

reason is that creating a preference order for the stories without having established their equivalence could result in data that gave a skewed signal

A second approach to the requirement for larger sub-groups is to ensure that more data are captured overall. A longer period of recruitment would increase the prospects of getting clinicians in certain roles as their workloads frequently have cycles of more than a couple of weeks.

#### **6.4.2 Improved or alternative assessment instrument**

Alternatives to the four point likert scales could be trialled to see if they made recruitment and retention any more difficult. Certainly, some reliability testing should be carried out on the scales used in any extension to this work.

#### **6.4.3 Patient stories**

The work to create suitable patient stories is non-trivial. So while increasing the range of stories could add support to any findings, the additional work involved should be weighed against the possible benefit. Certainly, an attempt to calibrate the stories or discover in what ways they may differ would be an improvement (see the point on stakes below).

#### **6.4.4 Differences between participants by role**

As we said in Section 5.5, it would be useful and interesting to develop an approach that was capable of reliably detecting any difference that might exist between participants with different roles. In this respect, we would be unsurprised to find differences between medics and other clinicians as well as between clinicians and non-clinicians.

#### **6.4.5 Differences between decision contexts by stakes**

As mentioned in Section 6.1.3.7, what is at stake in a given decision context is likely to have a complex relationship to role, experience, organisational context and other factors. Being able to isolate and control for this would improve the confidence we could place in any set of results.

# Bibliography

- [1] A. Giddens, *The Consequences of Modernity*. Cambridge UK: Polity Press, 1990.
- [2] D. Gambetta, *Trust. Making and Breaking Cooperative Relations*. Oxford, UK: Basil Blackwell, 1988.
- [3] —, “Trust: Making and breaking cooperative relations,” *The British Journal of Sociology*, vol. 13, 01 2000.
- [4] M. Witkowski, A. Artikis, and J. Pitt, “Experiments in building experiential trust in a society of objective-trust based agents,” 01 2000, pp. 111–132.
- [5] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995. [Online]. Available: <https://doi.org/10.5465/amr.1995.9508080335>
- [6] G. Dietz and D. Den Hartog, “Measuring trust inside organisations,” *Personnel Review*, vol. 35, no. 5, pp. 557–588, 2006.
- [7] J. Bryson, “No one should trust ai,” *Artificial Intelligence and Global Governance*, 2018.
- [8] —, “How do we hold ai itself accountable? we can’t.” 2019.
- [9] M. Y. Chandler, “Physician outcry on ehr functionality, cost will shake the health information technology sector,” *medicaleconomics.com*, Feb 2014. [Online]. Available: <https://www.medicaleconomics.com/view/physician-outcry-ehr-functionality-cost-will-shake-health-information-technology>
- [10] A. Gawande, “Why doctors hate their computers,” *The New Yorker, Annals of Medicine*, Nov. 2018. [Online]. Available: <https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers>

- [11] J. Kent, "50healthitanalytics.com, 2018. [Online]. Available: <https://healthitanalytics.com/news/50-of-physicians-not-satisfied-with-patient-data-access>
- [12] J. Bendix and L. Lutton, "Doctors sound off about ehr shortcomings," *medicaleconomics.com*, May 2020. [Online]. Available: <https://www.medicaleconomics.com/view/doctors-sound-about-ehr-shortcomings>
- [13] R. J. Keizer, E. Dvergsten, A. Kolacevski, A. Black, S. Karovic, S. Goswami, and M. L. Maitland, "Get real: Integration of real-world data to improve patient care," *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 722–725, 2020. [Online]. Available: <https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.1784>
- [14] J. D. Hron and E. Lourie, "Have you got the time? Challenges using vendor electronic health record metrics of provider efficiency," *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 644–646, 02 2020. [Online]. Available: <https://doi.org/10.1093/jamia/ocz222>
- [15] H. Thimbleby and P. Cairns, "Reducing number entry errors: solving a widespread, serious problem," *Journal of The Royal Society Interface*, vol. 7, no. 51, pp. 1429–1439, 2010. [Online]. Available: <https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2010.0112>
- [16] A. Downey, "Screening it systems 'unsafe' and require 'urgent' upgrades, review finds," *digitalhealth.net*, Oct 2019. [Online]. Available: <https://www.digitalhealth.net/2019/10/screening-it-unsafe-urgent-upgrades/>
- [17] B. Goldacre and B. MacKenna, "The nhs deserves better use of hospital medicines data," *BMJ*, vol. 370, 2020. [Online]. Available: <https://www.bmj.com/content/370/bmj.m2607>
- [18] Philips.com, "Future health index 2020: The age of opportunity. empowering the next generation to transform healthcare," *philips.com*, Philips, Tech. Rep., 2020. [Online]. Available: <https://www.philips.com/a-w/about/news/future-health-index/reports/2020/the-age-of-opportunity.html>
- [19] E. Topol, "Preparing the healthcare workforce to deliver the digital future," *nhs.uk*, Feb 2019. [Online]. Available: <https://topol.hee.nhs.uk/>

