ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

DEMONSTRATOR PROJECT

FINAL REPORT

SAM
SAFETY ASSURANCE OF AUTONOMOUS INTRAVENOUS MEDICATION MANAGEMENT SYSTEMS – REQUIREMENTS AND STRATEGIES

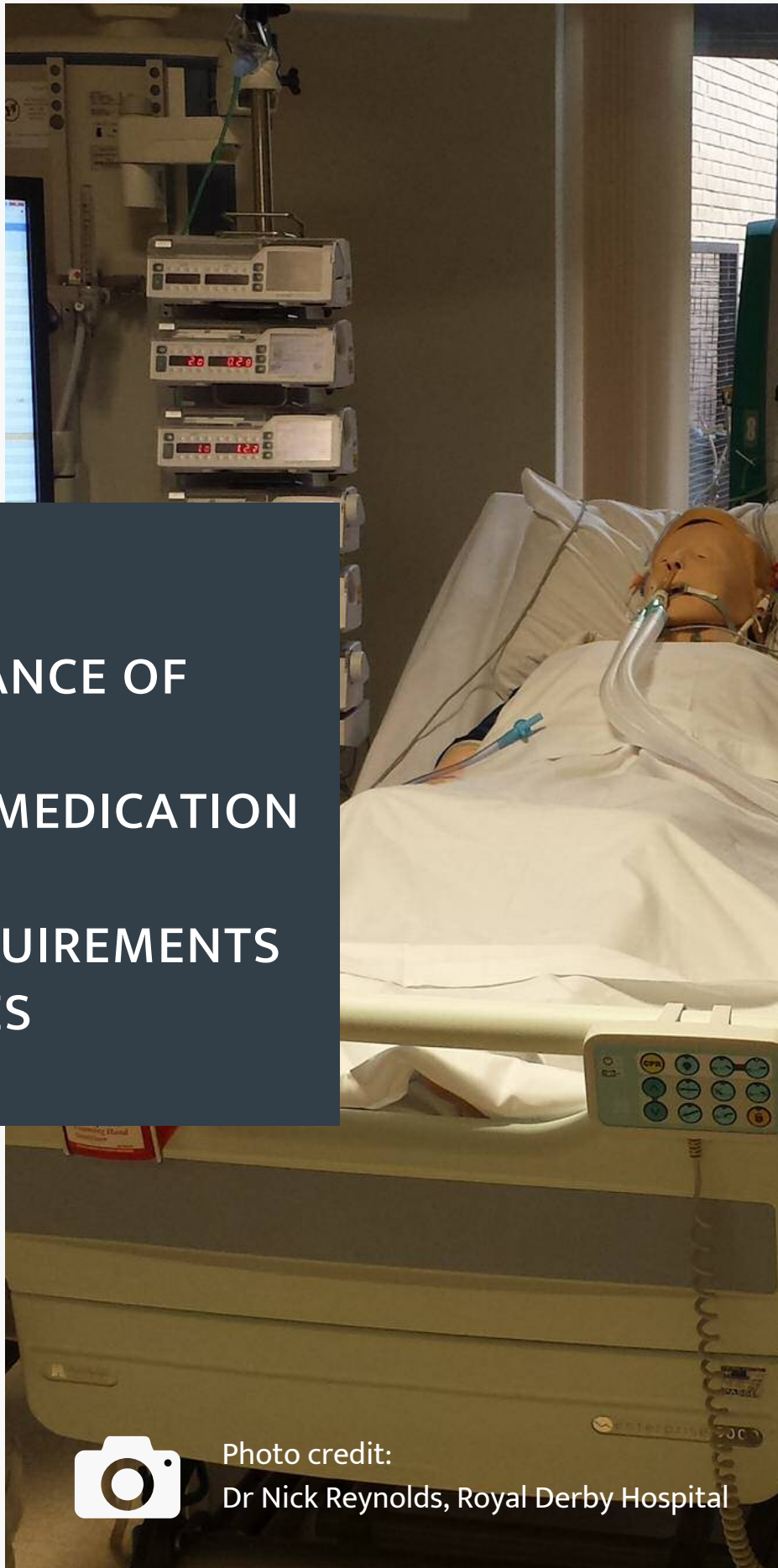

Photo credit:
Dr Nick Reynolds, Royal Derby Hospital

FEBRUARY 2022

SAFETY ASSURANCE OF AUTONOMOUS INTRAVENOUS MEDICATION MANAGEMENT SYSTEMS – REQUIREMENTS AND STRATEGIES (SAM)

Author affiliations

M Sujan[1]*, D Furniss[2], S White[3], S Laher[3], D Nelson[4], M Elliott[4], and N Reynolds[4]

*Corresponding author

[1]Human Factors Everywhere Ltd, Woking, UK

[2]Human Reliability Associates Ltd, Dalton, UK

[3]NHS Digital, Leeds, UK

[4]University Hospitals of Derby and Burton NHS Foundation Trust, Derby, UK

Contact details for corresponding author: mark.sujan@humanfactorseverywhere.com

Declaration of competing interests

All authors declare (1) no financial support for the submitted work from anyone other than their employer and Assuring Autonomy International Programme; (2) no financial relationships with commercial entities that might have an interest in the submitted work; (3) no spouses, partners, or children with relationships with commercial entities that might have an interest in the submitted work; and (4) no non-financial interests that may be relevant to the submitted work.

