Safety Science 155 (2022) 105870

Contents lists available at ScienceDirect

Safety Science

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

Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare

Mark A. Sujan [a,*], Sean White [b], Ibrahim Habli [c], Nick Reynolds [d]

[a] Human Factors Everywhere, Woking, UK
[b] NHS Digital, Leeds, UK
[c] University of York, York, UK
[d] University Hospitals of Derby and Burton NHS Foundation Trust, Derby, UK

ARTICLE INFO

Keywords:
Artificial intelligence
Safety
Healthcare
Attitudes
Innovation

ABSTRACT

Introduction: There is an increasing number of healthcare AI applications in development or already in use. However, the safety impact of using AI in healthcare is largely unknown. In this paper we explore how different stakeholders (patients, hospital staff, technology developers, regulators) think about safety and safety assurance of healthcare AI.

Methods: 26 interviews were undertaken with patients, hospital staff, technology developers and regulators to explore their perceptions on the safety and the safety assurance of AI in healthcare using the example of an AI-based infusion pump in the intensive care unit. Data were analysed using thematic analysis.

Results: Participant perceptions related to: the potential impact of healthcare AI, requirements for human-AI interaction, safety assurance practices and regulatory frameworks for AI and the gaps that exist, and how incidents involving AI should be managed.

Conclusion: The description of a diversity of views can support responsible innovation and adoption of such technologies in healthcare. Safety and assurance of healthcare AI need to be based on a systems approach that expands the current technology-centric focus. Lessons can be learned from the experiences with highly automated systems across safety-critical industries, but issues such as the impact of AI on the relationship between patients and their clinicians require greater consideration. Existing standards and best practices for the design and assurance of systems should be followed, but there is a need for greater awareness of these among technology developers. In addition, wider ethical, legal, and societal implications of the use of AI in healthcare need to be addressed.

1. Introduction

The use of artificial intelligence (AI) in healthcare is regarded a priority in national health policies to address challenges such as the COVID-19 pandemic as well as rising healthcare costs, staff shortages and burnout, and an increasingly elderly population with more complex health needs (Peek et al., 2020; Joshi and Morley, 2019). During 2015–2020 over 200 medical devices using machine learning (a type of AI) received regulatory approval in Europe and the US (Muehlematter et al., 2021). While most healthcare AI applications have been developed in diagnostics (e.g., breast cancer screening (McKinney et al., 2020), it is likely that all areas of healthcare will see the introduction of AI tools, e.g., in ambulance service triage (Blomberg et al., 2019), sepsis management (Komorowski et al., 2018), palliative care (Avati et al., 2018) and mental health (Fitzpatrick et al., 2017).

However, the impact of using AI in healthcare, and especially the safety impact, is largely unknown. Most healthcare AI applications have been evaluated retrospectively only, and the evidence base remains weak and at high risk of bias (Nagendran et al., 2020; Wu et al., 2021). Examples of prospective, real-world evaluations are few, and generally seem to conclude that the overall performance of the joint human – AI system does not improve to the extent suggested by retrospective studies (Beede et al., 2020; Blomberg et al., 2021). In addition, prospective evaluation can also help uncover unintended and unanticipated

* Corresponding author.
E-mail addresses: mark.sujan@humanfactorseverywhere.com (M.A. Sujan), sean.white@nhs.net (S. White), ibrahim.habli@york.ac.uk (I. Habli), nick.reynolds1@nhs.net (N. Reynolds).

https://doi.org/10.1016/j.ssci.2022.105870
Received 16 December 2021; Received in revised form 24 June 2022; Accepted 4 July 2022
Available online 9 July 2022

consequences of using healthcare AI (Cabitza et al., 2017).

Development and retrospective evaluation of healthcare AI are concerned predominantly with technical issues such as algorithm accuracy, data quality and the potential for bias in the data (Challen et al., 2019). While these issues are important, by themselves they are not sufficient to ensure that the use of AI in healthcare settings is safe, where socio-technical issues, such as trust, skill erosion, impact on workload and working practices, as well as ethical concerns around fairness, health equity and the wider societal impact are going to be critical aspects (Sujan et al., 2019; Sikstrom et al., 2022; Wawira Gichoya et al., 2021; Smallman, 2019).

The field of Science and Technology Studies (STS) suggests that the development, governance, and deployment of novel technologies should be studied as interacting socio-technical processes rather than as technical activities in isolation. Disruptive innovations, such as healthcare AI, inevitably create gaps, and they challenge established social and organisational structures and hierarchies, which must be bridged and repaired as part of the innovation process (Elish and Watkins, 2020). Assurance practices and regulatory frameworks for healthcare AI are still being developed, e.g., action plans and change programmes developed by the US Food and Drug Administration (FDA) and the UK Medicines and Healthcare Products Regulatory Agency (MHRA) around the use of AI in software as a medical device. The notion of responsible (research and) innovation recognises the difficulty and uncertainty of predicting the societal impact of such rapidly developing technological innovation and suggests embedding debate about societal concerns in the innovation process through anticipation, reflexivity, inclusion, and responsiveness (Stilgoe et al., 2013). Central to this is fostering a diversity of views and engaging in a broad dialogue with stakeholders about the risks, benefits and acceptability of healthcare AI (Macrae, 2019). Such inclusive dialogue can support anticipation, e.g., through identification of patient concerns about the impact of healthcare AI beyond the consequences of technical failures, and it can provide an opportunity for stakeholders to reflect on their respective assumptions, motivations, and priorities.

In line with this thinking, in this paper we explore how different stakeholders (patients, hospital staff, technology developers, regulators) think about safety and safety assurance of healthcare AI, using the example of AI-based intravenous infusion (IV) pumps within an intensive care unit (ICU) setting. There are an estimated 237 million medication errors in the UK NHS every year (Elliott et al., 2018), and studies have found that medication administration errors are five times more likely in IV doses than in non-IV doses (McDowell et al., 2010; McLeod et al., 2013). The use of AI might lead to the development of highly automated and partially autonomous infusion pumps with the potential to reduce errors and improve patient outcomes (Sujan et al., 2019).

The detailed objectives of this study were to describe stakeholder perceptions on: (1) the potential impact of using AI-based infusion pumps in intensive care with respect to perceived advantages and disadvantages, patient experience and working practices; and (2) the potential safety assurance practices for AI-based infusion pumps, approaches to regulation and incident investigation, and the gaps that exist. This rich description of perceptions from a diverse range of stakeholders can complement the current technological focus on healthcare AI to inform a more socio-technical and systems-based approach to AI safety.

2. Methods

2.1. Setting and clinical reference scenario

The organisation participating in this study was an English NHS hospital. The hospital serves a population of 600,000. It has a capacity of 1131 beds, and it employs over 8800 staff. The ICU within the hospital has 16 beds, and is staffed by approximately 35 medical staff, 100 nurses and 80 support staff. The ICU cares for 1300 patients annually.

The study focus was the use of AI to support intravenous medication management in ICU. The technology can carry out safety checks automatically, and it could potentially administer certain types of IV drug infusions autonomously (i.e., make changes to the infusion within certain hard limits or stop the infusion). Patients in ICU typically are very ill, and they can receive 6–12 infusions concurrently.