- [20] G. Eden, M. Jirotko, and B. C. Stahl, "Responsible research and innovation: Critical reflection into the potential social consequences of ict," *IEEE 7th International Conference on Research Challenges in Information Science (RCIS)*, pp. 1–12, 2013. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6577706>
- [21] J. Stilgoe, R. Owen, and P. Macnaghten, "Developing a framework for responsible innovation," *Research Policy*, vol. 42, no. 9, pp. 1568 – 1580, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0048733313000930>
- [22] B. Grimpe, M. Hartswood, and M. Jirotko, "Towards a Closer Dialogue between Policy and Practice: Responsible Design in HCI," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 2965–2974. [Online]. Available: <https://doi.org/10.1145/2556288.2557364>
- [23] B. Ribeiro and P. Shapira, "Responsible Research and Innovation (RRI): A Manchester perspective," University of Manchester, Tech. Rep., 07 2018. [Online]. Available: [https://www.researchgate.net/publication/326368331\\_Responsible\\_Research\\_and\\_Innovation\\_RRI\\_A\\_Manchester\\_perspective](https://www.researchgate.net/publication/326368331_Responsible_Research_and_Innovation_RRI_A_Manchester_perspective)
- [24] UKRI, "UKRI framework for responsible innovation," [epsrc.ukri.org](https://epsrc.ukri.org), Engineering and Physical Sciences Research Council (EPSRC), no date. [Online]. Available: <https://epsrc.ukri.org/research/framework/>
- [25] D. S. Wall, "Cybercrime and the culture of fear," *Information, Communication & Society*, vol. 11, no. 6, pp. 861–884, 2008. [Online]. Available: <https://doi.org/10.1080/13691180802007788>
- [26] A. Gray, "The robots are coming – for as many as 800 million jobs," [weforum.org](https://www.weforum.org), 2017. [Online]. Available: <https://www.weforum.org/agenda/2017/12/robots-coming-for-800-million-jobs>
- [27] J. Manyika, S. Lund, M. Chui, J. Bughin, J. Woetzel, P. Batra, R. Ko, and S. Sanghvi, "Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages," [mckinsey.com](https://www.mckinsey.com), Nov 2017. [Online]. Available: <https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages>

- [28] M. Bishop, "Fear artificial stupidity, not artificial intelligence," London, UK, Dec 2014. [Online]. Available: <https://www.newscientist.com/article/dn26716-fear-artificial-stupidity-not-artificial-intelligence/>
- [29] D. G. Johnson and M. Verdicchio, "Ai anxiety," *Journal of the Association for Information Science and Technology*, vol. 68, no. 9, pp. 2267–2270, 2017. [Online]. Available: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23867>
- [30] S. Poole, "The government called the exam algorithm 'robust'. how robust was that claim?" <https://www.theguardian.com>, Aug 2020. [Online]. Available: <https://www.theguardian.com/books/2020/aug/27/the-government-called-the-exam-algorithm-robust-how-robust-was-that-claim>
- [31] S. Coughlan, "A-levels and gcses: Boris johnson blames 'mutant algorithm' for exam fiasco," [bbc.co.uk](http://bbc.co.uk), Aug 2020. [Online]. Available: <https://www.bbc.co.uk/news/education-53923279>
- [32] L. Torcello, "Science denial, pseudoskepticism, and philosophical deficits undermining public understanding of science: A Response to Sharon E. Mason." *Social Epistemology Review and Reply Collective*, vol. 9, no. 9, pp. 1–9, Sep. 2020. [Online]. Available: <https://social-epistemology.com/2020/09/01/science-denial-pseudoskepticism-and-philosophical-deficits-undermining-public>
- [33] R. Darner, "How can educators confront science denial?" *Educational Researcher*, vol. 48, no. 4, pp. 229–238, 2019. [Online]. Available: <https://doi.org/10.3102/0013189X19849415>
- [34] J. M. I. P. Webb, H. and M. Patel, "Human centred computing approaches to embed responsible innovation in hci." 2019. [Online]. Available: <https://ora.ox.ac.uk/objects/uuid:e7e87d82-f96a-4b2e-921c-a3a364f33ccb>
- [35] S. Deng and A. Varzi, "Methodological blind spots in machine learning fairness: Lessons from the philosophy of science and computer science," [arxiv.com](http://arxiv.com), 2019. [Online]. Available: <https://arxiv.org/abs/1910.14210>
- [36] J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. Smith, "Show your work: Improved reporting of experimental results," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*



