BMJ Health & Care Informatics

Human factors challenges for the safe use of artificial intelligence in patient care

Mark Sujan ,[1,2] Dominic Furniss,[2] Kath Grundy,[3] Howard Grundy,[3] David Nelson,[4] Matthew Elliott,[4] Sean White,[5] Ibrahim Habli,[6] Nick Reynolds[4]

To cite: Sujan M, Furniss D, Grundy K, *et al*. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform* 2019;**26**:e100081. doi:10.1136/bmjhci-2019-100081

Received 31 May 2019
Accepted 14 November 2019

[1]Warwick Medical School, University of Warwick, Coventry, UK
[2]Human Reliability Associates, Dalton, UK
[3]Patient, Derby, UK
[4]Intensive Care Unit, University Hospitals of Derby and Burton NHS Foundation Trust, Derby, UK
[5]Clinical Safety Team, NHS Digital, Leeds, UK
[6]Department of Computer Science, University of York, York, UK

Correspondence to
Dr Mark Sujan;
mark.sujan@humanreliability.com

ABSTRACT

The use of artificial intelligence (AI) in patient care can offer significant benefits. However, there is a lack of independent evaluation considering AI in use. The paper argues that consideration should be given to how AI will be incorporated into clinical processes and services. Human factors challenges that are likely to arise at this level include cognitive aspects (automation bias and human performance), handover and communication between clinicians and AI systems, situation awareness and the impact on the interaction with patients. Human factors research should accompany the development of AI from the outset.

INTRODUCTION

The use of artificial intelligence (AI) in patient care currently is one of the most exciting and controversial topics. It is set to become one of the fastest growing industries, and politicians are putting their weight behind this, as much to improve patient care as to exploit new economic opportunities. In 2018, the then UK Prime Minister pledged that the UK would become one of the global leaders in the development of AI in healthcare and its widespread use in the National Health Service. The Secretary for Health and Social Care, Matt Hancock, is a self-professed patient registered with Babylon Health's GP at Hand system, which offers an AI-driven symptom checker coupled with online general practice (GP) consultations replacing visits at regular GP clinics.

GP at Hand is, arguably, one of the best-known AI-supported services currently in use in the UK. It is not without controversy, though, and a recent report has found evidence that, on average, patients attracted to GP at Hand tend to be younger and healthier than those at regular GP clinics.[1] This might have significant funding implications, which as yet have not been properly evaluated and understood.

Encouraging results have been achieved across a wide range of AI services, in particular in domains that rely on pattern recognition, classification and prediction. Examples include the use of deep neural networks (DNNs) to determine whether skin lesions are malignant or benign. In an evaluation study, the DNN outperformed doctors and achieved accuracy of around 70%.[2] Diabetes is a major public health concern, affecting around four million people in the UK, and researchers have developed an app based on DNNs that can detect changes in vascular activity using the light and camera on people's smartphones to determine whether a person is likely to suffer from diabetes.[3] Predictive use of DNNs has been demonstrated in a study that developed an algorithm to support palliative care by predicting mortality in the hospital.[4] Mental health is another area that might benefit significantly from the introduction of AI because access to mental health professionals remains challenging, and the perceived barriers to seeking help are frequently high. AI-based apps have been developed to deliver cognitive–behavioural therapy. A small, prospective trial of an AI chatbot found that for a limited sample size, the outcomes achieved by this app were superior to other forms of therapeutic contact.[5]

While all of these developments provide avenues for potentially significant patient benefit, it is also timely to take a step back and to consider whether it is safe to use AI in patient care or, more specifically, what kind of evidence is available, and what kinds of challenges might have to be addressed. While technical challenges, such as the quality of training data and the potential introduction of bias, have been recognised and discussed,[6] less emphasis has been given so far to the impact of integrating AI into clinical processes and services. It is at this level, where humans and AI come together, that human factors challenges are likely to emerge.