To focus data collection with study participants, a scenario was developed. Scenarios have been defined as "descriptions of possible futures that reflect different perspectives on the past, present and the future" (van Notten et al., 2003). Scenarios have been used in different fields, e.g., in Science and Technology Studies of emerging technologies (Selin, 2011), and in Computer Supported Collaborative Work (Bødker, 2000). The use of scenarios is intended to encourage a diversity of views and to enable broader participation in design. The scenario developed as part of this study concerned the management of a patient's blood sugar level via rapidly acting insulin administered intravenously through an infusion pump. The rationale for focusing on insulin administration was twofold: first, the management of blood sugar levels is a topic that is intuitive and widely understood even by non-clinical audiences; and second, the clinical protocol for the management of blood sugar levels is reasonably straightforward. The project team developed a clinical

Table 1
Clinical reference case.

Diabetic patient on ICU with sepsis requiring insulin – current, manual base line scenario	The patient is a 68-year old type-II diabetic with sepsis secondary to pneumonia. The patient's blood sugars require insulin control via rapidly acting IV insulin infusion. Patient identity, nurse identity, prescription and syringe formulation checks are all done manually (approx. 70 pieces of information). If the checks match, the nurse enters the pump programme and infusion limits, loads the syringe and starts the infusion pump by a manual button push. The nurse regularly reviews test results of blood sugar level, and adjusts the infusion rate as indicated by the clinical protocol when required.
Automated infusion technology	Patient identity, nurse identity, prescription and syringe formulation checks are all done by barcode. If checks match, the pump automatically programmes itself to start the infusion, displays medication identity and selects hard and soft programme infusion rate limits without further or final human confirmation. The nurse is only asked to intervene if factors known to the system (e.g., concurrent steroid prescribing, known previous insulin resistance from GP prescribing) may make the pre-programmed protocol likely to be ineffective based on this known information.
Autonomous infusion technology	The IV medication management system is further extended to allow it to select the best fit clinical protocol based on the known information about the patient in the information system. The pump controls the IV infusion rate of insulin in response to continuously measured blood sugar from a central venous sampling device. Within the pre-set limits it can adapt to the patient's actual insulin requirements and formulate an individualised protocol for the infusion rate based on the sugar readings to optimise sugars control through pre-emptive changes in infusion rates.

reference case with different levels of AI support, see Table 1.

2.2. Participants

Participants were patients, hospital staff, technology developers and regulators (including individuals involved in standardisation activities). Hospital staff were employees at the study hospital. Patients had received treatment at the study hospital. Other participants (technology developers and regulators) were independent of the study hospital. The sampling strategy aimed to recruit a diverse set of participants and roles, while being mindful of the practical constraints of the study. Data saturation was not an objective, in line with recent suggestions for thematic analysis (Braun and Clarke, 2021). In total, 26 participants were recruited. Table 2 provides a breakdown of participants by stakeholder group.

Patients and hospital staff were approached by the hospital-based clinical Principal Investigator. Regulators and technology developers were approached by the study Chief Investigator. Prior to the interview, potential participants received a participant information leaflet. Interviews took place in a meeting room at the hospital (patients and hospital staff), over the telephone or at the business offices of the interview participant. Participation was voluntary, and all participants provided written consent. Treatment in intensive care can be a very traumatic experience, and patients might find it emotionally challenging to discuss this. Support mechanisms were put in place at the hospital in case a patient participant might become distressed during the interviews. The support mechanism was used once by a patient participant and included use of the hospital counselling service and follow-up by a senior nurse known to the patient participant.

2.3. Data collection and data analysis

Data were collected through semi-structured interviews during May 2019–November 2019. Patient and hospital staff interviews were undertaken by two project team members, and technology developers and regulators interviews by one project team member. Interviews were semi-structured based on the topic guide shown in Table 3. Each interview lasted between 20 and 45 min. Interviews were audio-recorded. The audio recordings were subsequently transcribed, and during the transcription process all identifiers were removed to ensure anonymity.

Table 2
Interview participants by stakeholder group.

Stakeholder Group	Description	Participant ID
Intensive Care Patient	Sepsis	Patient-01
	Pancreatitis	Patient-02
	Diabetes	Patient-03
	Emergency surgery	Patient-04
Hospital Staff	ICU staff nurse	Staff-01
	ICU senior nursing staff	Staff-02
	Education	Staff-03
	ICU staff nurse	Staff-04
	Clinical pharmacist	Staff-05
	Anaesthetic trainee	Staff-06
	Education	Staff-07
	Business manager	Staff-08
	ICU senior nursing staff	Staff-09
	IT management	Staff-10
	Equipment management	Staff-11
	Education	Staff-12
	Clinical pharmacist	Staff-13
Technology Developers	Medical devices	Tech-01
	AI	Tech-02
	Medical devices	Tech-03
	Medical devices	Tech-04
	AI	Tech-05
Regulators	Health standards	Regulator-01
	Products certification	Regulator-02
	Health standards	Regulator-03
	Medical devices regulation	Regulator-04

Table 3
Topic guide for semi-structured interviews.

Introduction	Background to the study and the interview.
Participant background	Interviewee's professional background and current role, or their experience as a patient.
Impact on patient experience	Interviewee's perception of how the use of AI-based infusion pumps in intensive care might impact on patient experience.
Impact on working practices	Interviewee's perceptions of how working practices might be affected / changed when using AI-based infusion pumps.
Safety assurance	Safety assurance needs of interviewee, and suggestions for safety assurance activities of AI-based infusion pumps to meet these needs; description of gaps of existing assurance practice.
Regulation	Suggestions for regulatory practice for AI-based infusion pumps; description of gaps in current regulatory practice.
Incident investigation	Perceptions of how incidents involving AI-based systems should be investigated.
Ending	Expression of thanks for contribution

Data collection and analysis followed a sequential approach. All data were collected first and were then analysed in a subsequent step. Following transcription of the interview data, thematic analysis was undertaken (Braun and Clarke, 2006). The approach to thematic analysis used was towards the experiential and semantic end of the analytic spectrum (as opposed to the critical and latent), i.e., the focus was more on explicitly expressed meaning rather than implicit and conceptual level of meaning (Braun and Clarke, 2022). In a first step all interviews were read to allow familiarisation with the data. Subsequently, each interview was coded using descriptive and open coding (Saldaña, 2009). An analytic memo was produced for each interview summarising the researcher's thoughts and issues of particular interest. An analytic memo can be unstructured, and it is a tool to support reflection during the analysis. Using the initial set of codes and the analytic memos, themes were identified through clustering of codes in meetings of the project team. The high-level themes followed largely from the structure of the interview guide. Within each theme, sub-themes were developed by listing all codes within the theme, reviewing the actual data behind each code, and by using (and manipulating) visual representations (mind maps) to facilitate discussion within the project team to identify relationships and determine importance. These are subjective and interpretative analytic activities, i.e., they represent choices made by the project team.

2.4. Ethics

The study received institutional approval at the participating NHS hospital as a service evaluation study.

3. Results

Results are presented using descriptive themes following from the interview guide: (1) advantages, disadvantages, and impact on patient experience; (2) human – AI interaction; (3) safety assurance activities; (4) regulation; and (5) incident investigation. The description below presents rich detail, and a summary is provided in Table 4 for convenience.

3.1. Advantages, disadvantages, and impact on patient experience

3.1.1. Attitudes towards AI

Some of the hospital staff and patient participants expressed positive generic attitudes towards AI and technology, sometimes likening the situation to the release of new technology such as an iPhone or the use of robotics in car manufacturing. There were no negative generic views, but participants acknowledged that there might be a generational

Table 4
Summary of stakeholder perceptions.