- 
- Joint Conference on Natural Language Processing*, Nov 2019. [Online]. Available: <https://www.aclweb.org/anthology/D19-1224.pdf>
- [37] A. F. T. Winfield and M. Jirotko, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180085, 2018. [Online]. Available: <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2018.0085>
- [38] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourny, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-ai interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–13. [Online]. Available: <https://doi.org/10.1145/3290605.3300233>
- [39] G. Inc, "Perspectives on issues in ai governance," Google, Mountain View, CA 94043, techreport, Jan. 2019. [Online]. Available: <https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf>
- [40] M. Kwiatkowska, "Cognitive reasoning and trust in human-robot interactions," in *Proc. 14th Annual Conference on Theory and Applications of Models of Computation (TAMC 2017)*, ser. LNCS. Springer, 2017, to appear. [Online]. Available: <http://qav.comlab.ox.ac.uk/bibitem.php?key=Kwi17>
- [41] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human-Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020. [Online]. Available: <https://doi.org/10.1080/10447318.2020.1741118>
- [42] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang, "From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people," in *Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–6. [Online]. Available: <https://dl.acm.org/doi/10.1145/3334480.3381069>

- [43] K. Siau and W. Wang, "Building trust in artificial intelligence, machinelearning, and robotics," *Cutter Business Technology*, vol. 31, pp. 47–53, 2018. [Online]. Available: <https://www.cutter.com/sites/default/files/itjournal/2018/cbtj1802.pdf>
- [44] Z. C. Lipton, "The mythos of model interpretability," arxiv.com, 2016. [Online]. Available: <https://arxiv.org/abs/1606.03490>
- [45] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82 – 115, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1566253519308103>
- [46] D. Castelvechi, "Can we open the black box of ai," *Nature*, vol. 538, pp. 20–23, 2016. [Online]. Available: <https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731>
- [47] W. Guo, "Explainable artificial intelligence (xai) for 6g: Improving trust between human and machine," 2019. [Online]. Available: <https://arxiv.org/abs/1911.04542>
- [48] C. Cobey and J. Boillet, "How do you teach ai the value of trust?" ey.com, 2018. [Online]. Available: [https://www.ey.com/en\\_gl/digital/how-do-you-teach-ai-the-value-of-trust](https://www.ey.com/en_gl/digital/how-do-you-teach-ai-the-value-of-trust)
- [49] B. Heinrichs and S. B. Eickhoff, "Your evidence? machine learning algorithms for medical diagnosis and prediction," *Human Brain Mapping*, vol. 41, no. 6, pp. 1435–1444, 2020. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24886>
- [50] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar, "Estimating counterfactual treatment outcomes over time through adversarially balanced representations," arxiv.com, 2020. [Online]. Available: <https://arxiv.org/abs/2002.04083>
- [51] W. K. Diprose, N. Buist, N. Hua, Q. Thurier, G. Shand, and R. Robinson, "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator," *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 592–600, 02 2020. [Online]. Available: <https://doi.org/10.1093/jamia/ocz229>

- 
- [52] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1721–1730. [Online]. Available: <https://doi.org/10.1145/2783258.2788613>
- [53] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?': Explaining the predictions of any classifier," p. 1135–1144, 2016. [Online]. Available: <https://dl.acm.org/doi/10.1145/2939672.2939778>
- [54] O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz, and P. Blunsom, "Can i trust the explainer? verifying post-hoc explanatory methods," 2019. [Online]. Available: <https://arxiv.org/abs/1910.02065>
- [55] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019. [Online]. Available: <https://www.nature.com/articles/s42256-019-0048-x>
- [56] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arxiv.com, 2017. [Online]. Available: <https://arxiv.org/abs/1702.08608>
- [57] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.
- [58] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, Aug. 2018. [Online]. Available: <https://dl.acm.org/doi/10.1145/3236009>
- [59] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your ai: Expertise and explanations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 240–251. [Online]. Available: <https://doi.org/10.1145/3301275.3302308>

- [60] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," 2019. [Online]. Available: <https://arxiv.org/abs/1802.07810>
- [61] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–12. [Online]. Available: <https://doi.org/10.1145/3290605.3300509>
- [62] K. Yu, S. Berkovsky, D. Conway, R. Taib, J. Zhou, and F. Chen, *Do I Trust a Machine? Differences in User Trust Based on System Performance*. Cham: Springer International Publishing, 2018, pp. 245–264. [Online]. Available: [https://doi.org/10.1007/978-3-319-90403-0\\_12](https://doi.org/10.1007/978-3-319-90403-0_12)
- [63] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," ser. *Proceedings of Machine Learning Research*, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: <http://proceedings.mlr.press/v81/buolamwini18a.html>
- [64] J. Peters, "Ibm will no longer offer, develop, or research facial recognition technology," [www.theverge.com](http://www.theverge.com), Jun. 2020. [Online]. Available: <https://www.theverge.com/2020/6/8/21284683/ibm-no-longer-general-purpose-facial-recognition-analysis-software>
- [65] Amazon, "We are implementing a one-year moratorium on police use of rekognition," [amazon.com](http://amazon.com), Jun. 2020. [Online]. Available: <https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition>
- [66] J. Greene, "Microsoft won't sell police its facial-recognition technology, following similar moves by amazon and ibm," [washingtonpost.com](http://washingtonpost.com), Jun. 2020. [Online]. Available: <https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/>
- [67] S. Ampamya, J. M. Kitayimbwa, and M. C. Were, "Performance of an open source facial recognition system for unique patient matching in a resource-limited setting,"

