Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210847

Hazards for the Implementation and Use of Artificial Intelligence Enabled Digital Health Interventions, a UK Perspective

Stuart HARRISON[a,1], George DESPOTOU[a] and Theodoros N. ARVANITIS[a]
[a]*Institute of Digital Healthcare, WMG, University of Warwick, UK*

Abstract. Background: Artificial Intelligence (AI) has seen an increased application within digital healthcare interventions (DHIs). DHIs use entails challenges about their safety assurance. Exacerbated by regulatory requirements, in the UK, this places the onus of safety assurance not only on the manufacturer, but also on the operator of a DHI. Clinical Safety claims and evidencing safe implementation and use of AI-based DHIs require expertise, to understand and act to control or mitigate risk. Current health software standards, regulation, and guidance do not provide the insight necessary for safer implementation. Objective: To interpret published guidance and policy related to AI and justify clinical safety assurance of DHIs. Method: Assessment of UK health regulation policy, standards, and AI institution insights, utilizing a published Hazard Assessment framework, to structure safety justifications, and articulate hazards relating to AI-based DHIs. Results: AI enabled DHI hazard identification, relating to implementation and use within healthcare delivery organizations. Conclusion: By application of the method, we postulate that UK research of AI DHIs highlighted issues that may affect safety, in need of consideration to justify safety of a DHI.

Keywords. Digital health, safety, justification, health system, artificial intelligence, hazard analysis

1. Introduction

Digital Health Intervention manufacturers collate evidence and justification about the safety of their products. This has also been presented as a safety case [1], to aid the interpretation, understanding and communication across stakeholders (e.g., regulatory assessment, client engagement). In addition to safety, effectiveness and security are also foundational elements of the lifecycle of all DHIs [2]. In addition to the manufacturer justifying safety, in the UK, the operator of DHI also needs to justify safe implementation and use. AI hazards are unsubstantiated in guidance, standards, policy and overall practice [2]. Although they are often acknowledged, there is currently little published providing a comprehensive insight in the safe implementation and use of AI DHIs [3]. Many implementations of AI-based DHI's rely on expert resources and insights from the manufacturer, to provide safety assurance prior to implementation or use of the DHI [4]. The dynamic nature of the involved algorithms, in AI, poses a challenge to regulation, requiring "real time" post-market product change, implementation, and use of the DHI.

[1] Corresponding Author, S. Harrison, Institute of Digital Healthcare, WMG, University of Warwick, CV4 7AL, UK; E-mail: stuart.harrison@warwick.ac.uk

This will involve feedback and regulatory insight during post-deployment of the intervention, thus, to demonstrate the ongoing applicability of any safety justifications made. We investigate current recommendations in this field, aiming to collate best practice recommendations from sources. We identify a list of high level (generic) AI hazards, through the application of a framework for safety justification of DHIs [5].

2. Method

This study followed a 2-step approach. Firstly, the identification of UK-based standards, regulation, legislation, health policy and guidance sources, and concerns regarding AI were consolidated. The sources selected from UK Government Regulators, policy makers and national institutions provide standards, regulation and insight to AI DHI implementation and use. We document issues from each source, while sorting and grouping to align to common themes, technical and operational areas. Secondly, we identify high-level hazards relating to AI, along with their contributing factors. This was achieved through a hazard identification and safety justification process as part of the method applied [5], where issues related to hazards (e.g., directly affecting the patient care), or as classes of failures that constitute contributing factors to hazards were retrieved. Analyzing issues, presented in literature, enables the postulation of hazard, hazardous situation, harm, and effect. Assumptions, about likelihood of hazard, are out of scope. Medical device risk management standards, such as ISO 14971 [6], provide suitable defined terms.

3. Results

Table 1 sees the safety related issues, summarized by the application of the hazard assessment framework, with a rationale from sources. Extrapolation and application of this framework to express issues in the form of hazard table is presented in Table 2.

Table 1. Safety related recommendations or issues

Recommendation	Summary	
Access to sensitive data	Provide trust and confidence in data sharing. Reduction in transaction costs of accessing data [7–12].	Data
Improve the availability of data. Data Quality and Maturity.	Access to AI training data, in compliance with regulation [2,7–14].	
Bias and Discrimination. Minimizing Bias.	Training data and algorithms need to be verified for objectivity and inclusivity, extrapolated to the real world [2,12–14].	
Representativeness	Misrepresentation of groups within data samples [12,14].	
Fit-for-purpose and sufficiency. Self-fulfilling prediction.	Justification of data quality and quantity for intended purpose. Reinforcement learning bias [7,9,10,12,14,15].	
Source integrity and Measurement Accuracy	Ensuring data sources have reliable and impartial methods of collection [11,12,14].	
Timeliness and recency.	Accuracy and currency of datasets [11,12,14,15].	
Relevance, Appropriateness and Domain Knowledge	Utilization of domain experts [10–12,14].	
Outcome fairness	Methods to ensure unbiased deployment of AI based systems [14].	Operational use
Decision-Automation Bias/The Technological Halo Effect. Automation complacency.	Over reliance on the system and inability to respond to failures [14,15].	
Automation-Distrust Bias	Reluctance to trust AI based system decisions [14].	
Stakeholder Engagement. Trustworthiness. Accountability.	Aimed to build trust and confidence. Improved regulation and health care monitoring of AI systems [2,7,9,11–14].	