Theme 1: Advantages, disadvantages, and impact on patient experience	Attitudes towards AI are positive and are based on trust in the health system. AI can increase efficiency and reduce errors, but it can also contribute to delays and errors. There is still a need for human contact, and the use of AI systems should not disrupt the relationship between patients and clinicians.
Theme 2: Human – AI interaction	Training needs to enable clinicians maintain core clinical skills, and it needs to help clinicians build a baseline understanding of AI and its limitations. Clinicians in intensive care have a strong sense of autonomy. Clinicians need to build trust in AI. Feedback and alerts can provide clinicians with an awareness of what the AI is doing.
Theme 3: Safety assurance practices	Existing assurance practices are a good starting point for safety assurance of AI in clinical settings. AI evolution poses new challenges, but might be addressed through real-time monitoring and continuous feedback, or by locking algorithms. The use of synthetic data could complement real-world data to provide more comprehensive training data sets.
Theme 4: Regulation	Existing safety standards for medical devices are a good starting point for the regulation of AI in clinical settings. Regulation requires a culture change to deal with AI evolution. A more iterative approach to regulation will be required. Developers need to demonstrate they have competence and expertise in developing safe AI. Developers and regulators need to establish a dialogue.
Theme 5: Incident investigation	AI systems can enhance traceability and auditability. However, responsibility and accountability for incidents might be pushed onto clinicians. The incident investigation process needs to include additional actors such as AI experts and AI developers.

divide, with younger people potentially being more open to new forms of technology compared with older people. Other studies also report positive attitudes towards healthcare AI, which tend to be more pronounced where the use of AI is geared towards process automation rather than towards direct patient care (Scott et al., 2021).

However, participants often qualified their generic positive attitude towards AI by expressing certain underlying beliefs, such as if a technology is being used, they assume it must be safe and help their treatment. Trust is one such underlying belief, and one patient described the trust that people have in the NHS as an institution, and that as a result they tend to be more accepting of whatever form of technology is proposed. In this instance, trust is based on professional and institutional authority, and it is viewed as independent of the process of technological innovation itself. Despite the often very public discussion of the scale of patient harm, trust in healthcare professionals remains high, even if distrust in health systems as a whole to deliver high-quality care in the future is rising (Calnan and Sanford, 2004).

"I think some people will look at anything the NHS says as like okay, we can trust that, the NHS have said it. So, I think there is that unspoken kind of trust. A lot of people just think, oh well, if the doctors and the nurses and the NHS, the brand has said this is okay, it's a change but it's okay and these are the reasons why." Patient-04.

3.1.2. Efficiency and accuracy

Within all participant groups there were views that acknowledged the potential of AI to reduce error and to make care safer. Intensive care can be a very demanding and stressful environment, and clinicians can become tired or get distracted, and as a consequence they might forget to do something or make mistakes, such as drug dose calculation errors (Lyons et al., 2018). A perceived benefit of AI is that it would not be subject to such performance influencing factors, and it could reduce the opportunities for human error. Examples provided included the checking of drugs and drug doses, the reduction of data entry and data entry errors, and the provision of alerts of contraindications. The benefits might be particularly relevant in patients that require a lot of attention and receive several medications and interventions concurrently, as such situations are especially demanding. Similar expectations were held for previous generations of clinical decision support systems, as well as more broadly for highly automated systems across different industries, but numerous studies as well as accident investigations demonstrated that the assumption that automation reduces human error and thereby improves safety is overly simplistic (Cabitza et al., 2017; Bainbridge, 1983; Alberdi et al., 2004).

However, some participants also pointed out that AI has the potential to cause or contribute to patient harm. This might be because these technologies are still very novel and hence largely untested in clinical environments (Nagendran et al., 2020; Wu et al., 2021). In addition, some hospital staff pointed out that the existing infusion pumps sometimes failed, and they suggested that AI-based infusion pumps would not be perfect either. Furthermore, AI-based infusion pumps might have only a limited view onto the patient, their treatment and the wider clinical context. This limited contextual awareness might be inappropriate at times and could contribute to errors and sub-optimal treatment. Such unintended consequences of the introduction of technology is not limited to AI but has been documented previously, for example with the introduction of health information technology, such as electronic health records, computer-assisted physician order entry and smart infusion pumps (Ash et al., 2007; Black et al., 2011; Koppel et al., 2005; Sujan and Habli, 2021).

A nurse from the ICU described how human decision-making about care and treatments is not just based on a limited set of numeric indicators or variables. The nurse emphasised the importance of context, and the holistic aspect of clinical decision-making, which might not be available to an autonomous system. This could potentially affect patient care and patient outcomes.

"With regards to things automatically going up and down, sometimes it isn't just numbers. It's what you're looking at. Does that make sense? It's not just about numbers. It's what you're looking at […] Sometimes looking at a patient you just know what they need. You don't really need anything else to tell you. You can look at a patient and you can think, oh, God, this is not going to be a good day. This is going to end badly. You know just by looking at somebody, which obviously can't be quantified." Staff-01.

A key anticipated benefit of the use of AI in intensive care is improved efficiency of care and of clinical processes. By taking over specific tasks from nurses and doctors, AI can free up clinicians' time, and allow them to focus on other aspects of patient care. Several participants commented on this, suggesting that the use of AI would streamline processes, reduce the need for manual interventions, and provide more timely care.

However, some staff with direct patient contact questioned whether AI would improve efficiency and free up their time. For example, experiences with current health IT systems were not universally positive because these systems were sometimes very slow to respond, and it might take several minutes to enter data. More generally, the introduction of novel technology into a healthcare setting inevitably changes processes and can affect the existing professional and organisational structures. In order for novel technologies to work effectively in real-world clinical settings, such disruptions need to be accompanied by "repair" work to create a new set of practices and structures (Elish and Watkins, 2020). AI systems might have to communicate with other health IT systems, and such delays might need to be sorted out and dealt with by clinicians, thereby creating additional work. Also, clinicians

typically multi-task and make use of manual interventions to build up situation awareness and a holistic picture of the patient's care and the wider clinical system.

An ICU nurse challenged the view that the use of AI systems would free up significant amounts of nursing time, because changing settings manually on the existing pumps was not done in isolation. The nurse described how such interactions provided an opportunity to talk to the patient, look at the patient, and generally maintain an awareness of the overall situation. The nurse also alludes to wider implications such as the relationship and interaction with the patient's relatives.

"I don't spend ten minutes just looking at my pumps. You do it as part of your assessment […] So while I'm doing something with the patient, I'll just be checking the pumps and looking and make decisions, and then I don't spend a specific amount of time. It would probably only free up seconds, to go to the pump to alter it. That's all it would be. But then at least, when the family are looking at you, you are still doing something, whereas I feel like, if you weren't doing that, they'd be like, well, what are you doing?" Staff-01.

3.1.3. Relationship between patients and their clinicians

Even though patients expressed generally positive attitudes towards the use of AI, they also pointed out the need for human contact. All four patients suggested that they appreciated the companionship of nurses and the support they provide with needs other than purely medical. While the use of AI could improve efficiency by taking over tasks from clinicians, there is also a danger that clinicians might get pulled into other activities instead. These activities could potentially be away from the patient's bed side as less manual interaction is required. This has been demonstrated in studies of electronic health records, where it has been suggested that physicians now spend more time on data entry activities than on direct patient care (Hill et al., 2013). In this case, the introduction of AI might not lead to more humanised care, but rather leave patients feeling more isolated.

A patient described this trade-off of knowing, on the one hand, that the AI systems might be taking care of their medical needs, and, on the other hand, of having the comforting presence of a nurse.

