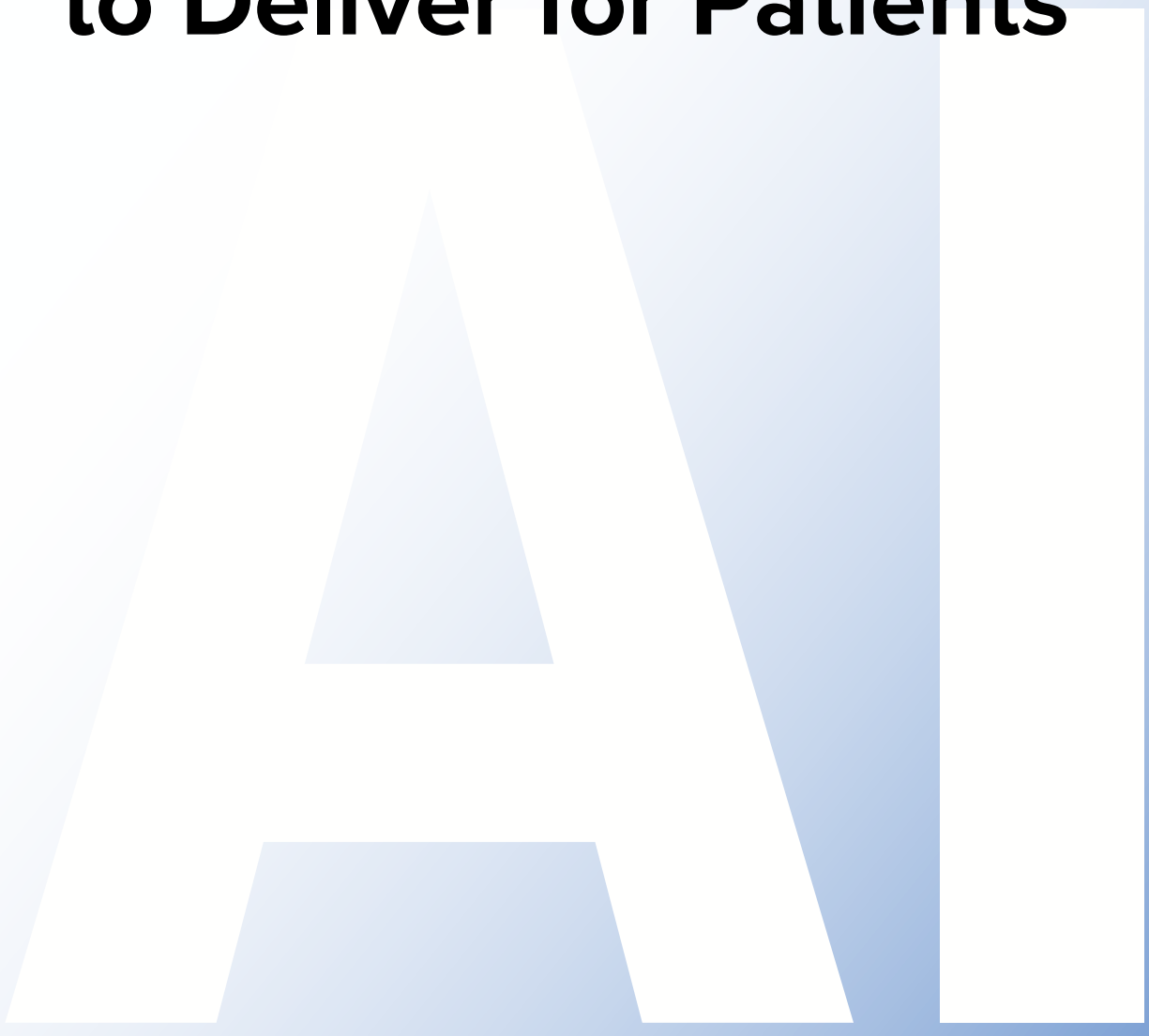
University of York | Improvement Academy | The MPS Foundation | Centre for Assuring Autonomy | NHS Bradford Teaching Hospitals NHS Foundation Trust

Avoiding the AI 'Off-Switch':

Make AI Work for Clinicians, to Deliver for Patients



FOREWORD

by the MPS Foundation

The MPS Foundation focuses on creating sustainable global change in patient safety and improving patient outcomes. We champion research that helps clinicians, healthcare professionals, and healthcare providers understand and mitigate the risks they face, and improve their wellbeing. To achieve this, we look to support ambitious cross-disciplinary research that can make a difference. The research that underpins this White Paper is an example of that.

The MPS Foundation understands that, more than in any other period in history, healthcare is undergoing rapid change, driven by advances in technology that fundamentally impact healthcare delivery, how patients interact with the profession, and patient outcomes. The potential opportunities provided by technology are only limited by one's imagination. However, it also presents risks for the healthcare profession - chief among them, the challenge of keeping up with the pace of change and ensuring clinicians remain informed users rather than servants of the technology. In response, one of the MPS Foundation's strategic priorities is to:

- Research the opportunities and risks to clinicians and healthcare professionals from digital integration and technology and its impact upon patient safety, care and outcomes.