SCIENTIFIC ABSTRACT

BACKGROUND AND AIMS

This report explores safety assurance challenges of robotic and autonomous systems (RAS) in healthcare using the example of intravenous (IV) medication management systems within an intensive care unit (ICU) setting. The report also investigates safety assurance strategies to address these challenges. Findings are presented from a multi-disciplinary qualitative study that investigated intravenous infusion practices in one ICU in an English National Health Service (NHS) hospital.

The focus of the study was the clinical system rather than the technology as such. The study, therefore, explored safety assurance challenges at the intersection of engineering and human factors.

The project addressed the following study questions:

- Q1: What are perceptions of different stakeholder groups of safety assurance of autonomous IV medication management systems in ICU?
- Q2: What are strengths and weaknesses of representative assurance methods for assuring the safety of autonomous IV medication management systems?

METHODS

The study design utilised a multi-disciplinary qualitative research approach organised into two research strands consisting of four research activities.

	Research Strand 1: Stakeholder perceptions of AI in healthcare
	The aim of this research strand was to describe stakeholder perceptions on safety assurance of AI and autonomous systems for IV medication management in ICU settings. The research strand consisted of one research activity.
Thematic Analysis	26 semi-structured interviews conducted with a purposive sample of stakeholders consisting of patients, hospital staff, technology developers and regulators.
	Research Strand 2: Safety assurance methods and strategies
	The aim of this research strand was to analyse relevant clinical scenarios using three different methods, and to identify strengths and weaknesses of these for addressing safety assurance challenges of RAS in healthcare.

Functional Resonance Analysis	Application of the Functional Resonance Analysis Method (FRAM) to the clinical scenarios.
Human Reliability Analysis	Application of the Systematic Human Error Reduction and Prediction Approach (SHERPA).
Hazard Analysis based on NHS clinical safety standards	Application of hazard analysis using bow-ties based on the logic of NHS clinical safety standards, and supported by the NHS Digital Safety Modelling, Assurance and Reporting Toolkit (SMART).

RESULTS

Stakeholder perceptions on safety assurance of RAS in clinical settings	Interviews with 26 patients, healthcare professionals, technology developers and regulators were carried out. Their views were grouped into 5 categories.
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 autonomous systems should not disrupt the relationship between patients and clinicians.
2. Human – RAS 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.
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. A risk-based approach to AI evolution should be taken. AI can present a black-box

	challenge, and this could be addressed through approaches towards more explainable AI. The use of synthetic data could complement real-world data to provide more comprehensive training data sets.
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. The type and rigour of the evidence expected depends on the intended use of the system and on the types of claims developers are making about their system.
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. The different regulatory bodies for medical devices, professional practice and health services need to come together to identify suitable processes for determining and managing accountability.
Safety assurance approaches for RAS in clinical settings	A descriptive scale of automation and autonomy levels was developed to enable reasoning about the capabilities of RAS in a clinical setting. Based on this, clinical scenarios were identified at different levels: baseline (level 1), automation (level 2) and autonomy (level 5). The clinical scenarios were analysed using three complementary approaches: FRAM, human reliability analysis and hazard analysis based on the NHS Digital clinical safety standards.

	The evidence generated in this way can be synthesised and summarised as follows:
6. FRAM can be used to understand work-as-done in a clinical system to inform the design of RAS	FRAM focuses on the performance variability of system functions, so what it does rather than its actual parts and composition. It has Safety-II foundations and so should be more aligned with how everyday safety is created the majority of the time, rather than trying to identify low frequency - high consequence events. It views deviations, goal conflicts and inherent trade-offs as necessary and normal. It tries to build a better understanding of work-as-done, not how work can fail. From this perspective an exemplar FRAM issue would be why a written prescription is rarely complete despite official guidance that says it should be. This issue is not written off as an error or non-compliance issue, but represents an opportunity for learning: to understand how this variability depends on the type of drug, the experience of the doctor and the nurse, the context, time pressure, etc. and why this adaptive behaviour happens for good reason.
7. Human Reliability Analysis provides a structured approach for investigating potential human – RAS interaction failures	Human Reliability Analysis techniques such as SHERPA focus on a detailed task analysis, human failure analysis and Performance Influencing Factors (PIF) analysis to understand what is driving human failure risks. This is very error orientated. However, consensus groups of subject matter experts (SMEs) are an explicit part of the method, so the task analysis is grounded in frontline worker experience while being informed by management and safety engineers. So, going beyond error management, this technique also looks at optimising system design and developing best practice. This method has cognitive science and task analysis as its foundation.

From this perspective an exemplar issue would be something like “right action on wrong object”, e.g. a label printed and placed on the wrong syringe. The method would then inspect the PIFs that influence this and seek to design the situation to eliminate these risks or make them less likely. Non-compliance would also be of interest, but more to understand the PIFs from the frontline that influence this rather than bluntly trying to reinforce the rules.