"People will still want a human contact. Machines might be able to do a damned good job, but I think people still need that little human contact. All right, while it's absolutely computerised and what's it and then people… I would feel lost a bit. I'd know they were doing all right, but I'd still need a little human… I don't think I could live with the machines continuously, although I know they were doing a good job." Patient-01 .

3.2. Human – AI interaction

3.2.1. Training

Clinician participants emphasised the need to retain core clinical skills. With the introduction of AI systems, such as partially autonomous AI-based infusion pumps, a range of tasks currently undertaken by clinicians would routinely be taken over by the AI. This includes, for example, dose calculations, taking and interpreting blood sugar level tests, and adjustments to IV infusions. Participants suggested that in the past when new systems were introduced, they observed a deterioration of skills among their colleagues. As a result, there is a danger of de-skilling clinical staff, and hence people might not be able to take over from the system in case of any failures or unavailability of the system. The potential negative impact on worker skills, skill retention, and the ability to take back control from an automated systems has not only been highlighted in healthcare (Cabitza et al., 2017) but is one of the classic "ironies of automation" (Bainbridge, 1983). Consideration of the potential for skill degradation has been proposed as one of the primary evaluative criteria for the design of automated systems (Parasuraman et al., 2000).

In addition to maintaining core clinical skills, participants widely agreed that staff require education about AI. Participants alluded to the potential transformation of staff roles from being users of an infusion pump to that of supervisor of an autonomous AI system. Monitoring and supervising an autonomous AI system requires a different skill set than simply using a passive system. When hospitals rely on staff to pick up problems with the AI, staff need to understand how to identify any issues, and they need to demonstrate competence in this. Different forms of real-time assistance and explanation, e.g., showing AI predictions alongside explanations and indications of AI confidence (Lai and Tan, 2019), as well as offline training using machine learning model driven tutorials and explanations can significantly improve users' understanding of the behaviour of the AI and overall task performance (Lai et al., 2020).

One of the technology developers likened this transformed role of clinical staff to control room operations in other sectors, such as the nuclear industry. The role of staff would be about monitoring and providing oversight rather than using the AI system.

"It might be more about checking that the system is configured correctly, more about monitoring the system, rather than a user-directed activity. Normally users are driving the system, and the use is much more intuitive. If it's autonomous, the mechanisms of training might be the same, but what you're training them to do is quite different. It's almost like training someone to monitor a nuclear power plant, in a control room." Tech-03.

Participants highlighted the dangers of becoming over-reliant on AI. Becoming too reliant on an AI system could lead to staff not being able to do tasks manually anymore in case of failure with the AI. Over-reliance on automation, or automation-induced complacency (Parasuraman and Riley, 1997), has been considered a contributory factor to accidents, such as the fatal collision between an Uber / Volvo self-driving test vehicle and a pedestrian in 2018. The notion of over-reliance as a contributory factor is contested because it narrowly focuses attention on the individual rather than the wider socio-technical, economic and political context (Stanton et al., 2019).

A participant from the hospital raised the issue of over-reliance and complacency when infusion pumps become truly autonomous. The participant expressed concern about whether staff might just take what the infusion pump reports at face value, without attempting to critically question it.

"My worry is, if you, as I've probably already alluded to… I think if you make something 100% automated, and there is a machine that's attached to the patient that says it's doing this, I worry that people will become complacent, trust it too much, whatever. But they'll just look at other pumps and think it's fine, and it's doing its job, sorted." Staff-03.

3.2.2. Autonomy, control, and trust

Clinician participants expressed a strong sentiment of being autonomous practitioners on intensive care. Nurses often spend their entire shift with a single patient, and they build up a strong relationship and bond with their patient. As a result, nurses feel a responsibility towards their patient, and they have a desire to be in control of the patient's treatment. With the introduction of AI systems, this sense of autonomy and control could be challenged, and this might cause anxiety and mixed feelings towards the technology. Participants suggested that options for manual overrides could contribute to their feeling of being in control and avoid confusion and potential complications in the delivery of care.

Participants also suggested that having alerts and feedback would help to maintain situation awareness and stay in control of the overall treatment and care for the patient. Situation awareness refers to the dynamic understanding of an ongoing situation, involving perception of relevant data and cues, integration of these features to comprehend their meaning, and projection into likely future states (Endsley, 1995). Distributed Situation Awareness represents a systems approach, which emphasises that situation awareness is distributed throughout a system and that it is built through interactions between actors (people as well as automation or AI) (Stanton et al., 2006). Alerts can help to raise awareness of a deteriorating situation so that clinicians can take over control to prevent patient harm. However, while an individual alert might appear a reasonable design feature, problems can occur when multiple alerts are raised concurrently or when there are frequent alerts

throughout a shift. Alarm fatigue, i.e., the delayed response or reduced response frequency to alarms, has been identified in major industrial accidents, such as the 1994 explosion and fires at the Texaco Milford Haven refinery. In intensive care it has been suggested that a healthcare professional can be exposed to over 1,000 alarms per shift, contributing to alarm fatigue, disruption of care processes and noise pollution, with potentially adverse effects on patient safety (Ruskin and Hueske-Kraus, 2015).

One participant identified as problematic how the AI would determine that it was doing something wrong, and how it would identify when to trigger an alert. The answer to this is most likely not straightforward as illustrated by the example given by the participant.

"Say you were planning on giving one to two ml of insulin an hour, and you have a regime that goes up to four, or something. And you're up at four, and the sugar's still 25 and it's not coming down after 15 min. Does it [the AI-based infusion pump] then go up to six? Does it go up to eight? Does it go up to 12? At what point does it say, there's something really not happening right here?" Staff-03.

Clinicians have their own mental model of the patient, their needs and the treatment. If the AI operates on a different model, or if it does not communicate what it is doing, then clinicians might feel out of the loop. Continuous feedback could help clinicians build situation awareness and gain trust in the AI. Participants suggested that to them it was important to know what the AI was doing, and that it would be looking for or checking the same things as they themselves would do. How such feedback could be designed and provided is far from straightforward, especially for AI that uses machine learning algorithms, which produce complex models such as Deep Learning. Many approaches to explainable AI focus on providing detailed accounts of how an algorithm operates (Miller, 2019), but for explanations to be useful they need to be able to accommodate and be responsive to the needs of different users across a range of situations, e.g., a patient might benefit from a different type of explanation compared with a healthcare professional. In this sense, rather than providing a description of a specific decision, explanation might be better regarded as a social process and a dialogue that allows the user to explore AI decision-making by interacting with the AI and by interrogating AI decisions (Weld and Bansal, 2019). Real-time feedback could also be complemented by offline coaching (as discussed above) (Lai et al., 2020).

Feedback and alerts are also important to build trust in the AI. Trust is important, because otherwise clinicians might not use AI, work against it or otherwise the intended benefits might not be realised. Some of the technology developers raised concerns that people might not trust AI, because they are unfamiliar with it and don't understand how it works. The way AI is represented in popular culture, e.g., as machines taking over or as artificial humans ("The Terminator" is a classic example) or as super-intelligent but ultimately destructive Artificial General Intelligence (such as HAL 9000 in "2001: A Space Odyssey"), might contribute to fuelling fears (Russell, 2019). This points to other avenues besides real-time feedback and explanation for considering trust. Trust can be conceptualised as arising from the socio-technical engagements between different stakeholders (including patients, clinicians, and technology developers) in the development and validation process of the AI (Elish, 2018). In this view, trust is not a property of the AI as a product, but arises from "trust work", i.e., how stakeholders interact, communicate, and make sense of the technology and their practice (Winter et al., 2022).