The MPS Foundation was proud to support the Shared CAIRE (Shared Care AI Role Evaluation) project in 2022 as part of our first annual Grant Programme. The research addressed one of our critical strategic research priorities and seemed very relevant. We did not understand quite how relevant until ChatGPT was launched in November 2022, after our decision to support the project had been made. Immediately the idea of Artificial Intelligence, as exemplified by Generative AI, moved from the conceptual to reality.

The MPS Foundation welcomes this White Paper, based on the findings of the Shared CAIRE research. While AI technologies and their use in healthcare continue to evolve rapidly, the recommendations made in this White Paper will remain relevant. AI offers great opportunities to improve patient outcomes and the working lives of healthcare professionals. There are, however, real risks at every stage – from AI development and regulation, through to informed acquisition and safe use. To maximise benefits for patients, healthcare AI technologies should be designed and developed in collaboration with clinicians and healthcare providers. Healthcare organisations and professionals purchasing or using AI decision-support tools should be aware of both their potential benefits and their associated risks.

Healthcare professionals will need to embrace the changes AI technologies bring and engage with these advancements. The genie cannot be put back in the bottle - nor should clinicians want it to. However, it is important that clinicians understand both opportunities and risks presented by AI, and how to interact with emerging technology in a way that ensures both they and their patients are the beneficiaries. This White Paper - its recommendations and the research findings – shows how the clinician and healthcare professional can engage effectively with AI in an uncertain world for the benefit of patients.

The MPS Foundation believes that the content of this paper will contribute significantly to the global conversation on healthcare AI technologies.

Professor Gozie Offiah
BSc, MB BCH BAO, LRCS/PI, MD, MMedSc, PhD
Chair of The MPS Foundation

Adrian Jackson
Head of The MPS Foundation

FOREWORD

by the Shared CAIRE Team

In a world where Artificial Intelligence (AI) technologies can work faster than people and spot patterns in data that humans may miss, policymakers, hospital managers, and clinicians alike are excited by AI's potential to improve the delivery of healthcare. But frontline clinicians are also wary of the impact of these technologies on their decision-making, their licences, and their patients.

Research studies typically focus on the impact of medical AI on patient safety and risk, but there is still a critical gap of research and analysis on AI's impact on clinicians. This gap must be addressed, otherwise clinicians could be overburdened by the technology, or they could simply ignore it, meaning that the benefits of AI will not be realised. If AI tools do not work for clinicians, they will not deliver for patients.

The focus of the Shared CAIRE (Shared Care AI Role Evaluation) project, funded by the MPS Foundation, was to consider the impact of AI decision-support tools on clinicians. The project team was multidisciplinary, bringing together researchers with expertise in medicine, AI, human-computer interaction, law, ethics, and safety science.

The project evaluated different ways AI tools could be used by clinicians in context, balancing out evaluations which only assess the performance of medical AI in the research lab. We assessed several models of AI decision-support, ranging from AI tools which surface information, to those which proffer concrete recommendations, to those which interact directly with patients outside the consultation room, with the clinician later expected to approve or veto decisions made.

The aim was to identify where AI decision-support tools best support clinicians to serve their patients effectively and where they risk increasing clinician stress and workload.

The project elicited a rich set of findings. We were struck by how much consensus there was between the two stakeholder groups involved in the study (clinicians and patient actors). Both groups shared concerns about what happens when there is disagreement between an AI tool and a clinician. There was also universal agreement that, whatever happens, the clinician will be liable. This being so, preserving clinicians' autonomy is crucial.

This White Paper outlines seven concrete, specific recommendations, which are grounded in the insights gained from the Shared CAIRE project.

We are at a turning point in the deployment of medical AI, as it moves from the margins to the mainstream of healthcare delivery. To unlock and realise AI's potential for patients, it is essential that we implement and deploy AI technologies in ways that work for those using them - the clinicians. If we fail to do so, we risk exacerbating the very challenges AI is supposed to address: clinician burnout, inefficiency, and uneven patient experiences and outcomes. We hope that this White Paper will help to address that urgent call.

The Shared CAIRE Project Team

Professor Tom Lawton, MBE
Principal Investigator, Clinical Lead, AI Lead

Dr Zoe Porter
Principal Investigator, Ethics Lead

Professor Ibrahim Habli
Principal Investigator, Safety Lead

Dr Vishal Sharma
Co-Investigator, Qualitative and Quantitative Analysis Lead

Professor Phillip Morgan
Co-Investigator, Law Lead

Dr Jo Iacovides
Co-Investigator, Human-Computer Interaction Lead

Dr Alice Cunningham
Co-Investigator, Research Fellow, Qualitative and Quantitative Analysis

Dr Yan Jia
Co-Investigator, Explainable AI Lead

Dr Yasmeen Hayat
Clinical Fellow

Dr Nee Wong
Clinical Fellow

Dr Muhammad Hussain
Research Associate, Human-Computer Interaction

Dr Nathan Hughes
Research Associate, Human Computer Interaction

Dr Shireen Gould
Clinical Fellow

TABLE OF CONTENTS

BACKGROUND

Artificial Intelligence (AI), which enables computers to perform cognitive tasks traditionally performed only by humans, has been used in healthcare for several decades.