8. NHS Digital clinical safety standards and SMART are useful to identify key hazards at a higher level of abstraction

The NHS Digital clinical safety standards and the SMART software tool focus on identifying hazards and their prevention barriers and mitigation barriers using the bowtie method. This looks at the number and quality of barriers to prevent the hazard and stop the ultimate outcome we are trying to avoid. Barriers can have degradation factors and controls. SMART also uses process diagrams to build up picture of the task as this is not captured in bowtie analyses. The main hazards and barriers can be identified without going into the details of a fine-grained task analysis. This type of analysis should be familiar to safety engineers and can be quite technical.

From this perspective an exemplar issue would be something like the autonomous infusion pump wrongly assumes it has authority to operate outside of clinical guidelines when in fact no authority has been granted. Typically, this approach is less likely to engage with the more intricate issues to do with trade-offs identified in FRAM and the psychological details that SHERPA engages with.

RECOMMENDATIONS

1. Strengthen the relationship between patients and their clinicians when RAS are introduced.	Behind every data point that is used to train algorithms for use in clinical settings there is a patient story and a human life. Patients in intensive care are particularly vulnerable and have a strong bond with their clinicians. The use of RAS in clinical settings should include consideration and design of the patient experience and protect and strengthen the relationship between patients and their clinicians. RAS can improve efficiency and free up clinicians' time, which could be used for patient care, but there is a danger that clinicians might be asked to supervise and "care" for several RAS instead. It is important that clinicians do not spend less time with patients as more tasks are taken over by RAS.
2. Deliver training to enable clinicians to maintain core clinical skills, to provide clinicians with a baseline understanding of AI, and to educate clinicians about limitations of AI.	When the RAS fails or becomes unavailable, staff need to remain vigilant and be able to take over. They require training and exposure to maintain their core clinical skills. Clinicians will become users as well as supervisors of RAS. The training needs to provide clinician with a baseline understanding of how AI works so that they are able to identify limitations and problems. Staff might rely too much on RAS. They require education about limitations of AI to help address over-reliance.
3. Consider introduction of new AI specialist roles	It is unreasonable to expect frontline clinicians to have an expert understanding of AI and ML technologies. In addition, they should not be expected to spend more time with the technology than with their patients. The introduction of RAS into clinical systems will create a wealth of context-specific data that could be used to enhance clinical processes as well as the performance of the RAS itself.

	Novel roles, such as an AI specialist nurse, should be developed with a remit to support the introduction, operation and maintenance of RAS in their respective clinical settings.
4. Perform hazard analysis at the level of the clinical system or pathway of which the RAS will be part of.	The focus of hazard analysis and safety assurance should move on from the narrow focus of RAS in isolation to consider how the RAS will be integrated into clinical systems. Hazard analysis should be based on a thorough understanding of work-as-done. FRAM can be used to study work-as-done and performance variability in everyday clinical work. Human Reliability Analysis approaches are useful to study systematically human – RAS interaction failures. Bowtie analysis can be used to investigate hazards at a higher level along the clinical pathway.
5. Design for situation awareness	Clinicians build situation awareness as an implicit by-product in everyday clinical work, e.g., due to the close and repeated interaction with prescriptions, the patient and their vital signs, and the adjustments they make to treatments. The introduction of RAS into a clinical system will automate some of these tasks, and this might disrupt the implicit maintenance of situation awareness by staff. Hence the design of clinical systems with integrated RAS needs to consider this explicitly. Design solutions include dashboards that follow good information visualisation principles. Alarms and information-only indicators can alert clinicians to important developments. There might also be times where situation awareness is needed more than other times, e.g., during handovers between staff and where the RAS is reaching a state where it can no longer cope with blood glucose management and may need to hand back control to staff.

	Improved situation awareness can also improve trouble shooting if there are issues and actions to support patient care.
6. Design for handover	The RAS needs to be able to recognise its own performance boundaries, project into the future clinical scenarios that will be beyond its performance boundaries and identify suitable ways to hand over control to the clinician. Handover includes consideration of: (a) when to hand over; (b) whom to hand over to; (c) what to hand over; and (d) how to hand over. A handover could occur if the RAS requests to operate outside of clinical guidelines, but authority to do so is not given by the human operators. This is a Human Factors design challenge because the designer needs to determine how early the system should make this request. Also, one should not assume that staff will answer immediately, and so how long the RAS should wait, what it does in the meantime and what it does should an answer not be forthcoming all need to be thought through. The mismanagement of handover could have significant adverse safety implications. These contingencies and timings should be investigated so best practices can be determined.
7. Design for performance variability	Clinicians need to manage competing organisational priorities and operational demands. They use their experience and judgement to make trade-offs based on the requirements of a specific situation. The RAS needs to support rather than constrain this performance variability and adaptive capacity. Many operational constraints (e.g., limited number of access points to infuse drugs) will not change (i.e., be resolved) with the introduction of RAS. The flexibility to deal with them appropriately needs to be designed into the clinical system integrating RAS.