- International Journal of Medical Informatics*, vol. 141, p. 104180, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1386505619312481>
- [68] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, no. 3, pp. 231–237, 2019. [Online]. Available: <https://qualitysafety.bmj.com/content/28/3/231>
- [69] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P. Ossorio, S. Thadaneys-Israni, and A. Goldenberg, "Do no harm: a roadmap for responsible machine learning for health care," *Nature Medicine*, vol. 25, pp. 1337–1340, 2019. [Online]. Available: <https://www.nature.com/articles/s41591-019-0548-6>
- [70] J. Yoon, J. Jordon, and M. van der Schaar, "INVASE: Instance-wise variable selection using neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: [https://openreview.net/forum?id=BJg\\_roAcK7](https://openreview.net/forum?id=BJg_roAcK7)
- [71] J. H. Pope, T. P. Aufderheide, R. Ruthazer, R. H. Woolard, J. A. Feldman, J. R. Beshansky, J. L. Griffith, and H. P. Selker, "Missed diagnoses of acute cardiac ischemia in the emergency department," *New England Journal of Medicine*, vol. 342, no. 16, pp. 1163–1170, 2000, PMID: 10770981. [Online]. Available: <https://doi.org/10.1056/NEJM200004203421603>
- [72] E. A. Howell, "Reducing disparities in severe maternal morbidity and mortality," *Clinical obstetrics and gynecology*, vol. 61, no. 2, pp. 387–399, 2018. [Online]. Available: <https://doi.org/10.1097>
- [73] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT\* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 295–305. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3351095.3372852>
- [74] A. Esteva, B. Kuprel, and R. Novoa, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. [Online]. Available: <https://www.nature.com/articles/nature21056>

- [75] A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Deep learning interpretation of echocardiograms," *npj Digital Medicine*, vol. 3, 2020. [Online]. Available: <https://www.nature.com/articles/s41746-019-0216-8>
- [76] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 12 2016. [Online]. Available: <https://doi.org/10.1001/jama.2016.17216>
- [77] D. Muoio, "Fda permits marketing of ai software that autonomously detects diabetic retinopathy," *mobihealthnews.com*, Apr 2018. [Online]. Available: <https://www.mobihealthnews.com/content/fda-permits-marketing-ai-software-autonomously-detects-diabetic-retinopathy>
- [78] S. Deng, X. Zhang, W. Yan, E. I.-C. Chang, Y. Fan, M. Lai, and Y. Xu, "Deep learning in digital pathology image analysis: a survey," *Frontiers of medicine*, vol. 14, pp. 470–484, Aug 2020. [Online]. Available: <https://doi.org/10.1007/s11684-020-0782-9>
- [79] J. Xu, K. Xue, and K. Zhang, "Current status and future trends of clinical diagnoses via image-based deep learning," *Theranostics*, vol. 9, no. 25, pp. 7556–7565, 2019. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6831476/>
- [80] S. Manaktala and S. R. Claypool, "Evaluating the impact of a computerized surveillance algorithm and decision support system on sepsis mortality," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 88–95, 05 2016. [Online]. Available: <https://doi.org/10.1093/jamia/ocw056>
- [81] E. Strickland, "Hospitals roll out ai systems to keep patients from dying of sepsis," *spectrum.ieee.org*, Oct 2018. [Online]. Available: <https://spectrum.ieee.org/biomedical/diagnostics/hospitals-roll-out-ai-systems-to-keep-patients-from-dying-of-sepsis>
- [82] K. M. Corey, S. Kashyap, E. Lorenzi, S. A. Lagoo-Deenadayalan, K. Heller, K. Whalen, S. Balu, M. T. Heflin, S. R. McDonald, M. Swaminathan, and M. Sendak,