Since the 1980s and 1990s, expert systems have been used in clinical decision-support for tasks such as diagnostics, treatment recommendations, and assessing the risk of events like heart attacks, strokes, and fractures.[1] These systems rely on "if-then" rules derived from human expertise. Since the start of the 21st century, advances in machine learning have led to a revolution in medical AI. Machine learning (ML) enables models to be trained on data. Advances in ML have led to systems that can detect diseases such as diabetic retinopathy[2] and skin cancer[3] at the level of human experts, recommend optimal treatment strategies,[4] and even have natural language interactions with patients.[5] With the recent emergence of Generative AI, which creates new content such as images and text, new frontiers are being explored. This includes systems which can generate synthetic medical images, such as chest X-rays,[6] to augment limited datasets and reduce bias,[7] and systems that can perform clinical summarisation tasks. The next transformative shift is the development of medical AI agents, which do not wait for human requests or prompts, but proactively monitor healthcare systems, track patient histories, retrieve data, identify issues, and suggest solutions.[8]

[1] https://qrisk.org/ ; https://frax.shef.ac.uk/FRAX/tool.aspx?country=9

[2] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. and Kim, R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), pp.2402-2410

[3] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), pp.115-118.

[4] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C. and Faisal, A.A., 2018. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 24(11), pp.1716-1720.

[5] de Pennington, N., Mole, G., Lim, E., Milne-Ives, M., Normando, E., Xue, K. and Meinert, E., 2021. Safety and acceptability of a natural language artificial intelligence assistant to deliver clinical follow-up to cataract surgery patients: proposal. JMIR Research Protocols, 10(7), p.e27227.

[6] Stanford MIMI, 2024. RoentGen: Radiology Report Generation. Available at: https://stanfordmimi.github.io/RoentGen/

[7] Koohi-Moghadam, M. and Bae, K.T., 2023. Generative AI in medical imaging: applications, challenges, and ethics. Journal of Medical Systems, 47(1), p.94.

[8] Zou, J. and Topol, E.J., 2025. The rise of agentic AI teammates in medicine. *The Lancet,* 405(10477), p.457.

The healthcare sector is one of the biggest areas of AI investment globally.[9] AI is at the heart of many nations' public policies for more efficient, more responsive healthcare systems. A recent study of healthcare AI maturity levels amongst member nations of the Organization for Economic Co-operation and Development (OECD) has shown that the UK and the US are the furthest ahead. These two nations are working towards specific goals for a collaborative and integrated healthcare AI ecosystem. Australia, Denmark, Finland, France, Germany, and Korea are at the next level of maturity, developing national healthcare AI policies and starting to set out best practice guidelines.[10]

In the UK, against a backdrop in which the Government has recently outlined strategies to 'turbocharge AI',[11] the National Health Service (NHS) Long Term plan in 2019 recommends the use of "decision-support and artificial intelligence (AI) to help clinicians in applying best practice, eliminate unwarranted variation across the whole pathway of care, and support patients in managing their health and condition."[12]

NHS England,[13] regulatory agencies[14] and international agencies such as the World Health Organisation[15] endorse the use of AI decision-support systems, rather than AI systems which replace a human clinician. The category of 'AI decision-support system' is quite broad. Decision-support systems range from AI tools that simply present relevant information to clinicians, to those that provide risk scores for clinician interpretation, to systems that make direct recommendations to the clinician about what treatment a patient should receive. Some perspectives categorise these systems across three levels of autonomy: Level 1, Level 2, and Level 3.[16]

[9] Meinhardt, C., Youssef, A., Thompson, R., Zhang, D., Kosoglu, R., Patel, K., & Langlotz, C., 2024. Pathways to governing AI technologies in healthcare. Stanford HAI. Available at: https://hai.stanford.edu/news/pathways-governing-ai-technologies-healthcare

[10] Castonguay, A., Wagner, G., Motulsky, A. and Paré, G., 2024. AI maturity in health care: An overview of 10 OECD countries. *Health Policy*, 140, p.104938

[11] Department for Science, Innovation and Technology, 2025. AI Opportunities Action Plan. [online] Available at: https://www.gov.uk/government/publications/ai-opportunities-action-plan

[12] NHS England, 2019. The NHS Long Term Plan. Available at: https://www.longtermplan.nhs.uk

[13] NHS England, 2022. Information governance guidance: artificial intelligence [online]. NHS England - Transformation Directorate. Available at: https://www.england.nhs.uk

[14] UK Government, 2023. Medical device stand-alone software (including apps), including IVDMDs. Available at: https://assets.publishing.service.gov.uk/media/64a7d22d7a4c230013bba33c/Medical_device_stand-alone_software_including_apps__including_IVDMDs_.pdf

[15] World Health Organization (WHO), 2021. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization. Available at: https://www.who.int/publications/i/item/9789240029200

[16] Bitterman, D.S., Aerts, H.J. and Mak, R.H., 2020. Approaching autonomy in medical artificial intelligence. *The Lancet Digital Health*, 2(9), pp.e447-e449.