	Lack of attention to the need for performance variability could not only lead to frustrations and inefficiencies, but also safety issues.
8. Promote existing best practice and establish an integrated safety governance framework for AI regulation in healthcare	Existing best practices in the development of safety-critical systems and medical devices should form the foundation for the development and assurance of RAS in healthcare. Awareness of these and capability in their use should be promoted so that new stakeholders (e.g., AI developers) in this area can draw on these experiences. Post-market surveillance for learning technologies, the management of AI evolution, the communication between manufacturers, users and regulators, and issues of ownership of data and liability aspects require a broad consensus. A dialogue has been started between national regulators and NHS stakeholders (including MHRA, NHSx, NHS Digital, CQC and BSI), professional bodies (e.g., Chartered Institute of Ergonomics and Human Factors) and researchers. Such a whole systems approach is required to define clear interfaces between the different AI safety facets, and to ensure ownership and acceptance. Specifically for the NHS, this should consider inclusion of the different nations.

FUNDING

The Assuring Autonomy International Programme (AAIP)

Word count = 2,410

ACKNOWLEDGEMENTS

We would like to thank all the participants involved in this research for their time and contributions to the interviews. We are grateful to the study organisation for agreeing to participate in this study.

We thank the members of the project advisory group for their input and advice. We are especially grateful to our patient representatives Kath Grundy and Howard Grundy.

We are also grateful to Dr David Embrey (Human Reliability Associates), who facilitated one of the SHERPA sessions.

DISSEMINATION

The project outputs have been published in:

SUJAN, M., WHITE, S., HABLI, I. & REYNOLDS, N. 2022. Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare. *SSRN* [Online]. Available from: https://dx.doi.org/10.2139/ssrn.4000675.

SUJAN, M. 2021. Muddling Through in the Intensive Care Unit. *In:* BRAITHWAITE, J., HOLLNAGEL, E. & HUNTE, G. (eds.) *Resilient Health Care V6 - Muddling Through With Purpose.* Boca Raton: CRC Press.

SUJAN, M., FURNISS, D., HAWKINS, R. & HABLI, I. 2020. Human Factors of Using Artificial Intelligence in Healthcare: Challenges that stretch across industries. *In:* PARSONS, M. & NICHOLSON, M. (eds.) *Safety-Critical Systems Symposium.* York: Safety-Critical Systems Club.

LAHER, S. & SUJAN, M. 2020. Assurance challenges for Artificial Intelligence and Machine Learning in Healthcare. *Safety Systems,* 28.

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. *Applied Ergonomics,* 86.

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 Informatics,* 26**,** e100081.

SUJAN, M., FURNISS, D., EMBREY, D., ELLIOTT, M., NELSON, D., WHITE, S., HABLI, I. & REYNOLDS, N. 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).* Hannover: CRC Press.

SUJAN, M. & FURNISS, D. 2019. Using AI in patient care. *The Ergonomist***,** 7.

APPENDIX 1 – IDENTIFYING HAZARDS (BOK 1.1)

Identifying hazards is an important function in the design and safety assurance of RAS. This function will be variable in its level of success. Some of the drivers for this performance variability will be intrinsic to the function (e.g. the experience of the analyst performing the function and the method they choose to use), some will be extrinsic to the function (e.g. the time and resources available for activities to identify hazards), and will be functionally coupled to wider functions upstream and downstream in the system (e.g. the analyst might have done a similar project before, this enhances their choice of method and the hazards they "see" for this project, this sparks more grounded debate and ideas with the subject matter experts and engineers, which leads to design improvements).

Here, we expand the notion of "identifying hazards" not just as a technical issue that focuses on the mechanical application of methods, but a sociotechnical issue that includes the skills, knowledge and experience of the analyst, who the rest of the team are and how they are involved, the processes that are followed, time and resources allowed, and the concepts and theory that guides thinking. As we see below, these drivers can be mapped so we have a better idea of what makes the performance of "identifying hazards" flourish rather than stall.

Scope of analysis	Identifying hazards for RAS in real world settings can be complex. In such cases simplifying assumptions might be made about working practices, and the scope of analysis. However, a study focused on the technology and the primary task would give a quite different perspective compared to a study focused on the context (e.g. clinical pathway) and primary and secondary tasks and broader related activities.
Granularity of analysis	Time, resources and perspective can also affect the granularity of the analysis. There is a trade-off between the efforts one expends and the value one gets back, presumably with diminishing returns. However, some subtle interactions and unintended consequences might only reveal themselves at a fine-grained level of detail.
Experience of analyst	The experience of the analyst leading the hazard identification exercise will have a significant effect on how it is organised,

	who is involved and what processes are followed. The analyst might also have specific skills and knowledge to enlighten the hazard analysis.
Engagement with subject matter experts (SME) and stakeholders	The analyst will only be able to "see" so much. SME's and stakeholders need to be engaged with effectively to bring their knowledge, experience and insight to enlighten the hazard analysis. Who is involved and how they are engaged will influence success.
Representations	Communicating how the task is currently done, and how the task might be reconfigured with a RAS, can be complex. Different representations can be used (e.g. process maps, task analyses and functional diagrams). Pictures and diagrams might also convey issues to do with the context, layout and interface design. All of these representations have strengths and limitations, they will shape the sort of dialogue and feedback that can be achieved with SME's and stakeholders.
Concepts, theory and guidewords	Different approaches and methods will have different concepts, theory and guidewords that will shape thought and dialogue. For example, more traditional engineering-based approaches might focus on technical issues, whereas human factors approaches might more readily draw attention to issues of situation awareness and attention. Methods focusing on a single task might miss issues with important goal conflicts and trade-offs between activities. Methods focused on failure might miss important resilience mechanisms that help to create safety.