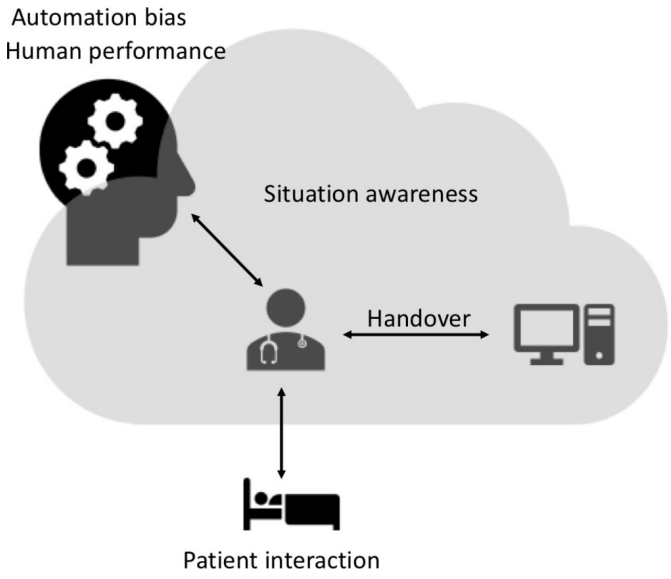

Figure 1 Overview of human factors challenges of using artificial intelligence in patient care.

TECHNOLOGY FOCUS OF EVALUATION STUDIES

Unsurprisingly, many of the published studies on AI in patient care focus on the technology itself because developers are keen to demonstrate that the technology is working. It is common to find claims that pit the performance of, for example, a DNN against that of clinicians at undertaking a well-defined and narrow task. Examples include identification of skin cancer,[7] identification of high-risk breast lesions not requiring surgical excision[8] and detection of diabetic retinopathy.[9] These studies have shown that AI systems often outperform humans at such tasks. However, the evidence base to date remains weak; sample sizes are often small; and prospective trials are infrequent.[10]

Compared with the large number of evaluation studies undertaken by the developers of AI algorithms, independent evaluation studies are relatively infrequent. Where there is an independent evaluation, the headline figures are not always reproduced. For example, an audit of 23 patient-facing symptom checkers found that the correct diagnosis was listed as the most probable one in only around one-third of the test cases.[11] There is a step change from validating the technology per se to evaluating its use in patient care.

Crucially, when delivering patient care, the integration into clinical systems needs to be considered, but prospective trials of AI remain the exception so far. Situations where the AI system delivers a service by itself will be far fewer than scenarios where clinical teams of healthcare professionals and AI systems will be cooperating and collaborating to provide patient care. It is likely that some of these AI systems will be autonomous agents that operate as part of the clinical team. An example is the future use of autonomous infusion pumps in intensive care, where the infusion pumps can adjust or stop infusions independently.[12] Clinicians remain in overall charge, but they need to manage and cooperate with these autonomous agents. This is not dissimilar to pilots supervising flight management systems in modern aircraft, with all the human factors issues and perils that recent major accidents, such as the Boeing 737 Max, have brought to attention.

INTEGRATION OF AI INTO CLINICAL SYSTEMS

When automation started to be deployed at scale in industrial systems, human factors research on 'automation surprises' and the 'ironies of automation' explained some of the problems that appeared with the introduction of automation.[13 14] The fundamental fallacy is the assumption that automation might replace people, but in actual reality, the use of automation changes and transforms what people do.[15] Clinical systems are not necessarily comparable to commercial aircraft or autonomous vehicles. However, a look across these different industries can be useful to highlight potential human factors challenges that are likely to require consideration when adopting AI in patient care. The human factors challenges discussed further relate to cognitive aspects (automation bias and human performance), handover and communication between clinicians and AI systems, situation awareness and the impact on the interaction with patients (see figure 1).