Because there is now more data than can reasonably be considered by a responsible human healthcare practitioner, the most successful AI technologies so far draw human attention to important data, whether highlighting suspicious areas on a chest X-ray or summarising key points from a meeting. But there has also been a consistent desire to use AI to replace another key element of healthcare - that of making decisions or recommendations for treatment. Despite much-celebrated successes on paper and in the lab,[17, 18] these have not yet fared well in the real world.[19]

The specific impact of recommendation systems, as a subset of AI decision-support systems, remains largely unexplored. Their failure will contribute to a more general mistrust in AI amongst patients and practitioners. AI companies that respond by simply inserting a human clinician to act as a final checkpoint at the end of the chain risk creating a generation of “liability sinks.”[20]We use the term "liability sink" as an analogy to a "heat sink" in engineering, which absorbs heat from a component. Similarly, in this context, the clinician faces the risk of being used to absorb liability, drawing it away from other parties who also contribute to adverse consequences from the AI system, through their upstream decision-making beyond the clinician’s control.

This White Paper proposes recommendations which address the impact of decision-support tools on clinicians. The greatest threat to AI uptake in healthcare is the “off” switch, if frontline clinicians refuse to engage with technology they see as burdensome or unfit for purpose. Given competing priorities for funding, political pressures and the need for good governance, it is more vital than ever to focus resources on AI solutions which will generate the most benefit. It is by understanding how AI can genuinely support clinicians that this benefit will most likely be achieved.

[17] GP Online, 2024. Babylon's AI 'outperforms average doctor' in MRCGP exam. Available at: https://www.gponline.com
[18] Razzaki, S., Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D., Sangar, D., Taliercio, M., Butt, M., Majeed, A. and DoRosario, A., 2018. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. arXiv preprint arXiv:1806.10698.
[19] The Independent, 6 March 2021. Regulator has concerns over symptom checker app. Available at: https://www.independent.co.uk/news/health/nhs-symptom-checker-app-safety-complaints-b1813142.html
[20] Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N., Iacovides, I., Jia, Y., Sharma, V. and Habli, I., 2024. Clinicians risk becoming ‘liability sinks’ for artificial intelligence. Future Healthcare Journal, 11(1).

EXECUTIVE SUMMARY

With AI at the heart of many nations' healthcare policies, understanding its potential and risks is critical. To translate this understanding into meaningful policy and practice, it is time to move beyond an awareness of the general issues AI raises in healthcare toward much more targeted evaluations of its impact.

The recommendations of this White Paper are based on the findings of the Shared CAIRE (Shared Care AI Role Evaluation) project, whose principal focus was the impact of AI decision-support tools on the clinicians who use them. The research was conducted in the UK, with NHS clinicians working within NHS hospitals. Our findings appear to be in agreement with research into the impact of AI on clinicians being carried out elsewhere in the world.[21] It should be noted, however, that this global applicability has not been specifically tested.

The multidisciplinary Shared CAIRE project focused specifically on the interaction between a clinician and an AI decision-support tool during a patient consultation.

The standard or most prevalent model of AI decision-support is that an AI tool uses electronic data to calculate a course of action or treatment for the patient, and then makes a recommendation to the clinician, who then decides whether or not to act on it, after dialogue with the patient. The prevalence of this model can lead to two things being overlooked. First, other types and models of AI decision-support are possible. Systems which make recommendations do not have to be the default AI decision-support tool. Second, the prevalent model may not be the best one for those using the technology (nor for patients who have their own beliefs, values and preferences that need to be taken into account). The Shared CAIRE project assessed the impact of the prevalent model on clinicians, and the impact of some of its alternatives. We evaluated how different models affected clinical decision-making and whether some models of using AI tools to support decision-making were particularly prone to generating clinician stress, confusion, worry and overload - thereby making the "off switch" the most attractive option.

[21] McCradden MD, Thai K, Assadi A, et al. What makes a 'good' decision with artificial intelligence? A grounded theory study in paediatric care BMJ Evidence-Based Medicine doi: 10.1136/bmjebm-2024-112919

The project proceeded as follows

We developed six models of AI decision-support, including the most prevalent model. The six models involved varying degrees of information and direction from the AI tool. The models are described in Chapter 3, with diagrams presented in the Appendix.

We devised three different scenarios for each of two different specialties: diabetes and obstetrics. These were scenarios in which a specific clinical decision had to be made (namely, whether to move a patient to insulin, and whether to advise a patient to have a C-section). The scenarios were refined in consultation with the Patient/Public Involvement/Engagement (PPIE) panel and clinical experts.

We created a high-fidelity prototype of an AI decision-support tool, in consultation with working clinicians. "Wizard of Oz" prototypes of the tool (prototypes which appear to be, but are not in fact, 'real' AI), based on each of the six models, were developed to simulate the functionality of the AI, without requiring full system implementation.

Real clinicians, with enough years of experience to be making independent decisions in their specialty, then took part in simulated (i.e., mock) consultations of three scenarios from their specialty. The three mock consultations were also based on a randomised selection of three of the six models of AI decision-support. Actors played the role of patients.