Indeed, there is some suggestion from recently literature that to ensure system safety we must not only attend to identifying hazards and reducing risks following the ALARP principle (Safety-I), but that we must also understand the (sometimes hidden and implicit) positive behaviours that create safety

(Safety-II). We must have a good understanding about how safety is normally created in everyday work, otherwise the introduction of RAS might inadvertently erode resilience behaviours. For example, the official view of the system might be clear that verbal medication orders should not be taken and medication prescriptions should always be complete, however enforcing these things could lead to delayed medication, workarounds, non-compliance and disuse. Sometimes seemingly erroneous behaviour is practiced to keep the system safe.

Identifying hazards will not be perfect and factors driving its performance need to be understood.

APPENDIX 2 – DEFINING THE OPERATING ENVIRONMENT (BOK 1.1.2)

Healthcare is a complex and diverse setting with many different operating environments. A family doctor's practice is very different from a hospital setting, and even within a hospital there is diversity across operating environments such as surgery or the hospital pharmacy. Reflecting this broad range of potential operating environments is the large number of different types of artificial intelligence (AI) and machine learning (ML) applications in healthcare. Examples include clinician-facing applications (e.g. breast cancer screening algorithms), patient-facing mobile phone apps (e.g. symptom checkers) and tools to support healthcare business processes (e.g. missed appointment predictors).

The definition of the operating environment can, therefore, be challenging for developers of AI and ML applications in healthcare. Drawing an accurate boundary around the AI / ML system and the operating environment is not straightforward, and can be done in different ways. To date, most developers have bounded the AI / ML system very narrowly and assumed a well-defined task or function in order to reduce complexity. For example, one way of looking at an algorithm for breast cancer screening is to consider only a set of mammograms as input and the likelihood of malignancy as the output. However, this approach runs into difficulties quickly when the wider use context needs to be considered, for example when an algorithm trained on data from a specific population or health system (e.g. patients in the NHS in the UK) is deployed in another population or health system (e.g. patients in the US). Performance figures tend to drop quickly in these situations.

Another option is to define the operating environment as the clinical system within which the AI / ML will be used. This perspective recognises that the AI / ML interacts with other technology and with people. Care is generally delivered by teams of healthcare professionals working as clinical teams, and supported by a large number of tools and technologies. AI and ML systems, even with increasing autonomy, might be best understood as part of such clinical teams.

A useful approach to model clinical systems at the functional level is the Functional Resonance Analysis Method (FRAM). FRAM decomposes the clinical system into functions, to move away from "what a system is" to "what it does". Each function is examined for its potential performance variability, then interactions between functions are examined. "Functional resonance" is used to describe how outcomes can "emerge" from everyday variability of many functions, to move away from simple notions of "cause and effect". FRAM is built on four principles:

- The principle of equivalence of success and failure – Success and failure come from the same source, i.e. they are not fundamentally different in nature. Approximate adjustments mean that people adapt successful most of the time but sometimes variability in performance will lead to unsatisfactory outcomes.

- The principle of approximate adjustments – Due to limitations in resource, uncertainties, underspecified systems and variance demands people will adjust to suit the situation. This gives rise to performance variability which is inevitable, ubiquitous and necessary.

- The principle of emergence – Complex systems with many links and fluctuating approximate adjustments become intractable as it is impossible to predict what will happen precisely beyond expecting regular events.

- The principle of functional resonance – Functions represent the different things a system does. Due to approximate adjustments these will exhibit performance variability. Functional resonance refers to how functions may impact each other's performance variability. Small changes could lead to disproportionally large effects and vice versa.

The strength of FRAM is that it supports the analyst or system designer in reasoning about interactions. For example, when introducing an autonomous infusion pump into the intensive care unit, FRAM encourages consideration of not just the algorithmic performance (e.g. whether the infusion pump can control a patient's blood sugar levels by giving insulin), but also of how the autonomous infusion pump communicates with nurses and doctors as well as other systems, such as the electronic patient record. This provides a more realistic representation of the complexity of the operational environment in healthcare settings.

APPENDIX 3 – DEFINING OPERATING SCENARIOS (BOK 1.1.3)

The designers of artificial intelligence (AI) and machine learning (ML) applications need to scope, bound and articulate clearly the situations for which the application is going to be used, and how it is going to be used. In the case of clinical settings, it is very likely that even autonomous systems will have a significant degree of interaction with people. For example, an autonomous infusion pump will require interaction with the nurse in case of unexpected patient deterioration.