- “Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (pythia): A retrospective, single-site study,” *PLOS Medicine*, vol. 15, no. 11, pp. 1–19, 11 2018. [Online]. Available: <https://doi.org/10.1371/journal.pmed.1002701>
- [83] O. K. Nguyen, A. N. Makam, C. Clark, S. Zhang, B. Xie, F. Velasco, R. Amarasingham, and E. A. Halm, “Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison,” *Journal of Hospital Medicine*, vol. 11, no. 7, pp. 473–480, 2016. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jhm.2568>
- [84] J. Yoon, C. Davtyan, V. Der, and M. Schaar, “Discovery and clinical decision support for personalized healthcare,” *IEEE journal of biomedical and health informatics*, vol. 21, no. 4, pp. 1133–1145, 2017. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/27254875/>
- [85] M. K. Ross, J. Yoon, A. van der Schaar, and M. van der Schaar, “Discovering pediatric asthma phenotypes on the basis of response to controller medication using machine learning,” *Annals of the American Thoracic Society*, vol. 15, no. 1, pp. 49–58, 2018, PMID: 29048949. [Online]. Available: <https://doi.org/10.1513/AnnalsATS.201702-101OC>
- [86] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis,” *The Lancet Digital Health*, vol. 1, no. 6, pp. e271 – e297, 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2589750019301232>
- [87] Pickhardt, “Automated ct biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: a retrospective cohort study,” *The Lancet Digital Health*, vol. 2, no. 4, p. 1, 2020. [Online]. Available: [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(20\)30025-X/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30025-X/fulltext)
- [88] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, “Cardiovascular disease risk prediction using automated machine learning: A

- prospective study of 423,604 uk biobank participants," *PLOS ONE*, vol. 14, no. 5, pp. 1–17, 05 2019. [Online]. Available: <https://doi.org/10.1371/journal.pone.0213653>
- [89] M. Abroshan, A. M. Alaa, O. Rayner, and M. van der Schaar, "Opportunities for machine learning to transform care for people with cystic fibrosis," *Journal of Cystic Fibrosis*, vol. 19, no. 1, pp. 6–8, 2020. [Online]. Available: [https://www.cysticfibrosisjournal.com/article/S1569-1993\(20\)30008-4/fulltext](https://www.cysticfibrosisjournal.com/article/S1569-1993(20)30008-4/fulltext)
- [90] E. Cenko, M. van der Schaar, J. Yoon, O. Manfrini, Z. Vasiljevic, M. Vavlukis, S. Kedev, D. Milicic, L. Badimon, and R. Bugiardini, "Sex differences in acute heart failure and cardiovascular outcomes after myocardial infarction," *Circulation*, vol. 140, no. Suppl\_1, pp. A13 669–A13 669, 2019. [Online]. Available: <https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2677062>
- [91] E. Cenko, M. Van Der Schaar, J. Yoon, Z. Vasiljevic, S. Kedev, M. Vavlukis, M. Bergami, M. Scarpone, D. Milicic, O. Manfrini, L. Badimon, and R. Bugiardini, "P6419Machine learning in critical care: the role of diabetes and age in acute coronary syndromes," *European Heart Journal*, vol. 40, no. Supplement\_1, 10 2019, ehz746.1013. [Online]. Available: <https://doi.org/10.1093/eurheartj/ehz746.1013>
- [92] D. Jarrett, J. Yoon, and M. van der Schaar, "Dynamic prediction in clinical survival analysis using temporal convolutional networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 424–436, 2020. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/31331898/>
- [93] A. Alaa, J. Yoon, S. Hu, V. Der, and M. Schaar, "Personalized risk scoring for critical care prognosis using mixtures of gaussian processes," *IEEE transactions on bio-medical engineering*, vol. 65, no. 1, pp. 207–218, 2018. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/28463183/>
- [94] A. M. Alaa, K. H. Moon, W. Hsu, and M. van der Schaar, "Confidentcare: A clinical decision support system for personalized breast cancer screening," *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 1942–1955, 2016. [Online]. Available: <https://ieeexplore.ieee.org/document/7506232>
- [95] H. Thimbleby, "Improving safety in medical devices and systems," in *2013 IEEE International Conference on Healthcare Informatics*, 2013, pp. 1–13. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6680455>



- [96] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1893–1905, 2014. [Online]. Available: <https://ieeexplore.ieee.org/document/6868201>
- [97] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019. [Online]. Available: <https://science.sciencemag.org/content/363/6433/1287>
- [98] D. Sittig, A. Wright, E. Coiera, F. Magrabi, R. Ratwani, D. Bates, and H. Singh, "Current challenges in health information technology-related patient safety," *Health informatics journal*, vol. 26, no. 1, pp. 181–189, 2020. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/30537881/>
- [99] M. Sendak, "How machine learning is transforming clinical decision support tools," *Health IT Analytics*, Mar 2020. [Online]. Available: <https://healthitanalytics.com/features/how-machine-learning-is-transforming-clinical-decision-support-tools>
- [100] D. S. W. Ting, H. Lin, P. Ruamviboonsuk, T. Y. Wong, and D. A. Sim, "Artificial intelligence, the internet of things, and virtual clinics: ophthalmology at the digital translation forefront," *The Lancet Digital Health*, vol. 2, pp. e8 – e9, Sep 2020. [Online]. Available: [https://doi.org/10.1016/S2589-7500\(19\)30217-1](https://doi.org/10.1016/S2589-7500(19)30217-1)
- [101] A. Downey, "Nhs consultation on digital gps could stifle babylon's expansion," *digitalhealth.net*, Jul 2019. [Online]. Available: <https://www.digitalhealth.net/2019/07/nhse-consultation-digital-gp-babylon/>
- [102] —, "Manchester ccg objects to babylon expansion 'due to safety concerns'," *digitalhealth.net*, December 2019. [Online]. Available: <https://www.digitalhealth.net/2019/12/manchester-ccg-objects-to-babylon-expansion-due-to-safety-concerns/>
- [103] D. Oliver, "Lessons from the babylon health saga," *BMJ*, Jun 2019. [Online]. Available: <https://www.bmj.com/content/bmj/365/bmj.l2387.full.pdf>
- [104] Z. Obermeyer and E. J. Emanuel, "Predicting the future — big data, machine learning, and clinical medicine," *New England Journal of Medicine*,