3.3. Safety assurance activities

3.3.1. Existing assurance practices

Several participants suggested that existing safety assurance practices would remain relevant for AI systems in clinical environments, and that such safety assurance practices might provide an excellent starting point. Participants referred to existing assurance practices for medical devices such as having independent review, being clear about intended use and limitations of the system, ensuring traceability and audit, working with domain experts, and creating a safety argument. This should be underpinned by processes that demonstrate that the developer is a competent (with respect to systems development) organisation.

Even though participants recognised limitations of existing approaches regarding AI systems, there was no disagreement about the value of retaining existing good practices. One of the technology developers suggested that in many respects AI systems should not be treated any differently from other medical devices, with the need for rigorous approaches, sound safety evidence and independent oversight.

"You need a robust design process. Autonomous systems shouldn't really be very different […] In reality AI is not much different, and it should be treated in the same way. There shouldn't be a different approach. The procedures that apply to non-AI should apply here as well […] The evidence should be the same for AI systems as for standard system. […] An independent level of assurance needs to be provided by regulators and notified bodies, but this could be the same as for current medical devices." Technology-02.

3.3.2. Managing AI evolution

Existing safety assurance practices have been developed to deal with mostly static and stable systems. With the move towards learning and evolving systems, it becomes necessary to manage this evolution in a safe way. Participants suggested that there need to be continuous monitoring and appropriate feedback mechanisms in place that allow developers of AI systems to track and monitor that the AI system is not violating any assumptions or its safety envelope. In addition, providing developers with a wider range of contextual clinical data could help improve the performance of algorithms.

However, all of this would fundamentally change the relationship between developers and users. Participants suggested that at present, developers and manufacturers of medical devices tend to have rather sporadic contact with their users, and often only in response to operational problems. Development, procurement and safety assurance processes between manufacturers and users (e.g., hospitals) do not appear to be linked up well, in part because manufacturers often target multinational markets, which might operate to a diverse range of standards and processes.

There are political, ethical and financial issues that require resolution if manufacturers and users are to share operational data more extensively. One of the technology developers raised the issue that, on the one hand, it would be useful to have access to data to improve the systems, but that, on the other hand, there need to be contractual mechanisms in place, and ownership and payment for data provision would need to be settled. Clinical data will increasingly become a valuable commodity, and it is very much an open question who owns it, who can access it (e.g., to improve systems), and who can distribute it.

"There's a heritage of building to contract and delivering it, but in the new world of AI there might be value of gathering data, [we] would need an arrangement with clients, to get better systems out of it. Things like, if the system has not behaved properly, we'd want to know about that. Arrangements would need to be worked out and formalised contractually. Who is paying for the service [of collecting and sharing data to improve systems]? Do we pay for it?" Technology-03.

One approach of managing AI evolution might be to lock down algorithms, and simply not allow evolution. The operational data could be used to train an offline version of the algorithm, which could be released subsequently after extensive testing. However, this might not be a suitable solution for all cases, e.g., in situations where insufficiencies in the AI behaviour were detected or where fast learning and adaption is required to deal with novel situations.

3.3.3. Data

Participants identified many potential issues and problems with data that might not have been present to the same extent in traditional systems. For example, bias in the training data, which is unrepresentative of the real world or the specific clinical setting can cause patient harm.

Similarly, participants with a clinical background expressed concern about the breadth and complexity of scenarios that an autonomous infusion pump might have to consider.

One of the participants from the hospital explained the potential complexity of the scenario of a diabetic patient with sepsis in intensive care. Every patient has a unique physiology and responds differently to interventions, and patients might have a multitude of underlying conditions all potentially interacting and affecting both treatment choices and effectiveness of the treatment. An algorithm driving an autonomous infusion pump would need to be capable of dealing with the complexity of such scenarios, and relevant training data would need to be available.

"That's going to be difficult, because you're going to have to work out the pharmacological interactions of every drug that you're giving at the same time, also the time frames are all different […] So if you've got some patients with sepsis, as you said, they may well be on steroids, which some make your sugar go awry. They may, on occasion, be on adrenaline, which can make your glucose metabolism change. We may well then start feed at some point, which will then change, or we may well stop feed if they start vomiting. It will also depend on what the other drugs are actually made up in, some are made up of 5% dextrose, and some are made up in normal saline." Staff-03.

A participant from the regulatory space referred to ongoing research around the use of synthetic data, which might be able to address these issues with data quality. Real-world data can contain bias, it can be incomplete or duplicated. An alternative or complementary approach might be to generate synthetic data through simulation. In this way, the AI system might have access to an almost infinite amount of data covering also very infrequent scenarios. A second participant from the regulatory stakeholder group summarised these concerns in a useful list of questions that should be asked about AI data quality.

"Personally, I would like to know about the quality of the training data, what was done to scrub the training data, remove duplicates, fix or exclude partial records etc., its performance data, e.g.*, what is the rate of false positives? The rate of false negatives? What is the consequence of a false positive or negative? I want to know what biases exist in the data; all data has bias."* Regulator-03.

3.4. Regulation

Participants touched upon aspects of regulation that were closely related to the issues discussed regarding safety assurance practices including the continued relevance of existing standards and the challenge of managing AI evolution. Participants felt that the existing standards described good safety engineering reflecting best practice in the sector, which should be applied to AI and autonomous systems. As an estimate, a participant suggested that around 80% of what is being done for medical devices would equally apply to AI, and that maybe 20% would need to be covered by new standards. Regulators would expect developers to demonstrate that they have identified relevant standards, and that they have understood the breadth of standards that might be applicable.

One of the technology developers called for a culture change within the regulatory space in order to accommodate evolving AI systems. Regulatory approaches assume static systems, which do not change once assessed and certified. The interview participant alluded to the situation with safety and security, where potentially there might be pointers as to how to manage such situations.

"How does it [AI evolution] fit with regulators? It will be, regulators like the idea of a static system, which is built and certified, if you change anything, that change needs to be carefully assessed. There needs to be a cultural change among regulators to accept a much more dynamic situation. There are parallels with systems that have security implications. Safety-Security trade-off, static versus dynamic system. Safety doesn't want to change the baseline, but security wants to update frequently." Technology-03.

Participants from the regulatory stakeholder group acknowledged that AI evolution presented a challenge. Participants suggested that the regulatory framework had to move away from pre-market approval and post-market surveillance towards a much more iterative approach. They pointed to the standardisation work that has already been launched and referred to safety assurance practices such as learning in batches with strict version control of algorithms so that any potential issues can be tracked. Most importantly, however, regulators would expect developers to demonstrate competency.

This focus on competence, expertise and processes was echoed by technology developers. Participants from this group suggested that it was important to have a dialogue with regulators around the impact of changes to the system as the underlying algorithms learn and evolve. If it was possible to agree a defined change envelope, where the impact of AI evolution on the system and on the intended use was reasonably minor, then additional regulatory approval might not be required for every iteration. What was important was to agree with the regulator a process for managing such changes, and then to provide evidence that this process was executed accordingly.

3.5. Incident investigation

3.5.1. Traceability and auditability

Several participants felt that the ability of AI systems to log every piece of data and every action would facilitate the incident investigation process, because there would be an auditable and traceable history of what the AI did. The participants suggested that the incident investigation process would, therefore, not be reliant as much on human memory, which tends to be less accurate. In addition, documentation done by people is frequently incomplete, whereas the electronic logs would be fully available.