Human centered implementation processes	AI development specific Human factors considerations [10–12,14].	
Accuracy and performance metrics. Unsafe Failure Modes.	Error rates for AI generated outputs. Data to predict an output and performance (e.g., accuracy) [14,15].	Performance
Reliability. Negative side effects. Reward hacking.	Consistency of behavior. Supervision for longer-term operational reliability goals and control [11,12,14,15].	
Robustness	A measure of a systems integrity [11,14].	
End-to-end AI Safety. Unscalable oversight.	Regular verification and validation of AI throughout its lifecycle [2,10–12,14].	
Concept Drift or Distributional Shift	Training data mismatch over time [10–12,14,15].	
Brittleness	Undetectable changes in input data leading to failures [14].	
Model hardening	Securing the AI system to combat adversarial attack [14].	
Misdirected Reinforcement Learning Behavior. Insensitivity to impact. Unsafe exploration.	Insufficient controls placed upon trial-and-error processing methods. Inefficient supervision or monitoring impacts outcome and efficiency aims [14,15].	
Transparency and explain-ability.	Explanation of processes, services and decision making by AI [7,9–15].	

Table 2 summarizes associated failures as a hazard table. Hazardous situation is the circumstance exposing patients/users hazard(s) aligned to medical device standards [6].

Table 2. Hazard Analysis & Results

	Hazard	Hazardous Situation	Harm	Effects
1	Incorrect clinical decision result or diagnosis.	Incorrect treatment plan or decision selected.	Incorrect diagnosis	Availability and quality of data; Insufficient data sample. Training error (e.g., bias). Incorrect usage of intervention; Inconsistent intervention performance; Recall, precision accuracy training for users.
2	Failure to operate as intended (annunciated).	Software failure of data error.	Delay in diagnosis and treatment.	Corrupt files and data, hardware failure, algorithmic (expected) errors.
3	Incorrect use of intervention as intended.	Effectiveness and suitability of intervention.	Delay or ineffective impact of patients.	Communication and validation of intervention. Insufficient documentation of use within the pathway. Poor adherence to guidelines. Inadequate supervision. Lack of algorithmic explain-ability.
4	Ineffective use of digital intervention.	Ineffective provision of healthcare services.	Delay or ineffective provision of service.	Unsuitable planning of use. Intervention suitability justification. Lack of buy-in from users. Wrong timeframe of evidence basis. Lack of co-production of intervention.

4. Discussion and Conclusion

The results, Table 2, highlight the increased clinical safety risk AI presents to Healthcare Delivery Organizations (HDO). The timeliness and accuracy of decision-making are two specific hazards highlighted by this analysis. By presenting the potential causes of such hazards, we can direct effort applied in risk mitigation to relevant sources with greater likelihood of success. Inadequate utilization of data, social acceptance or trust of AI technology influences clinical effectiveness and operational benefit to larger patient cohorts. Issues with algorithmic functions and data specific facets of bias, validation and revalidation have the potential to cause direct harm to a patient. There is risk in assuming increased accuracy/efficiency from AI decision-making and limited human involvement. A commissioning HDO has a direct relationship with effectiveness, safety, and clinical outcomes. By application of the method, we postulate that UK sources have highlighted

issues that may affect justification of safety of a DHI, Table 2. The complexity and dynamic nature of risk, together with increased HDO control of potential hazards, enhances the need for more effective methods for communicating safety justification. Addressing the causes of these issues, a body of evidence (sources) will support the clinical safety of the AI-based DHI. The correct use of DHIs adds greater burden on organizations, as unintended operation / use exposes risk beyond the immediate decision support function, and into future decision-making and algorithmic learning of the DHI. Operating AI-based DHIs as intended, correct implementation within the clinical/patient pathway and effective periodic review of clinical outcomes would enhance safety claims. Manufacturers must consider collaborative engagement with HDOs to establish proven in use safety, reliability, and efficacy claims. Simplifying the explanation of decision-making, diagnosis and foundation of operation may enable intelligent supervision to mitigate the clinical risk. The results align with current EU policy and industry body recommendations for the use of AI DHI, including medical devices [16].

References

[1] Royal Academy of Engineering. Establishing high-level evidence for the safety and efficacy of medical devices and systems. 2013;(January 2013):1-40.
[2] Rowley A. The emergence of artificial intelligence and machine learning algorithms in healthcare: Recommendations to support governance and regulation. BSI Gr. 2019;1-18.
[3] Glauner P. An Assessment of the AI Regulation Proposed by the European Commission. 2021;(May 2021).
[4] Shaw J. Artificial Intelligence and the Implementation Challenge. J Med Internet Res. 2019;21(7).
[5] Despotou G, Ryan M, Arvanitis TN, Rae AJ, White S, Kelly T, et al. A framework for synthesis of safety justification for digitally enabled healthcare services. Digit Heal. 2017;3(April):205520761770427.
[6] ISO. ISO - ISO 14971:2019 - Medical devices - Application of risk management to medical devices.
[7] UK Gov. Accelerating Artificial Intelligence in health and care: results from a state of the nation survey. 2018;1–64.
[8] Hall W, Pesenti J. Growing the Artificial Intelligence Industry in the UK. 2017;
[9] Thorpe GJ. The U.K.: ready, willing, and able. J Palliat Care. 1988;4(1–2):26–8.
[10] UK Gov. A guide to good practice for digital and data-driven health technologies. 2021.
[11] UK NHSx. A Buyer's Guide to AI in Health and Care.
[12] NHSX. Artificial Intelligence: How to get it right. 2019;(October):1–55.
[13] Harwich EE. Thinking on its own: AI in the NHS. Reform. 2018; Jan:1–60.
[14] Leslie D. Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. SSRN Electron J. 2020;
[15] Challen R. Artificial intelligence, bias and clinical safety. BMJ. 2019.
[16] Gudeppu M. Medical device regulations. In: Trends in Development of Medical Devices. Elsevier Inc.; 2020. p. 135–52.