It is important that the definition of operating scenarios is done based on operational realities (work-as-done) rather than through an abstract view of what should be done in principle (work-as-imagined). Typically, a range of situations needs to be considered, such as routine operational scenarios, exceptional or emergency response scenarios, and maintenance and inspection scenarios. Understanding of the operational scenario includes consideration of what specifically needs to be done by the application and by any users, in what kind of order different activities need to be done, what kinds of information are required to complete an activity, what forms of interactions and communication take place, and what other activities people interacting with the application might be engaged with at the same time.

Definition of operating scenarios can make use of analysis techniques for understanding and representing clinical work. Examples include Hierarchical Task Analysis (HTA) and Functional Resonance Analysis (FRAM).

HTA represents human activities based on a theory of goal-directed behaviour, and includes a hierarchy of goals and sub-goals linked by plans, which describe how sub-goals combine to achieve the higher-level goal. Plans can be used to express any kind of algorithm, e.g. simple sequential ordering (such as do step 1 to step 3 in order), free ordering (do steps 1, 2, 3 in any order), as well as more complex loops (such as do step 1 and step 2 in order until signal A is active, then do step 3). This representation creates a tree-like structure, where the leaves represent task steps that are considered elementary (e.g. basic manual operations) or where further decomposition is not considered necessary.

FRAM decomposes the clinical system into functions, to move away from “what a system is” to “what it does”. Each function is examined for its potential performance variability, then interactions between functions are examined. “Functional resonance” is used to describe how outcomes can “emerge” from everyday variability of many functions, to move away from simple notions of “cause and effect”.

APPENDIX 4 – IDENTIFYING HAZARDOUS SYSTEM BEHAVIOUR (BOK 1.2)

One of the main mechanisms for identifying hazards and error prone conditions are the methods used to help identify hazardous system behaviour. Methods shape thinking and dialogues, and influence what can be "seen" in the context before the RAS intervention and what may happen when the RAS intervention is deployed. Methods will influence requisite variety, i.e. the ability to foresee issues that may arise in future systems that do and do not yet exist.

Understanding the coverage, strengths and weaknesses of a method is important for its determining its adequacy for identifying hazardous system behaviour. However, it is impossible to run method comparison studies that do not suffer from confounding variables. For example, there is always the "evaluator effect", and even if you keep the same evaluator then they learn over successive applications of different methods to the same area, which means that the study is then confounded. Furthermore, where some methods engage with stakeholders and subject matter experts (SMEs) then their contributions does not necessarily have to be aligned with the method, serendipity may help discover insights. Accepting these limitations, we may still compare the foundational theory, concepts and representations that are tied up in the use of methods, which has consequences for understanding system safety.

1) Functional Resonance Analysis Method (FRAM)

FRAM focuses on the performance variability of system functions, so what it does rather than its actual parts and composition. It has Safety-II foundations and so should be more aligned with how everyday safety is created the majority of the time, rather than trying to identify low frequency - high consequence events. It views deviations, goal conflicts and inherent trade-offs as necessary and normal. It tries to build a better understanding of work-as-done, not how work can fail.

From this perspective an exemplar FRAM issue would be why a written prescription is rarely complete despite official guidance that says it should be. This issue is not written off as an error or non-compliance issue, but represents an opportunity for learning: to understand how this variability depends on the type drug, the experience of the doctor and the nurse, the context, time pressure, etc. and why this adaptive behaviour happens for good reason.

2) Systematic Human Error Reduction and Prediction Approach (SHERPA)

SHERPA focuses on a detailed task analysis, human failure analysis and Performance Influencing Factors (PIF) analysis to understand what is driving human failure risks. This is very error orientated. However, consensus groups of subject matter experts (SMEs) are an explicit part of the method, so the task analysis is grounded in frontline worker experience while being informed by management and safety engineers. So, going beyond error management, this technique also looks at optimising system design and developing best practice. This method has cognitive science and task analysis as its foundation.

From this perspective an exemplar SHERPA issue would be something like "right action on wrong object", e.g. a label printed and placed on the wrong syringe. The method would then inspect the PIFs that influence this and seek to design the situation to eliminate these risks or make them less likely. Non-compliance would also be of interest, but more to understand the PIFs from the frontline that influence this rather than bluntly trying to reinforce the rules. Something more out of scope of SHERPA would be technical issues like the autonomous infusion pump fails to communicate with the health IT system because the network is down, or updates to health IT software meaning current request for authority to operate outside of clinical guidelines (extended autonomy) is cancelled.

3) Safety Modelling, Assurance and Reporting Toolset (SMART)

SMART focuses on identifying hazards and their prevention barriers and mitigation barriers using the bowtie method. This looks at the number and quality of barriers to prevent the hazard and stop the ultimate outcome we are trying to avoid. Barriers can have degradation factors and controls. SMART also uses process diagrams to build up picture of the task as this is not captured in bowtie analyses. The main hazards and barriers can be identified without going into the details of a fine-grained task analysis. This type of analysis should be familiar to safety engineers and can be quite technical.

From this perspective an exemplar SMART issue would be something like the autonomous infusion pump wrongly assumes it has authority to operate outside of clinical guidelines when in fact no authority has been granted. Typically, SMART is less likely to engage with the more intricate issues to do with trade-offs identified in FRAM and the psychological details that SHERPA engages with.