Immediately after the simulated consultations, we interviewed clinicians to understand the impact of the different models of AI decision-support on their decision-making, and how having an AI tool influenced their behaviour in the consultation.

Post-consultation surveys were also given to actor patients. These surveys looked at their experiences in the consultation, whether they had been made aware of the use of an AI tool, and who they thought would be responsible if a patient came to harm.

To assess the legal implications on clinicians of the six models of AI decision-support, we asked another set of clinicians to review the transcripts of the simulated consultations and complete a questionnaire. The aim was to see whether the clinician participants' decision-making during the mock consultations was deemed appropriate by an external group of peers, similar to "a responsible body of medical opinion," as referred to in the `Bolam test' of medical negligence.[22] This part of the study provides some initial indication of how responsible bodies of medical opinion might approach the AI/clinician interface.

The qualitative data, from interviews and questionnaires, were then reviewed and analysed using Thematic Analysis. Quantitative data collected from patient actors and external peer reviews were analysed for descriptive statistics and outputs.

From the results of the study (which are discussed in Chapter 3), we propose the following recommendations. To be clear, these recommendations are based on current and near-future AI decision-support tools and the real-world context in which clinicians are actually working. These recommendations are not based on the assumption of an ideal future in which AI technologies are infallible or fair systems of liability attribution for AI have already been established.

[22] Bolam v Friern Hospital Management Committee [1957] 1 WLR 582 (QB)

Recommendation 1:

AI tools should provide clinicians with *information*, not *recommendations*

AI decision-support systems in healthcare should provide information to support the clinician making a decision about a patient's treatment or care, but not a recommendation on treatment or care. While the legal weight of an AI recommendation remains unclear, Recommendation 1 will help to reduce the chance of clinicians being pushed into accepting recommendations they deem suboptimal (but defensible). At the same time, AI tools can surface salient information that saves the clinician valuable time, enabling them to engage more closely with the patient. As such, we recommend the use of medical AI tools which provide information, but not recommendations, to clinicians (until product liability has been revised).

Recommendation 2:

Revise product liability for AI tools before allowing them to make recommendations

Only after a revision of product liability for AI tools in healthcare should AI-based recommender systems be deployed. Our view is not that recommender systems will never have value, but that the main impediment to realising their value is that there are significant difficulties in applying the current product liability regime to an AI tool. Without reforms to the product liability regime to accommodate AI systems, there is a risk that clinicians will act as a 'liability sink' (i.e., absorb all of the liability) even where the system is a major cause of the wrong. Until product liability for AI tools has been revised, healthcare systems and organisations should limit deployment to AI tools that provide, distill, and present information to clinicians rather than deploy AI tools that make direct recommendations.

Recommendation 3:

AI companies should provide clinicians with the training and information required to make them comfortable accepting responsibility for an AI tool's use

Clinicians are responsible for delivering the best possible care to their patients. To assess whether the information provided by an AI tool, such as risk scores and statistics, or, in due course, concrete AI recommendations, are appropriate for the patient in front of them, clinicians need training and information from the companies that develop AI tools. In particular, clinicians need to understand the intended purpose of the system, the contexts it was designed and validated to perform in, the scope and limitations of its training dataset (with reference to bias), and its decision thresholds.

Recommendation 4:

AI tools should not be considered akin to senior colleagues in clinician-machine teams

An AI decision-support system should not be regarded as a 'senior colleague' in a human-machine team. Clinicians should not always be expected to agree with or defer to an AI output, whether that output is a direct recommendation, a classification, or an analysis of the data. It should be made explicit in new healthcare AI policy guidance and in guidance from healthcare organisations how clinicians should approach conflicts of opinion with the AI. This is particularly important given that AI outputs are unlikely to be obviously wrong, so in cases where there is a mismatch between the clinician's judgement and the AI's output, clinicians are more likely to be disagreeing with an output that they deem to be suboptimal, but defensible or plausible.

Recommendation 5:

Disclosure should be a matter of well-informed discretion

Presently, clinicians use many tools which are not AI-based, and often do not disclose this to the patient (although they might, depending on the patient and the clinical decision being made). During the simulations, the clinicians' approach was broadly similar. Given that the clinician is responsible for patient care, and that disagreement with an AI tool could end up worrying the patient, it should be at the clinician's discretion, depending on context, whether to disclose to the patient that their decision has been informed by an AI tool. However, regulatory officials and healthcare organisations should provide clinicians with guidance on the exercise of this discretion, and provide patients with information on their rights.

Recommendation 6:

AI tools that work *for* users need to be designed *with* users

The need for co-design is widely agreed across the Responsible Research and Innovation landscape, and this is important to emphasise and reaffirm. In healthcare contexts, which are safety-critical and fast-moving, engaging clinicians in the design of all aspects of an AI tool – from the interface, to the balance of information provided, to the details of its implementation – can help to ensure that these technologies deliver more benefits than burdens.