- vol. 375, no. 13, pp. 1216–1219, 2016, pMID: 27682033. [Online]. Available: <https://doi.org/10.1056/NEJMp1606181>
- [105] C. Wallis, “How Artificial Intelligence Will Change Medicine,” *Nature*, vol. 576, no. 5, p. 48, Dec. 2019. [Online]. Available: <https://www.nature.com/articles/d41586-019-03845-1>
- [106] Amisha, M. P., M. Pathania, and V. K. Rathaur, “Overview of artificial intelligence in medicine,” *Journal of family medicine and primary care*, vol. 8, pp. 2328—2331, 2019. [Online]. Available: [https://doi.org/10.4103/jfmpc.jfmpc\\_440\\_19](https://doi.org/10.4103/jfmpc.jfmpc_440_19)
- [107] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, pp. 44–56, Jan 2019. [Online]. Available: <https://doi.org/10.1038/s41591-018-0300-7>
- [108] D. Bates, J. Teich, J. Lee, D. Seger, G. Kuperman, N. Ma’luf, D. Boyle, and L. Leape, “The impact of computerized physician order entry on medication error prevention,” *Journal of the American Medical Informatics Association : JAMIA*, vol. 6, no. 4, pp. 313–321, 1999. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC61372/>
- [109] C. Tekin, O. Atan, and M. Van Der Schaar, “Discover the expert: Context-adaptive expert selection for medical diagnosis,” *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 2, pp. 220–234, dec 2015. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6998045>
- [110] S. Davies, “Health 2040 – better health within reach,” NHS, Tech. Rep., 2018. [Online]. Available: <https://www.gov.uk/government/publications/chief-medical-officer-annual-report-2018-better-health-within-reach>
- [111] C. Willyard, “Can AI fix medical records?” *Nature*, vol. 576, no. 559–562, Dec. 2019. [Online]. Available: <https://www.nature.com/articles/d41586-019-03848-y>
- [112] J. Ross, C. Webb, and F. Rahman, “Artificial intelligence in healthcare,” Academy of Medical Royal Colleges, Tech. Rep., Jan 2019. [Online]. Available: [https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial\\_intelligence\\_in\\_healthcare\\_0119.pdf](https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf)

- [113] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artificial Intelligence in Medicine*, vol. 104, p. 101822, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0933365719310437>
- [114] S. Saria, A. Butte, and A. Sheikh, "Better medicine through machine learning: What's real, and what's artificial?" *PLoS Med*, vol. 15, no. 12, p. 1002721, 2018. [Online]. Available: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002721>
- [115] F. Magrabi, E. Ammenwerth, J. B. McNair, N. F. de Keizer, H. Hyppönen, P. Nykänen, M. Rigby, P. J. Scott, T. Vehko, Z. S.-Y. Wong, and A. Georgiou, "Artificial intelligence in clinical decision support: Challenges for evaluating ai and practical implications," *Yearbook of Medical Informatics*, vol. 28, pp. 128 – 134, 2019. [Online]. Available: <https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0039-1677903>
- [116] J. Powell, "Trust me, i'm a chatbot: How artificial intelligence in health care fails the turing test," *J Med Internet Res*, vol. 21, no. 10, p. e16222, Oct 2019. [Online]. Available: <http://www.jmir.org/2019/10/e16222/>
- [117] M. Sendak, M. C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, A. Bedoya, S. Balu, and C. O'Brien, ""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*, ser. FAT\* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 99–109. [Online]. Available: <https://doi.org/10.1145/3351095.3372827>
- [118] N. Summerton and M. Cansdale, "Artificial intelligence and diagnosis in general practice," *British Journal of General Practice*, vol. 69, no. 684, pp. 324–325, 2019. [Online]. Available: <https://bjgp.org/content/69/684/324>
- [119] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY,

- USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: <https://doi.org/10.1145/3313831.3376718>
- [120] M. B. A. McDermott, S. Wang, N. Marinsek, R. Ranganath, M. Ghassemi, and L. Foschini, “Reproducibility in machine learning for health,” *CoRR*, vol. abs/1907.01463, 2019. [Online]. Available: <http://arxiv.org/abs/1907.01463>
- [121] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, M. S. Board, L. Waldron, B. Wang, C. McIntosh, A. Kundaje, C. S. Greene, M. M. Hoffman, J. T. Leek, W. Huber, A. Brazma, J. Pineau, R. Tibshirani, T. Hastie, J. P. A. Ioannidis, J. Quackenbush, and H. J. W. L. Aerts, “The importance of transparency and reproducibility in artificial intelligence research,” 2020. [Online]. Available: <https://arxiv.org/abs/2003.00898>
- [122] D. C. Angus, “Randomized Clinical Trials of Artificial Intelligence,” *JAMA*, vol. 323, no. 11, pp. 1043–1045, 03 2020. [Online]. Available: <https://doi.org/10.1001/jama.2020.1039>
- [123] O. Atan, W. R. Zame, and M. van der Schaar, “Sequential patient recruitment and allocation for adaptive clinical trials,” in *Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 1891–1900. [Online]. Available: <http://proceedings.mlr.press/v89/atan19a.html>
- [124] I. Marshall, J. Kuiper, and B. Wallace, “Automating risk of bias assessment for clinical trials,” *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1406–1412, 2015. [Online]. Available: <https://ieeexplore.ieee.org/document/7104094>
- [125] O. Francon, S. Gonzalez, B. Hodjat, E. Meyerson, R. Miikkulainen, X. Qiu, and H. Shahrzad, “Effective reinforcement learning through evolutionary surrogate-assisted prescription,” *arxiv.com*, 2020. [Online]. Available: <https://arxiv.org/abs/2002.05368>
- [126] S. Davies, “Health, our global asset – partnering for progress,” NHS, Tech. Rep., 2019. [Online]. Available: <https://www.gov.uk/government/publications/chief-medical-officer-annual-report-2019-partnering-for-progress>

- [127] M. D. McCradden, "Ethical limitations of algorithmic fairness solutions in health care machine learning," *The Lancet Digital Health*, vol. 2, pp. e221 – e223, May 2020. [Online]. Available: [https://doi.org/10.1016/S2589-7500\(20\)30065-0](https://doi.org/10.1016/S2589-7500(20)30065-0)
- [128] D. M. Lloyd-Jones, L. T. Braun, C. E. Ndumele, S. C. Smith, L. S. Sperling, S. S. Virani, and R. S. Blumenthal, "Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease," *Journal of the American College of Cardiology*, vol. 73, no. 24, pp. 3153–3167, 2019. [Online]. Available: <https://www.jacc.org/doi/abs/10.1016/j.jacc.2018.11.005>
- [129] C. Andrade, "Understanding relative risk, odds ratio, and related terms: As simple as it can get," *The Journal of clinical psychiatry*, vol. 76, no. 07, pp. e857–e861, July 2015.
- [130] —, "The numbers needed to treat and harm (nnt, nnh) statistics: what they tell us and what they do not," *The Journal of clinical psychiatry*, vol. 76, no. 3, p. e330—3, March 2015. [Online]. Available: <https://doi.org/10.4088/JCP.15f09870>
- [131] D. Mosquera, N. Chiang, and R. Gibberd, "Evaluation of surgical performance using v-possum risk-adjusted mortality rates," *ANZ journal of surgery*, vol. 78, pp. 535–9, 08 2008.
- [132] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients," *Critical care medicine*, vol. 34, no. 5, pp. 1297–1310, 2006.
- [133] M. Hussain-Gambles, K. Atkin, and B. Leese, "Why ethnic minority groups are under-represented in clinical trials: a review of the literature," *Health & Social Care in the Community*, vol. 12, no. 5, pp. 382–388, 2004. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2524.2004.00507.x>
- [134] M. McCradden, M. Mazwi, S. Joshi, and J. A. Anderson, "When your only tool is a hammer: Ethical limitations of algorithmic fairness solutions in healthcare machine learning," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 109. [Online]. Available: <https://doi.org/10.1145/3375627.3375824>