The choice of method will impact the understanding of system safety, which will in turn impact design and safety management.

APPENDIX 5 – CONSIDERING HUMAN-MACHINE INTERACTION (BOK 1.2.1)

Artificial intelligence (AI) and machine learning (ML) applications in healthcare are often evaluated on narrowly defined tasks. However, the real challenges for the adoption of AI and ML will arise when algorithms are integrated into clinical systems to deliver a service in collaboration with clinicians as well as other technology. It is at this clinical system level, where teams consisting of healthcare professionals and AI systems cooperate and collaborate to provide a service, that human factors challenges will come to the fore.

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. 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. 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 and ML in patient care. Such human factors challenges 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.

The table provides an illustration of human factors issues that might require consideration in the example of the design of an autonomous infusion pump to be deployed in the intensive care setting.

HF Challenge	**Description**	**Example**
Handover	The autonomous system needs to be able to recognise its own performance boundaries, project into the future clinical scenarios that will be beyond its performance boundaries, and identify suitable ways to hand over control to the clinician. Handover includes consideration of: (a) when to hand over; (b) whom to hand over to; (c)	The patient's blood sugar levels do not respond sufficiently to the insulin given by the autonomous infusion pump. The pump predicts and recognises that it will not be able to control the patient's blood sugar. The pump triggers an alert on the electronic health record, raises an audible alarm, and requests the nurse to take over. The nurse can review the reason for the alert, the history of the pump's insulin management,

	what to hand over; and (d) how to hand over.	and its projection into the future, and act accordingly.
Performance Variability	Clinicians need to manage competing organisational priorities and operational demands. They use their experience and judgement to make trade-offs based on the requirements of a specific situation. The autonomous system needs to support rather than constrain this performance variability and adaptive capacity.	The nurse realises that insulin has not yet been prescribed for the patient even though they will likely need it. The nurse goes and finds the doctor, explains the situation, and the doctor issues a verbal medication order and will follow this up with the written prescription later (performance variability). The autonomous system requires an electronic medication order, but allows for a manual override. The autonomous system sends reminders to the doctor with a request for completing the electronic medication order.
Automation bias	When a system works well most of the time, clinicians start to rely on it. In some situations, this can lead to overreliance, for example when the system takes an inappropriate action but the clinician does not recognise this because they trust the system.	Due to sepsis the patient requires tighter control of blood sugar levels than usual. The autonomous system has managed successfully septic patients before but, in this instance, fails to recognise the need for tighter glycaemic control. The autonomous system provides clinician interpretable justification and explanation of its decisions, and the clinician, who has received training on potentially inappropriate behaviours of the autonomous system, is able to spot the discrepancy and act accordingly.
Supervision	Clinicians are both users and supervisors of the autonomous system. They need to understand not only how to operate the autonomous system (e.g. loading a	The autonomous infusion pump is operating on the sliding scale algorithm for administering insulin. It classifies the patient's response to the current insulin infusion as requiring transition to another

	syringe), but also how to recognise potential failure modes or deviations from appropriate behaviour or changes in the environment that might move the autonomous system outside of its design envelope.	scale with 70%, as opposed to 30% for staying within the current scale. The autonomous system initiates and the transition, and activates an “uncertainty marker” to alert the clinician.

APPENDIX 6 – VALIDATION OF SAFETY REQUIREMENTS (BOK 1.3.1)

The introduction of artificial intelligence (AI) and machine learning (ML) applications into clinical systems can create challenges for traditional design approaches that require clearly defined and precise specifications of the operating environment, operational scenarios and of the resulting safety requirements that bound the behaviour of the AI / ML system. Healthcare is a complex domain, and clinical systems are made up of many different actors and technologies all interacting with one another in ways that can be very dynamic and responsive to the requirements of a specific situation.

One way of both eliciting and validating safety requirements (though not the only one, nor even a standalone approach) is to seek input and feedback from stakeholders, e.g. through simulation or interviews. This approach is particularly appropriate where human – machine interaction and training are concerned.

For example, in the case of the design of an autonomous infusion pump to be used in intensive care, input from stakeholders might produce training requirements as shown below. Feedback from stakeholders on a design prototype can then provide information about the extent to which the requirements have been met.

Clinicians need to maintain core clinical skills	When an autonomous system fails or becomes unavailable, staff need to remain vigilant and be able to take over. They require training and exposure to maintain their clinical skills.
Clinicians need to build a baseline understanding of AI and its limitations	Clinicians will become users as well as supervisors of AI systems. They shall be provided with a baseline understanding of how AI works so that they are able to identify limitations and problems.
Training needs to address over-reliance on AI	Staff might rely too much on AI. They shall receive training in core clinical skills and education about limitations of AI to help address over-reliance.

Similarly, high level safety requirements relating to autonomy and control might be validated through feedback on a proposed interaction design.