Recommendation 7:

AI tools need to provide an appropriate balance of information to clinician users

How much information a clinician receives from an AI tool matters. Too much information, and the time it takes to review detracts from paying attention to the patient. Too little, and clinicians do not trust the machine. One way of discovering the `sweet spot' between too much and too little information is to involve clinicians in the design and development of AI decision-support tools.

CHAPTER 1

DETAILED RECOMMENDATIONS

This White Paper offers seven concrete recommendations on the use of AI decision-support tools based on the results of the Shared CAIRE study.

The recommendations are bold and specific, yet nuanced. The Shared CAIRE research highlights the need for a thoughtful approach to integrating AI decision-support tools into real-world clinical settings – ensuring they genuinely support the clinicians using them while preserving the important human touch in patient care. The detailed descriptions below provide the justification for each of the seven recommendations. These are supported with quotations from clinicians, patient actors, external clinical peer reviewers, and Patient-Public Involvement and Engagement (PPIE) Panel members.

PPIE Panel members and the patient actors are entirely anonymised. Clinician participants are identified by number (e.g., D01 or C01, where D and C refer to their specialty, i.e., D to Diabetes and C to Caesarean). Where feasible, the quotations are labelled according to the specific scenario and model of AI decision-support to which they refer. The key is in Tables 1 and 2 below. Further details on these scenarios and models are available in Chapter 3.

Scenario D1 – Diabetes Scenario 1	Scenario C1 – Caesarean Scenario 1
Scenario D2 – Diabetes Scenario 2	Scenario C2 – Caesarean Scenario 2
Scenario D3 – Diabetes Scenario 3	Scenario C3 – Caesarean Scenario 3

Table 1: Key for the labelling of scenarios

M1 – Model 1: Traditional (no AI) Model
M2 – Model 2: Prevalent AI Model
M3 – Model 3: No Recommendation Model
M4 – Model 4: Recommendation with Information Model
M5 – Model 5: Conversational AI Efficiency Model
M6 – Model 6: Conversational AI Quality Model

Table 2: Key for the labelling of models of AI decision-support

Recommendation 1

AI tools should provide clinicians with *information*, not *recommendations*

AI decision-support systems in healthcare should provide information to support the clinician making a decision about a patient's treatment or care, but not a recommendation on treatment or care. While the legal weight of an AI recommendation remains unclear, Recommendation 1 will help to reduce the chance of clinicians being pushed into accepting recommendations they deem suboptimal (but defensible). At the same time, AI tools can surface salient information that saves the clinician valuable time, enabling them to engage more closely with the patient. As such, we recommend the use of medical AI tools which provide information, but not recommendations, to clinicians (until product liability has been revised).

Of the six models tested, clinicians generally preferred Model 3, which was the model where the AI tool highlighted relevant information from the electronic data (e.g., risk scores for uterine rupture), but did not give the clinician a direct recommendation (e.g., "recommend a caesarean section").

"I much preferred the output that was just stats to use rather than a recommendation."
(Participant C01, Model 3)

"The one where it had sort of the chat with the patient was too wordy [Model 6] and didn't help, when it had concise bullet point sort of information, that was probably the most useful [Model 3]."
(Participant C05, Model 3)

"I think the last one [Model 3] was probably the closest you can get to my ideal."
(Participant D09B, Model 3)

There are several reasons for this preference.

For one thing, clinicians cannot be sure that an AI recommendation adequately incorporates patient preferences and values:

“It didn’t account for her mental health and her trauma from the last delivery which I think is an important reason to take on board the decision of mode of delivery.”
(Participant C04, Model 4)

“It [AI] hasn’t made the recommendation taking into account patient preference. I suppose [how did that feel?] -- impersonal.”
(Participant C05)

“It should be kind of a decision made for the patient, taking into account his medical, psychosocial, and other things as well. It’s not just about a medical decision, I don’t think the medical decision can be taken without the psychosocial component of it... I think it will be very difficult to input the data about the psychosocial impact into AI, and the level of the patient is anxious, okay, you can put that information, but how anxious is the patient. Like why are they anxious?”
(Participant D01)

Clinicians also do not like being given a black and white answer to complex clinical scenarios.

“I felt quite uncomfortable with the AI rec... when it gave you a recommendation, particularly the one that wasn’t backed up with anything because it’s much more of a grey area... like it made it very black and white when it’s not a black and white decision.”
(Participant C01, Scenario C2, Model 2)

“The scenario was so much more complicated than what the AI seems to think it was, that you worry about that it’s not picked up on the context of it. I think it’s fair enough to have a kind of straightforward answer, if the scenario’s quite straightforward, but if it’s not then you worry it’s missed something.”
(Participant D03, Scenario D1, Model 4)

“I think the last one was the most difficult, Mr Smith, because it gave you a very black and white answer to a question that’s not that simple, without any of the other context of the way it made the decision. So, he wanted to definitely know if there were other options, but the decision support tool just said "no this is the recommendation.”
(Participant D03, Scenario D1, Model 4)

The clinician participants also expressed a desire to understand and investigate the reasoning behind the AI's recommendations to ensure it aligns with their professional judgment. As a result, rather than alleviating their workload, AI-based recommender systems may inadvertently increase clinicians' cognitive load:

"... there's always a question of where are the numbers from? And how are they derived?"
(Participant D02)

"A bit surprising [the recommendation], and I think in real life I would have taken a bit of extra time to run through some guidelines, and written support."
(Participant D11)

More positively, an AI tool which *just* surfaces information (of the right kind, at the right level of detail - see Recommendation 7 below) can help clinicians to cut through the vast quantities of data they typically have to sift in a time-limited clinical consultation:

"I think if I'm doing the consultation, having the stats in front of me was great for that situation."
(Participant C06, Model 3)

"I've not got a great memory for stats...so that probably would speed up my clinic quite a lot, not having to check guidelines, not having to put out notes."
(Participant C01)

"I think anything that improves efficiency it's great, and if... you have a decision support tool it can help you go through the clinic more easily."
(Participant C08)

"I thought having the actual numbers in terms of risks was really helpful, because they're risks that I know about, but I don't have the numbers on the top of my head, so actually having them there is really helpful and you can show them to the patient and talk them through what it means."
(Participant D03)

Concerns around liability also provide a strong justification for Recommendation 1. Liability for patient harm caused by implementing an AI recommendation (e.g., "recommend a caesarean section") presently rests with the clinician. Clinicians are ultimately responsible for the outcome, regardless of whether they choose to follow or reject an AI tool's recommendation. If a clinician rejects an AI recommendation (and something subsequently goes wrong), the burden of justification on why they rejected the recommendation will fall on them, as it will be an obvious line of questioning for cross-examination in a legal forum such as a court. The benefit of providing clinicians with AI recommendations therefore seems limited:

"Am I going to put my GMC registration on the line from the statistics from this model? No. So in a way that one that I trust most is probably the one with least information."
(Participant D02, Model 2)

"They'd think that they've been given the wrong information. Because you're not going to blame the AI are you, cos that's a thing. You'll be blaming whoever's presented that to you, that you have access to it."
(PPIE Panel Member 1)

"I think because the information is there, you know, she's got a physical up-down scar, it says that it was born at 26 weeks, that is part of our fact-finding information, it's well established that that's a contraindication to vaginal birth. If you stuck that before a panel of experts, they'd say oh, she made the wrong decision, she was totally off put by that output. So yeah, it would be my fault."
(Participant C01)

Indeed, an AI recommendation may sometimes be actively unhelpful. For example, it created difficulties for clinician participants in borderline cases, where the `right' course of action for a patient is unclear, and in cases of disagreement with the AI:

"I don't see often, for example the second case where you have this pregnant lady with abnormal sugar levels, so I think that was a bit difficult, especially when it recommends treatment that I don't quite agree with."
(Participant D06, Scenario D2, Model 4)

"A bit bewildered, I looked at the decision, and I thought, oh, I don't agree with that. And then it left me feeling a lack of confidence in the consultation, subsequent consultation, I just felt lots of uncertainty."
(Participant D10, Scenario D1, Model 6)

Amongst other issues here, an "anchoring effect" might arise which clinicians would have to resist.[23] The 'anchoring effect' is the cognitive bias of relying too heavily on the first piece of information offered:[24]

"The AI kind of reinforces that potential incorrect decision."
(Participant C02)

AI tools can surface salient information that saves the clinician valuable time, enabling them to engage more closely with the patient. As such, we recommend the use of AI decision-support tools which provide information, but not recommendations, to clinicians. The rationale for Recommendation 1 stems from both clinician preferences and concerns about the legal implications of deploying AI tools that provide direct recommendations (until product liability has been revised). The aim is to protect clinicians from being pushed into accepting recommendations that in their judgement are suboptimal. This is discussed further in Recommendation 2 below.

[23] Gaube, S. et al. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digital Medicine 4,(31),289
[24] Sherif, M., Taub, D., & Hovland, C. I., 1958. Assimilation and contrast effects of anchoring stimuli on judgments. Journal of Experimental Psychology, 55(2), pp. 150–155.

Recommendation 2

Revise product liability for AI tools before allowing them to make recommendations

Only after a revision of product liability for AI tools in healthcare should AI-based recommender systems be deployed. Our view is not that recommender systems will never have value, but that the main impediment to realising their value is that there are significant difficulties in applying the current product liability regime to an AI tool. Without reforms to the product liability regime to accommodate AI systems, there is a risk that clinicians will act as a 'liability sink' (i.e., absorb all of the liability) even where the system is a major cause of the wrong. Until product liability for AI tools has been revised, healthcare systems and organisations should limit deployment to AI tools that provide, distill, and present information to clinicians rather than deploy AI tools that make direct recommendations.

A recent survey by the Alan Turing Institute found that only 30% of surveyed clinicians had a clear understanding of who was responsible if an erroneous decision was made using an AI system.[25] By contrast, the Shared CAIRE study revealed clear and widespread agreement amongst clinicians and patient representatives that clinicians would be liable for patient outcomes.