- [135] A. A. Montgomery, T. Fahey, T. J. Peters, C. MacIntosh, and D. J. Sharp, "Evaluation of computer based clinical decision support system and risk chart for management of hypertension in primary care: randomised controlled trial," *BMJ*, vol. 320, no. 7236, pp. 686–690, 2000. [Online]. Available: <https://www.bmj.com/content/320/7236/686>
- [136] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Hdpm: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133 034–133 050, 2020.
- [137] P. Boddington, *Towards a Code of Ethics for Artificial Intelligence*, 1st ed., ser. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, 2017. [Online]. Available: <https://www.springer.com/gp/book/9783319606477>
- [138] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds & Machines*, vol. 28, no. 4, pp. 689–707, 2018. [Online]. Available: <https://link.springer.com/article/10.1007/s11023-018-9482-5>
- [139] European Commission , "White Paper on Artificial Intelligence: a European approach to excellence and trust," [ec.europa.eu](https://ec.europa.eu), 2019. [Online]. Available: [https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020\\_en.pdf](https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf)
- [140] AIHLEG, "Ethics guidelines for trustworthy ai," [ec.europa.eu](https://ec.europa.eu), April 2019. [Online]. Available: [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419)
- [141] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," Berkman Klein Center, Tech. Rep., 2020. [Online]. Available: <https://cyber.harvard.edu/publication/2020/principled-ai>
- [142] G. Iacobucci, "Babylon app will be properly regulated to ensure safety, government insists," *BMJ*, Jul 2018. [Online]. Available: <https://www.bmj.com/content/362/bmj.k3215/article-info>

- [143] E. Larosa and D. Danks, "Impacts on trust of healthcare ai," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 210–215. [Online]. Available: <https://dl.acm.org/doi/10.1145/3278721.3278771>
- [144] C. Paton and S. Kobayashi, "An open science approach to artificial intelligence in healthcare," *Yearb Med Inform*, pp. 047–051, aug 2019. [Online]. Available: <https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0039-1677898>
- [145] M. Ashoori and J. D. Weisz, "In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes," *arxiv.com*, 2019. [Online]. Available: <https://arxiv.org/abs/1912.02675>
- [146] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Muller, L. Ju, and H. Su, "Trust in automl," *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Mar 2020. [Online]. Available: <http://dx.doi.org/10.1145/3377325.3377501>
- [147] O. Lahav, N. Mastronarde, and M. van der Schaar, "What is interpretable? using machine learning to design interpretable decision-support systems," *arxiv.com*, 2018. [Online]. Available: <https://arxiv.org/abs/1811.10799>
- [148] S. M. Merritt, "Affective processes in human–automation interactions," *Human Factors*, vol. 53, no. 4, pp. 356–370, 2011, pMID: 21901933. [Online]. Available: <https://doi.org/10.1177/0018720811411912>
- [149] Sauf Pompier Limited. [Online]. Available: <https://www.mindmup.com/>
- [150] M. Madsen and S. Gregor, "Measuring human-computer trust," in *Proceedings of the 11th Australasian Conference on Information Systems*, vol. 53, 01 2000, pp. 6–8. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.93.3874>
- [151] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," 2019. [Online]. Available: <https://arxiv.org/abs/1812.04608v2>
- [152] W. J. Conover, "Rank tests for one sample, two samples, and  $k$  samples without the assumption of a continuous distribution function," *Ann. Statist.*, vol. 1, no. 6, pp. 1105–1125, 11 1973. [Online]. Available: <https://doi.org/10.1214/aos/1176342560>

- [153] A. R. Baggaley and A. L. Hull, "The effect of nonlinear transformations on a likert scale," *Evaluation and the health professions*, vol. 6, no. 4, pp. 483—491, 1983. [Online]. Available: <https://doi.org/10.1177/016327878300600408>
- [154] C. J. W. III, "Dispelling three myths about likert scales in communication trait research," *Communication Research Reports*, vol. 30, no. 4, pp. 366–372, 2013. [Online]. Available: <https://doi.org/10.1080/08824096.2013.836937>
- [155] P. A. Bishop and H. R. L, "Use and misuse of the likert item responses and other ordinal measures," *International journal of exercise science*, vol. 8, no. 3, pp. 297–302, Jul. 2015. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833473/>
- [156] P. Kero and D. Lee, "Likert is pronounced "lick-urt" not "lie-kurt" and the data are ordinal not interval." *Journal of applied measurement*, vol. 17, pp. 502–509, 2016.
- [157] I. D. Cooper and T. P. Johnson, "How to use survey results," *Journal of the Medical Library Association : JMLA*, vol. 104, pp. 174–177, Apr. [Online]. Available: 10.3163/1536-5050.104.2.016
- [158] O. Filonik and S. Winters, "Semantic priming effect on survey results," *Scientific Notes of Ostroh Academy National University: Philology Series*, no. 9(77), pp. 81–83, Jul. 2020. [Online]. Available: <https://journals.oa.edu.ua/Philology/article/view/2819>
- [159] A. Clifford, A. Holmes, I. R. Davies, and A. Franklin, "Color categories affect pre-attentive color perception," *Biological Psychology*, vol. 85, no. 2, pp. 275 – 282, 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0301051110002139>
- [160] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 6th ed. Pearson, 2016. [Online]. Available: <https://dl.acm.org/doi/book/10.5555/3033040>
- [161] N. Scott and M. F. Weiner, ""patientspeak": An exercise in communication." *Journal of Medical Education*, vol. 59, no. 11, pp. 890–893, 1984. [Online]. Available: <https://psycnet.apa.org/record/1985-26258-001>