Clinicians need to be able to maintain autonomy.	Clinicians feel responsible for their patient and want to remain in control. Autonomous systems can challenge this sense of autonomy, and clinicians need to be allowed to remain in charge, e.g. through manual override options.
Feedback and alerts shall provide clinicians with an awareness of what the AI is doing	Feedback and alerts can help to maintain situation awareness and stay in control of the overall treatment and care for the patient. The design shall determine clearly when an alert is raised. The system shall avoid alert fatigue or overload.
Clinicians need to be able to build trust in AI	Clinicians have to trust AI in order to realise its benefits. The interaction design shall include training and feedback. The AI system shall be introduced gradually in low-risk areas over time.

APPENDIX 7 – DEFINING SAFE SYSTEM RESPONSES TO CHANGES (BOK 2.6)

Healthcare is delivered in a highly dynamic and non-deterministic environment, and successful outcomes are dependent on actions and decisions made by humans. However, humans are fallible and unintentional errors and mistakes have led to unsafe care outcomes. Application of artificial intelligence (AI) offers great potential in this domain by: automating routine tasks that are susceptible to human error e.g. transcribing prescriptions and printing syringe labels; and working autonomously to monitor and manage care scenarios e.g. optimising insulin infusion regime.

However, by removing the HCP from the real-time, closed loop, care pathway there is a significant risk that they will lose their situational awareness and their ability to deliver effective care could be compromised. So, whilst there is a great opportunity to improve efficiency within healthcare, careful consideration needs to be given to monitoring and handover protocols such that effective human intervention occurs should the AI's behaviour exhibit characteristics that could cause or contribute to patient harm.

The following guidelines provide a framework which will support effective monitoring and handover between AI technology and HCPs

Upskill HCPs	HCPs will need to establish an understanding of the technology, its capabilities and weaknesses so they are better placed to recognise anomalous behaviour.
Baseline and understand care pathway	Care pathways need to be defined and baselined (representing work as is done) so that the contribution and authority of AI is clearly expressed and understood within the clinical team.
Define AI capability	The specific capability that the AI is providing needs to be defined and characterised in the context of supporting the care pathway. This must consider the interaction between the AI and the HCP both as a user and also a supervisor.

	This must consider the authority limits autonomous AI can have. This must consider the monitoring and alerting mechanisms and whether these are undertaken by the AI itself or independently by another element of the care system.
Conduct pre-emptive hazard analysis	Need to understand the potential patient-level harm effects that could occur in the care-pathway and the specific contributions AI could make. The severity of harm outcome and the significance of the AI contribution will impact the definition of the following key activities.
Develop monitoring SOP	Need to develop a regime within the care-pathway which will ensure the continued safe operation of the AI. This will be dictated by the capability that the AI is providing (automation and/or autonomy) but typically would need to consider: Pro-active or re-active: HCP routinely monitors behaviour of AI or responds to an alert or alarm. Frequency: sufficient to maintain situational awareness but not so frequent that it compromises efficiency Trends: does the monitoring indicate progression to a unsafe state or imminent point of handover. Escalation: is there a need for a monitoring HCP to seek second opinion or high authority before initiating any action Monitoring architecture: this needs to be defined and consideration given to whether the AI simply monitors itself, whether the AI is monitored independently by another element of the care system or whether it's a combination of both.

Develop hand-over SOP	Need to develop a regime within the care pathway, which will ensure timely HCP intervention when required. This will be dictated by the capability that the AI is providing (automation and/or autonomy) but typically would need to consider: Definition of safe limits: the limit of authority the AI can have before HCP intervention is required needs to be defined. Definition should consider the need for soft limits i.e. those that can be transgressed but signify an impending handover requirement. Hard limits i.e. those that must never be exceeded need to be defined. The protocol needs to consider the degree of authority the AI has; does it have the same as a HCP or is it restricted to a lower level. Definition of transfer state: the AI's behaviour whilst a handover is being determined needs to be specified. Does the AI continue to perform its function which may result in a change in outcome, does it maintain a steady state of output or default to previous know good (safe) output. The period of time that the transfer state can persist needs to be defined. Definition of safe state: the AI's behaviour once authority has been relinquished needs to be defined. Does the AI revert to a "off" state and dissociate itself from the care pathway or does it continue to function in "hot standby" in order to support a subsequent transfer of authority back from the HCP. Definition of re-engagement criteria: the criteria and process for re-engagement of the AI needs to be defined. Definition of audit log: an audit log may be needed to support informed and safe handover of authority to a HCP. It will be necessary to identify those clinical variables, environmental conditions and system parameters that influenced the learning. The HCP will need to be able to quickly assimilate the clinical scenario and take effective mitigating action.

Simulation and dry-runs	Need to train HCPs in execution of the SOPs potentially through simulation (outside of the care environment) and dry-runs (inside the care environment) of hazard scenarios. This needs to verify the effectiveness of the SOP and the ability of the organisation to follow it in real-life scenarios.
Re-active incident management	There is a need for an organisation to recognise when handover between the AI and HCPs has resulted in an incident or near-miss. This needs to be accommodated in the organisation's existing service / safety management process. Such events need to be reviewed and the impact on the organisation's safety case and understanding of AI technology considered.

ASSURING
AUTONOMY
INTERNATIONAL PROGRAMME