Automation bias

Studies in aviation dating back to the 1980s and 1990s and analysis of incident reports recorded in the Aviation Safety Reporting System found that pilots frequently failed to monitor important flight indicators or did not disengage the autopilot and automated flight management systems in the cockpit in case of malfunction.[16 17] For example, in 1985, the US National Transportation Safety Board (NTSB) investigated an incident involving a China Airlines Boeing 747 SP-9 flying from Taipei to Los Angeles. When the aircraft was close to San Francisco, an engine failed. The autopilot took mitigating actions but did not alert the pilots to this problem. The pilots only became aware of the engine failure when they disengaged the autopilot and the aircraft started rolling over and dived into an uncontrolled descent. The NTSB report concluded that 'the probable cause of this accident was the captain's preoccupation with an inflight malfunction and his failure to monitor properly the airplane's flight instruments […] Contributing to the accident was the captain's over-reliance on the autopilot […]' (NTSB report AAR-86–03).

This phenomenon is referred to as automation bias or automation-induced complacency, and represents an example of inappropriate decision-making as a result of over-reliance on automation.[18] Automation bias can lead to omission errors, where people do not take a required action because the automation failed to alert them, and it can lead to errors of commission, where people follow the inappropriate advice of an automated system.[16]

The speed with which people start to rely and over-rely on automation and AI-driven autonomous systems might come as a surprise to many as a recent study in the automotive domain has demonstrated.[19] A sample of 49 experienced drivers were instructed on the limitations of a partially autonomous car and were asked to complete a 30 min commute for 1 week. By the end of the week, most of the drivers were not watching the road anymore and spent instead about 80% of their time on their smartphones or reading books and documents.

Healthcare is transitioning towards digital and AI-supported clinical environments at a rapid pace, and we can and should expect clinicians to come to trust and rely on the technology. This brings with it the risk of automation bias, and this can potentially affect clinician decision-making for millions of patients. Automation bias introduced with clinical decision support systems has been highlighted in a number of studies. An early study comparing the performance of radiologists interpreting mammograms found that under certain situations, the performance of expert radiologists deteriorated when supported by a decision support system that highlighted specific areas to focus on.[20] A study investigating the impact of decision support on the accuracy of ECG interpretation found that while correct decision support classification increased clinician (non-cardiologist) accuracy, incorrect decision support classification decreased the accuracy of clinicians from 56% to 48%.[21] Similar findings of the effects of clinical decision support were produced by another study looking at the impact of clinical decision support in electronic prescribing systems.[22] The study found that clinical decision support reduced prescribing errors when working correctly but also increased prescribing errors by around one-third in cases where the system either did not alert the clinician to a potential problem or provided the wrong advice. A review of the literature on automation bias in healthcare identified six studies investigating the impact of automation bias on errors.[23] The study concluded that task complexity (eg, diagnosis supported by a clinical decision support system) and task load (ie, the number of task demands) increased the likelihood of over-reliance on automation.

Many, if not most, AI systems will be advertised as having ultrahigh reliability, and it is to be expected that in due course, clinicians will come to rely on these systems. However, studies on automation bias suggest that the reliability figures by themselves do not allow prediction of what will happen in clinical use, when the clinician is confronted with a potentially inaccurate system output.[20] How easy or difficult will it be to spot this, and how will the potential for automation bias be guarded against?

Impact on human performance

Expertise is built through frequent exposure and training. The current generation of human car drivers is reasonably skilled in managing complex traffic situations because many of us do it on an everyday basis. Will the generation that has grown up with autonomous vehicles have the same levels of basic driving skills that enable them to retake control in potentially highly time-critical and complex traffic situations when the AI system fails? This is particularly relevant in healthcare, where healthcare professionals take pride in their professional skill sets. Will the expertise of radiographers deteriorate when they are exposed only to specific images specifically selected by an AI system rather than the broad range of images they currently train on day by day?[24]

Ironically, AI algorithms are frequently trained and validated against baseline data developed from human performance (eg, radiologist reading of images), and the erosion of training opportunities and hands-on skills for clinicians as a result of introducing AI systems might create a vicious circle where the quality of baseline data deteriorates in the long term.