The ability of AI to produce such an audit trail was not contested by other participants. However, some of the participants raised concerns about how these data would be used during the incident investigation process and what it might mean for clinicians.

One of the hospital staff felt that when it came to the investigation of an incident, it was unclear to whom an AI system would be answerable to. They felt that, in the end, responsibility for safe patient care would rest with the clinician. Another participant from the technology development group expressed a similar feeling, suggesting that almost by default clinicians would bear responsibility for an incident unless they could prove otherwise. This might be a significant burden on clinicians, and it might reduce their willingness to trust and to accept AI systems.

3.5.2. Additional actors in the incident investigation process

Participants suggested that the incident investigation process would need to involve and have input from a wider range of stakeholders. With the introduction of AI systems, it would be necessary to involve AI experts to provide an informed opinion on the quality and safety aspects of AI systems. In addition, manufacturers would have to be brought in much more closely than is currently the case, because the clinical decision-making process is much more blurred with AI systems than it is with current systems. However, this could also create problems. On the one hand, developers of AI systems might be concerned about their reputation and the detrimental economic impact if their systems were found to be responsible for causing patient harm. As a result, there is a danger that even though manufacturers have an interest in improving and fixing their systems, they might drag out and delay the incident investigation process.

One of the technology developers explained further that intellectual property rights might become problematic and prevent manufacturers from sharing their data freely during an incident investigation process. Data can carry a significant economic value.

"If there's a court litigation, then data would need to be made available anyway, e.g.*, test data. It wouldn't be made public from the start, but would be shared if there's a need. If there was any IP that would be shared, that could make it trickier, training sets might be very valuable."* Technology-03.

A nursing participant questioned accountability for incidents. The participant expressed concerns about whether they would be held

responsible for incidents involving an autonomous system.

"Who's going to be accountable for that then because would it… if your patient dies and you end up in Coroners because the noradrenaline wasn't being administered and they had no blood pressure and then they arrest. Is it good enough to say well the pump didn't tell me? It's my pin number, it's still my patient. It's accountability." Staff-04.

4. Discussion

The analysis of the interviews with a diverse set of stakeholders provides a rich description of people's perceptions of and attitudes towards the potential impact of healthcare AI, requirements for human-AI interaction, safety assurance practices and regulatory frameworks for AI and the gaps that exist, and how incidents involving AI should be managed. Reflecting on this description, four important lessons for the safe design and adoption of healthcare AI stand out: (1) a socio-technical systems approach to understanding and designing healthcare AI; (2) the continued relevance of lessons learned from highly automated systems; (3) the critical role of the patient – clinician relationship; and (4) the need for a cultural change around the safety of healthcare AI and of digital health more generally.

Health systems are socio-technical systems where people interact with one another to achieve shared goals, using tools and technologies to complete specific tasks, which are performed in physical spaces, and are situated within organisational contexts as well as wider professional, legal, and societal rules and expectations. The design and implementation of any technology, including healthcare AI, should be based on a thorough understanding of the operational realities of health systems and how the elements of a work system interact with one another (i.e., work-as-done (Hollnagel et al., 2015). An example of a widely used conceptual systems framework developed specifically for healthcare is the Systems Engineering for Patient Safety (SEIPS) model, which can be helpful in describing the interactions of work system elements, and how these interactions deliver healthcare processes, which in turn lead to system outcomes such as patient health and patient safety, and staff wellbeing (Carayon et al., 2020; Holden and Carayon, 2021). Accident models based on systems thinking include Systems-Theoretic Accident Model and Processes (STAMP) along with the corresponding analysis method Systems-Theoretic Process Analysis (STPA) (Leveson, 2011). The Functional Resonance Analysis Method (FRAM) is another systems-based analysis method that can be used to explore interactions in everyday work as a basis for design (Hollnagel, 2012). For example, FRAM has been used to study IV infusion practices in ICU to highlight performance variability, which can inform design requirements for AI technology that supports rather than erodes the adaptive capacity within this system (Furniss et al., 2020). Systems frameworks and systems analysis methods are essential for ensuring that AI is integrated meaningfully and safely into health systems (Sujan et al., 2022).

Important lessons for the design and use of healthcare AI can be derived from the experiences with highly automated systems (Macrae, 2019). For example, Bainbridge's seminal short paper on the "ironies of automation" remains very relevant (Bainbridge, 1983): even though automation is introduced to reduce reliance on people, human operators are still part of the system and are expected to intervene in exceptional and critical situations. However, people then must respond under time pressure and will have less practical experience than previously, and they might not have had opportunities to build a full understanding of the situation they are expected to manage. Humans are not well suited to passive monitoring tasks over extended periods of time. A recent tragic illustration of these ironies of automation in a modern context is the Uber / Volvo self-driving test vehicle accident in March 2018 when a pedestrian pushing a bicycle across the road was killed. The AI in the Volvo test vehicle initially failed to correctly classify the pedestrian, and the safety driver failed to take over manual control as they were using their mobile phone (Stanton et al., 2019). However, the classic ironies of automation also present themselves in slightly different and even more complex form in modern AI technologies. An example of this is situation awareness (SA), which refers to the dynamic understanding of an ongoing situation (Endsley, 1995). From a systems perspective, distributed situation awareness (DSA) (Stanton et al., 2006) regards SA as being distributed around the socio-technical system, and SA is built through interactions between agents – both human (e.g., clinicians) and non-human (e.g., a healthcare AI system). The design and safety assurance of healthcare AI need to consider how both people and AI can build and maintain appropriate SA by designing suitable interactions. In the case of AI-based partially autonomous infusions pumps, for example, the AI might need to know about other medications the patient is receiving as well as other relevant circumstances (e.g., recent food intake, which will affect blood sugar levels). Research around explainable AI might also provide insights into how people can build better awareness of the decisions and actions of the AI, but this is an area of current debate about the best ways to achieve this (Weld and Bansal, 2019; Shneiderman, 2020), or whether it can be achieved meaningfully at all in specific situations (Ghassemi et al., 2021).

The interviews also highlighted the importance of considering the impact on (human) relationships when AI is used. Patients on ICU are particularly vulnerable and distressed, and the human aspect of care is critically important to them. While patients generally had positive attitudes and trust in the technical aspects of AI, they expressed concern about the potential impact on the relationship with their clinicians. A literature review suggested that unfortunately patient attitudes towards AI are not usually included in clinical trials of healthcare AI applications (Scott et al., 2021), and some studies found that fear of dehumanisation of the relationship between patients and their clinicians was a significant concern (Esmaeilzadeh, 2020; Sisk et al., 2020). Such concerns echo findings from research about patient perceptions on their care, e.g., in the management of deterioration. In this domain, concerns have been raised that patients who experience a deterioration in their condition find it challenging to be heard by their clinicians, partly because clinicians give preference to "objective" parameters such as early warning scores (Subbe et al., 2021). The design and use of AI would normally be based precisely on such parameters, and, therefore, consideration needs to be given to patient experience and the impact of the introduction of AI to protect and strengthen the relationship between patients and their clinicians.