Handover

A key argument for the safety of autonomous vehicles is that the driver is able to take control in case of emergencies or unforeseen situations. However, the well-publicised fatal Tesla accidents of Josh Brown in 2016 and more recently of Jeremy Banner in March 2019 tragically demonstrate that drivers do not always take control from the autopilot when required. Research has put into question whether such an assumption is realistic in the first place, considering the short reaction time available.[25]

Handover is a well-recognised safety critical task in the delivery of care, although in traditional conception, we think of handover between clinicians or teams of clinicians.[26] In the future, handover between humans and autonomous AI systems will become increasingly important, and one might assume that this will be even more complex than the handover between the autopilot and the driver of an autonomous vehicle.

The AI system needs to recognise the need to hand over. While this might be achievable, the AI also needs to figure out what to hand over, how this should be done and when. In human handover, we have recognised the need for structured communication protocols to convey clearly the salient features of a situation, for example, age, time, mechanism, injuries, signs, treatments (ATMIST) for emergency care or situation, background, assessment, recommendation (SBAR) more generally. Should there be an equivalent for human—AI handover?

For example, if an autonomous infusion pump delivering insulin starts to recognise that it is struggling to maintain blood sugar levels, at what point should it trigger an alarm to initiate handover? Identifying the precise moment requires trading off accuracy with timeliness. Should the handover simply convey the infusion pump's inability to maintain blood sugar levels, or should the infusion pump provide further information about prior adjustments it made? Is the best strategy to wait for the infusion pump to trigger an alarm and initiate handover, or should we ensure that the clinician is enabled to recognise that a need to retake control will arise?

These questions are fundamentally about how the AI will support clinicians and clinical teams, and how their interaction can be optimised.

Situation awareness

Individuals and teams perform more successfully when they have good situation awareness.[27] Traditional handover contributes to the development of shared situation awareness, and it enables discussion and dialogue.[28] While it might be possible to create autonomous agents that have high reliability, questions arise about what the autonomous system should communicate to clinicians during normal operation to enable the clinician to maintain situation awareness. This is not straightforward to answer by looking simply at one AI system in isolation, because clinicians might be interacting with many autonomous agents (eg, multiple infusion pumps) concurrently, and the design of communication has to consider human information needs and limitations.

Autonomous agents need to build situation awareness, too. An autonomous infusion pump needs to know if the patient receives other medications that might affect the patient's physiology and response. These medications might come via other infusion pumps or they might be given by the clinician. The saying 'if it's not documented, it didn't happen' applies here with critical consequence: if there are relevant activities going on that are not documented and communicated to the autonomous agent (eg, infusion pump), then as far as the AI is concerned, these literally did not happen because the system has no way of knowing about it. The results could be catastrophic.

Patient interaction

AI can improve efficiency of clinical processes and free up clinician time to undertake other tasks. This is potentially very useful in a pressured health system. However, another way of looking at this is that there might be smaller numbers of clinicians that have other tasks to do, potentially away from the patient. Might AI-enabled intensive care units make do with fewer nurses and therefore increase the number of patients per nurse? This might be a worry for patients, because they might see less of their clinicians, and they might find it harder to provide feedback about their care and their condition. For example, if a needle comes unstuck, the patient might be aware of this before the AI system—and could potentially avoid and mitigate any adverse effect—but who do patients communicate this to?

Providing healthcare means being responsive to a patient's physiological as well as personal and emotional needs. In some clinical settings, such as the intensive care unit, the bond between nurse and patient is very strong, and for many patients, their episode in intensive care is traumatic. How will the introduction of AI and autonomous systems in these environments affect this unique relationship? It has been argued that AI might actually create more opportunities for empathy and caring because it might allow clinicians to focus more on these aspects of care.[29] However, whether this is the case, or whether the caring aspect is eroded by transforming, for example, nursing care into AI specialist nurses who 'care' for autonomous systems (ie, supervise them), remains to be seen.

The introduction of AI at scale has the potential to fundamentally change and disrupt communication between patients and their clinicians. Will hospitals become similar to automated supermarket checkouts, with frustrated customers waiting for an overstretched employee to attend to the frequent hassles at the checkout? To date, these issues have received too little attention compared with the focus on accuracy and performance of the AI in isolation.