There was broad agreement among interview participants from the technology and regulatory sectors that existing safety assurance practices and regulatory frameworks provided an excellent starting point for the development and regulation of AI in clinical settings. The consensus among participants was that no revolution in standards was required, rather refinements were necessary and additional work to address some of the gaps left in existing standards. These gaps relate to the management of AI evolution and the assurance and regulation of data. In the UK, relevant standards include, for example, the clinical risk management standards (referred to as DCB 0129 and DCB 0160, respectively) managed by NHS Digital. Similar views are expressed in the literature, where calls for "digital exceptionalism" have been rejected (Greaves et al., 2018; The Lancet, 2018). However, questions remain as to how familiar technology developers are with these standards and the concepts underpinning them, especially new developers coming from the AI domain who might have little previous experience with the design of health information technology (Habli et al., 2018). There is a need to build capacity and knowledge about safety assurance practices for AI and digital technologies within the health sector (Sujan and Habli, 2021). In addition, the use of AI might have wider ethical, legal and societal implications, such as fairness and impact on different stakeholder groups Wawira Gichoya et al, 2021; Burton et al., 2020. These go beyond the scope of standard failure analysis techniques. Traditional regulatory frameworks and the culture around digital safety need to be suitably extended to be sensitive to these broader issues.

4.1. Limitations

The study set out to explore how different stakeholders think about safety and safety assurance of healthcare AI using a specific example. While the findings were rich, they are limited by the selection of participants and the scenario considered. All participants came from a UK context, and there might be national and cultural differences in perceptions internationally. The scenario was one of care delivery, but the potential for application of healthcare AI is much broader and different factors might be of greater relevance in other areas of healthcare, such as flow optimisation, logistics and drug development. Data were collected during 2019 and before the coronavirus pandemic, and attitudes and perceptions might have changed since then.

5. Conclusion

The paper provides a rich and contextualised description of how different stakeholders think about safety and assurance of healthcare AI. This description of a diversity of views can support responsible innovation and adoption of such technologies in healthcare by encouraging anticipation through consideration of a broader range of concerns. Safety and assurance of healthcare AI need to be based on a systems approach that expands the current technology-centric focus. Lessons can be learned from the experiences with highly automated systems across safety-critical industries, but issues such as the impact of AI on the relationship between patients and their clinicians requires greater consideration. Existing standards and best practices for the design and assurance of systems should be followed, but there is a need for greater awareness of these among technology developers. In addition, wider ethical, legal, and societal implications of the use of AI in healthcare need to be addressed.

CRediT authorship contribution statement

Mark A. Sujan: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Sean White:** Writing – review & editing, Investigation, Funding acquisition, Conceptualization. **Ibrahim Habli:** Writing – review & editing, Funding acquisition, Conceptualization. **Nick Reynolds:** Writing – review & editing, Resources, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

References

Alberdi, E., Povyakalo, A., Strigini, L., Ayton, P., 2004. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Acad. Radiol. 11 (8), 909–918.

Ash, J.S., Sittig, D.F., Dykstra, R.H., Guappone, K., Carpenter, J.D., Seshadri, V., 2007. Categorizing the unintended sociotechnical consequences of computerized provider order entry. Int. J. Med. Inf. 76 (Suppl 1), S21–S27.

Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., Shah, N.H., 2018. Improving palliative care with deep learning. BMC Med. Inf. Decis. Making 18 (S4). https://doi.org/10.1186/s12911-018-0677-8.

Bainbridge, L., 1983. Ironies of automation. Automatica. 19 (6), 775–779.

E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2020. p. 1–12.

Black, A.D., Car, J., Pagliari, C., Anandan, C., Cresswell, K., Bokun, T., McKinstry, B., Procter, R., Majeed, A., Sheikh, A., Djulbegovic, B., 2011. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS Med. 8 (1), e1000387. https://doi.org/10.1371/journal.pmed.1000387.

Blomberg, S.N., Folke, F., Ersbøll, A.K., Christensen, H.C., Torp-Pedersen, C., Sayre, M.R., Counts, C.R., Lippert, F.K., 2019. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation. 138, 322–329.

Blomberg, S.N., Christensen, H.C., Lippert, F., Ersbøll, A.K., Torp-Petersen, C., Sayre, M. R., Kudenchuk, P.J., Folke, F., 2021. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. JAMA Network Open. 4 (1), e2032320. https://doi.org/10.1001/jamanetworkopen.2020.32320.

Bødker, S., 2000. Scenarios in user-centred design—setting the stage for reflection and action. Interact. Comput. 13 (1), 61–75.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol.. 3 (2), 77–101.

Braun, V., Clarke, V., 2021. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. Qual. Res. Sport, Exercise Health. 13 (2), 201–216.

Braun, V., Clarke, V., 2022. Thematic Analysis: A Practical Guide. Sage Publications Ltd, London.

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., Porter, Z., 2020. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artif. Intell. 279, 103201. https://doi.org/10.1016/j.artint.2019.103201.

Cabitza, F., Rasoini, R., Gensini, G.F., 2017. Unintended Consequences of Machine Learning in Medicine. JAMA, J. Am. Med. Assoc. 318 (6), 517. https://doi.org/10.1001/jama.2017.7797.

Calnan, M.W., Sanford, E., 2004. Public trust in health care: the system or the doctor? Qual. Safety Health Care. 13, 92.

Carayon, P., Wooldridge, A., Hoonakker, P., Hundt, A.S., Kelly, M.M., 2020. SEIPS 3.0: Human-centered design of the patient journey for patient safety. Appl. Ergon. 84, 103033. https://doi.org/10.1016/j.apergo.2019.103033.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., Tsaneva-Atanasova, K., 2019. Artificial intelligence, bias and clinical safety. BMJ quality & safety. 28 (3), 231–237.

Elish, M.C., Watkins, E.A., 2020. Reparing innovation: a study of integrating AI in clinical care. Data & Society Research Institute, New York.

Elish, M.C., 2018. The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care. Ethnographic Praxis in Industry Conference Proceedings, 2018, pp. 364-80.

Elliott, R.A., Camacho, E., Campbell, F., Jankovic, D., St James, M.M., Kaltenthaler, E., et al., 2018. Prevalence and economic burden of medication errors in the NHS in England. Policy Research Unit in Economic Evaluation of Health & Care Interventions, Sheffield.

Endsley, M.R., 1995. Toward a Theory of Situation Awareness in Dynamic Systems. Hum. Factors 37 (1), 32–64.

Esmaeilzadeh, P., 2020. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med. Inf. Decis. Making 20, 170.

Fitzpatrick, K.K., Darcy, A., Vierhile, M., 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health. 4 (2), e19. https://doi.org/10.2196/mental.7785.

Furniss, D., Nelson, D., Habli, I., White, S., Elliott, M., Reynolds, N., Sujan, M., 2020. Using FRAM to explore sources of performance variability in intravenous infusion administration in ICU: A non-normative approach to systems contradictions. Appl. Ergon. 86, 103113. https://doi.org/10.1016/j.apergo.2020.103113.

Ghassemi, M., Oakden-Rayner, L., Beam, A.L., 2021. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digital Health. 3 (11), e745–e750.

Greaves, F., Joshi, I., Campbell, M., Roberts, S., Patel, N., Powell, J., 2018. What is an appropriate level of evidence for a digital health intervention? The Lancet. 392 (10165), 2665–2667.

Habli, I., White, S., Sujan, M., Harrison, S., Ugarte, M., 2018. What is the safety case for health IT? A study of assurance practices in England. Saf. Sci. 110, 324–335.

Hill, R.G., Sears, L.M., Melanson, S.W., 2013. 4000 Clicks: a productivity analysis of electronic medical records in a community hospital ED. The American Journal of Emergency Medicine 31 (11), 1591–1594.

Holden, R.J., Carayon, P., 2021. SEIPS 101 and seven simple SEIPS tools. BMJ Qual. Safety 30 (11), 901–910.

Hollnagel, E., 2015. Why is Work-as-Imagined different from Work-as-Done? In: Wears, R., Hollnagel, E., Braithwaite, J. (Eds.), The Resilience of Everyday Clinical Work. Ashgate, Farnham.

Hollnagel, E., 2012. FRAM, the functional resonance analysis method: modelling complex socio-technical systems: Ashgate Publishing, Ltd.; 2012.

Joshi, I., Morley, J., 2019. Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care. NHSX, London.

Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C., Faisal, A.A., 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat. Med. 24 (11), 1716–1720.

Koppel, R., Metlay, J.P., Cohen, A., Abaluck, B., Localio, A.R., Kimmel, S.E., et al., 2005. Role of computerized physician order entry systems in facilitating medication errors. JAMA, J. Am. Med. Assoc. 293, 1197–1203.

Lai, V., Tan, C., 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA: Association for Computing Machinery; 2019. p. 29–38.
Lai, V., Liu, H., Tan, C., 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery, pp. 1–13.
Leveson, N.G., 2011. Applying systems thinking to analyze and learn from events. Saf. Sci. 49 (1), 55–64.
Lyons, I., Furniss, D., Blandford, A., Chumbley, G., Iacovides, I., Wei, L.i., Cox, A., Mayer, A., Vos, J., Galal-Edeen, G.H., Schnock, K.O., Dykes, P.C., Bates, D.W., Franklin, B.D., 2018. Errors and discrepancies in the administration of intravenous infusions: a mixed methods multihospital observational study. BMJ Quality Safety. 27 (11), 892–901.
Macrae, C., 2019. Governing the safety of artificial intelligence in healthcare. BMJ Quality Safety. 28 (6), 495–498.
McDowell, S.E., Mt-Isa, S., Ashby, D., Ferner, R.E., 2010. Where errors occur in the preparation and administration of intravenous medicines: a systematic review and Bayesian analysis. Qual. Safety Health Care. 19 (4), 341–345.
McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F.J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C.J., King, D., Ledsam, J.R., Melnick, D., Mostofi, H., Peng, L., Reicher, J.J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K.C., De Fauw, J., Shetty, S., 2020. International evaluation of an AI system for breast cancer screening. Nature 577 (7788), 89–94.
McLeod, M.C., Barber, N., Franklin, B.D., 2013. Methodological variations and their effects on reported medication administration error rates. BMJ quality & safety. 22 (4), 278–289.
Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38.
Muehlematter, U.J., Daniore, P., Vokinger, K.N., 2021. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digital Health 3 (3), e195–e203.
Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H., et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ (Clinical research ed). 2020;368: m689.
Parasuraman, R., Riley, V., 1997. Humans and Automation: Use, Misuse, Disuse. Abuse. Human Factors. 39 (2), 230–253.
Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2000. A model for types and levels of human interaction with automation. IEEE Trans. Syste., Man Cybern. - Part A: Syst. Hum. 30 (3), 286–297.
Peek, N., Sujan, M., Scott, P., 2020. Digital health and care in pandemic times: impact of COVID-19. BMJ Health Care Inform.. 27 (1), e100166. https://doi.org/10.1136/bmjhci-2020-100166.
Ruskin, K.J., Hueske-Kraus, D., 2015. Alarm fatigue: impacts on patient safety. Current Opinion. Anesthesiology 28 (6), 685–690.
Russell, S., 2019. Human compatible: Artificial intelligence and the problem of control. Penguin, London.
Saldaña, J., 2009. The coding manual for qualitative researchers. Sage, London.
Scott, I.A., Carter, S.M., Coiera, E., 2021. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform. 28 (1), e100450. https://doi.org/10.1136/bmjhci-2021-100450.
Selin, C., 2011. Negotiating plausibility: intervening in the future of nanotechnology. Sci. Eng. Ethics 17 (4), 723–737.
Shneiderman, B., 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. Int. J. Hum.-Comput. Inter. 36, 495–504.
Sikstrom, L., Maslej, M.M., Hui, K., Findlay, Z., Buchman, D.Z., Hill, S.L., 2022. Conceptualising fairness: three pillars for medical algorithms and health equity. BMJ Health Care Inform. 29 (1), e100459. https://doi.org/10.1136/bmjhci-2021-100459.
Sisk, B.A., Antes, A.L., Burrous, S., DuBois, J.M., 2020. Parental Attitudes toward Artificial Intelligence-Driven Precision Medicine Technologies in Pediatric Healthcare. Children. 7 (9), 145. https://doi.org/10.3390/children7090145.
Smallman, M., 2019. Policies designed for drugs won't work for AI. Nature 567 (7746), 7.
Stanton, N.A., Stewart, R., Harris, D., Houghton, R.J., Baber, C., McMaster, R., Salmon, P., Hoyle, G., Walker, G., Young, M.S., Linsell, M., Dymott, R., Green, D., 2006. Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology. Ergonomics 49 (12-13), 1288–1311.
Stanton, N.A., Salmon, P.M., Walker, G.H., Stanton, M., 2019. Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian. Saf. Sci. 120, 117–128.
Stilgoe, J., Owen, R., Macnaghten, P., 2013. Developing a framework for responsible innovation. Res. Policy 42 (9), 1568–1580.
Subbe, C.P., Ahsan, S., Smith, L., Renggli, J.F., 2021. An audible patient voice: How can we ensure that patients are treated as partners in their own safety? Future Healthcare J. 8 (3), e564–e566.
Sujan, M., Furniss, D., Embrey, D., Elliott, M., Nelson, D., White, S., et al., 2019. Critical barriers to safety assurance and regulation of autonomous medical systems. In: Beer, M., Zio, E. (Eds.), 29th European Safety and Reliability Conference (ESREL 2019). CRC Press, Hannover.
Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., Reynolds, N., 2019. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inform. 26 (1), e100081. https://doi.org/10.1136/bmjhci-2019-100081.
Sujan, M., Habli, I., 2021. Safety cases for digital health innovations: can they work? BMJ Qual. Safety. 30 (12), 1047–1050.
Sujan, M., Pool, R., Salmon, P., 2022. Eight Human Factors and Ergonomics Principles for Healthcare AI. BMJ Health Care Inform. 29, e100516. https://doi.org/10.1136/bmjhci-2021-100516.
The Lancet, 2018. Is digital medicine different? The Lancet. 392 (10142), 95. https://doi.org/10.1016/S0140-6736(18)31562-9.
van Notten, P.W.F., Rotmans, J., van Asselt, M.B.A., Rothman, D.S., 2003. An updated scenario typology. Futures. 35 (5), 423–443.
Wawira Gichoya, J., McCoy, L.G., Celi, L.A., Ghassemi, M., 2021. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform. 28 (1), e100289. https://doi.org/10.1136/bmjhci-2020-100289.
Weld, D.S., Bansal, G., 2019. The challenge of crafting intelligible intelligence. Commun ACM. 62 (6), 70–79.
Winter, P., Carusi, A., 2022. 'If You're Going to Trust the Machine, Then That Trust Has Got to Be Based on Something': : Validation and the Co-Constitution of Trust in Developing Artificial Intelligence (AI) for the Early Diagnosis of Pulmonary Hypertension (PH). Sci. Technol. Stud., 2022.
Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J., 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat. Med. 27 (4), 582–584.