CONCLUSION

The use of AI in patient care is a disrupter of unprecedented scale, affecting all areas of the health system. Understandably, much effort is devoted to the development of the new technologies. However, it is crucially important that research around human factors, the integration of AI into clinical processes and services, and rigorous evaluation studies are not left behind; they should accompany and inform the development of these exciting innovations from the outset.

Twitter Mark Sujan @MarkSujan and Dominic Furniss @domfurniss

Contributors MS, DF, KG and HG developed the idea for this paper. KG and HG drafted the first version of the manuscript. MS drafted subsequent revisions to the manuscript. All authors reviewed and critiqued the draft manuscripts and contributed to subsequent versions. All authors approved the final version of the manuscript.

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

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Mark Sujan http://orcid.org/0000-0001-6895-946X

REFERENCES

1 Ipsos MORI. *Evaluation of Babylon GP at hand*. London: Ipsos MORI, 2019.
2 Haenssle HA, Fink C, Schneiderbauer R, *et al*. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* 2018;29:1836–42.
3 Avram R, Tison G, Kuhar P, *et al*. Predicting diabetes from PHOTOPLETHYSMOGRAPHY using deep learning. *J Am Coll Cardiol* 2019;73:16.
4 Avati A, Jung K, Harman S, *et al*. Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018;18:122.
5 Fitzpatrick KK, Darcy A, Vierhile M. 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* 2017;4:e19.
6 Challen R, Denny J, Pitt M, *et al*. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28:231–7.
7 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
8 Bahl M, Barzilay R, Yedidia AB, *et al*. High-Risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. *Radiology* 2018;286:810–8.
9 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs. *JAMA* 2016;316:2402–10.
10 Yu K-H, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf* 2019;28:238–41.
11 Semigran HL, Linder JA, Gidengil C, *et al*. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 2015;351.
12 Sujan M, Furniss D, Embrey D, *et al*. 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)*. Hannover: CRC Press, 2019.
13 Bainbridge L. Ironies of automation. *Automatica* 1983;19:775–9.
14 Sarter NB, Woods DD, Billings CE. Automation surprises. In: Salvendy G, ed. *Handbook of Human Factors & Ergonomics*. Wiley, 1997: 1926–43.
15 Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. *IEEE Trans. Syst Man Cybern A* 2000;30:286–97.
16 Mosier KL, Skitka LJ, Heers S, *et al*. Automation bias: decision making and performance in high-tech cockpits. *Int J Aviat Psychol* 1998;8:47–63.
17 Riley V. Operator reliance on automation: Theory and data. In: Parasuraman R, Mouloua M, eds. *Automation and human performance: theory and applications*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1996: 19–35.
18 Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. *Hum Factors* 1997;39:230–53.
19 Burnett G, Large DR, Salanitri D. *How will drivers interact with vehicles of the future?* London: RAC Foundation, 2019.
20 Alberdi E, Povyakalo A, Strigini L, *et al*. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Acad Radiol* 2004;11:909–18.
21 Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J Am Med Inform Assoc* 2003;10:478–83.
22 Lyell D, Magrabi F, Raban MZ, *et al*. Automation bias in electronic prescribing. *BMC Med Inform Decis Mak* 2017;17:28.
23 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association* 2016;45:ocw105–31.
24 Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in MedicineUnintended consequences of machine learning in MedicineUnintended consequences of machine learning in medicine. *JAMA* 2017;318:517–8.
25 Hancock PA. Some pitfalls in the promises of automated and autonomous vehicles. *Ergonomics* 2019;62:479–95.
26 Sujan MA, Chessum P, Rudd M, *et al*. Emergency care handover (echo study) across care boundaries: the need for joint decision making and consideration of psychosocial history. *Emerg Med J* 2015;32:112–8.
27 Endsley MR. Toward a theory of situation awareness in dynamic systems. *Hum Factors* 1995;37:32–64.
28 Sujan M, Spurgeon P, Inada-Kim M, *et al*. Clinical handover within the emergency care pathway and the potential risks of clinical handover failure (echo): primary research. *Health Serv Deliv Res* 2014;2:1–144.
29 Topol E. *Deep medicine: how artificial intelligence can make healthcare human again*. New York: Hachette, 2019.