When asked who would be liable, clinicians said: "Me" (Participant D07), "Me" (Participant D09), "I would be" (Participant D03), "Me" (Participant D11). This view was widely felt by clinicians:

"Myself as the clinician because it would ultimately be my decision."
(Participant C04)

"Well, the doctor would be liable, [...] I mean even NICE guidelines and everything, these are purely recommendations, they're not things which, you know, have to be followed. So, I think yeah, that they're all recommendations, and it's the doctor's decision to follow, so I think if anything was done wrong or missed, it would always fall back on the doctor."
(Participant D04)

[25] Hashem, Y., Esnaashari, S., Morgan, D., Francis, J., Poletaev, A., Enock, F., Bright, J., 2024. One in four UK doctors are using Artificial Intelligence: Exploring doctors' perspectives on AI after the emergence of large language models . The Alan Turing Institute.

“It would be me. Because I’m relying on this, I need to still exercise my clinical judgement, I am wholly liable.”
(Participant D08)

“I know we’re trying to step away from a blame culture, but ultimately when you’re stood in court, it’s still someone who needs to take responsibility, I think it’s still very much the clinicians who will have to take the responsibility.”
(Participant D02)

“I think it is the clinician, the person who makes the decision, because it’s a tool. We have many tools already and you can’t blame a tool for the decision you make.”
(Participant D13)

“*You [the clinician] can't blame anyone. If it's recorded, all the information the clinician's given and the AI's decision, they wouldn't have much grounds would they?”*
(PPIE Panel Member 5)

“If it’s not going to give you the right information, why was it even there for me to use? Because you’re always going to be looking to blame a person in the end, not a thing.”
(PPIE Panel Member 1)

This perception that the clinicians are liable is correct. Under current liability regimes, clinicians are likely to be liable for harms which materialise.[26] AI-based recommender systems put them in a tough position: they are liable for outcomes despite having no control over how the AI reaches its conclusions. Clinicians are in a particularly difficult situation if they reject the AI output and something were subsequently to go wrong. But even if they were to follow the AI recommendation and a negative outcome resulted, so that the ‘fault’ is in some way shared between the system and the clinician, then, given the current problems with applying the product liability regime in the United Kingdom to AI, the clinician is likely to be a “liability sink” (i.e., absorb all of the liability for the outcome), with little prospect of the designers or producers of the system being challenged, or sharing liability.

[26] Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N., Iacovides, I., Jia, Y., Sharma, V. and Habli, I., 2024. Clinicians risk becoming ‘liability sinks’ for artificial intelligence. Future Healthcare Journal, 11(1).

The inadequacy of the 1985 European Product Liability Directive[27] (implemented in the UK through the Consumer Protection Act 1987) in dealing with AI is well acknowledged. Reforms to the directive at the European level have been made to deal with some of these acknowledged problems,[28] but post-Brexit these reforms will not apply to the UK regime. Without reform, clinicians, or those harmed by decisions made which resulted from poor AI outputs being provided to clinicians, will have little prospect at pointing the finger at those who design and provide AI systems.

An alternative distribution of liability might lead to better outcomes. The sharing of responsibility between clinicians and those responsible for the system's design, development, production, and deployment might facilitate better results. This will ensure pressure is placed on other actors to ensure the system is well-designed and works well. These other actors may be in a better place to effect systemic change than individual clinicians (or their employers). We can term this 'Clinician responsibility plus'. Such a reform of the product liability regime for medical AI may help to generate greater clinician confidence in the AI, as well as greater confidence amongst leadership teams in healthcare organisations. Further, it may be appropriate to reconsider standards of care in the light of AI if its use materially affects best clinical practice.

We should be clear that we are not suggesting that clinicians should be exempt from liability when AI is involved. Rather, for clinicians to take meaningful responsibility, they should be properly supported with the right information and understanding. In the future, if AI systems take on greater decision-making roles, product liability laws must be reformed to align liability with control. In particular, shared liability between clinicians and those responsible for a system's design, development, production and deployment should be considered. Until there has been a revision of product liability for AI tools, healthcare systems and organisations should only deploy AI tools that provide, distil and present information (rather than recommendation) to clinicians.

[27] Council of the European Communities (1985) Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. Official Journal L 210, 07/08/1985, pp. 29–33. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31985L0374

[28] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC. Official Journal L 2024/2853, 18.11.2024 Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024L2853

Recommendation 3

AI companies should provide clinicians with the training and information required to make them comfortable accepting responsibility for an AI tool's use

Clinicians are responsible for delivering the best possible care to their patients. To assess whether the information provided by an AI tool, such as risk scores and statistics, or, in due course, concrete AI recommendations, are appropriate for the patient in front of them, clinicians need training and information from the companies that develop AI tools. In particular, clinicians need to understand the intended purpose of the system, the contexts it was designed and validated to perform in, the scope and limitations of its training dataset (with reference to bias) and its decision thresholds.

The importance of training clinicians to use AI effectively has been emphasised by, in the UK, Health Education England[29] and the Information Commissioner's Office (ICO),[30] and globally, by the World Healthcare Organisation (WHO).[31] The Shared CAIRE study confirms the need for training, and goes deeper: this training is needed in order to support clinicians to deliver the best possible care to their patients, and feel comfortable taking responsibility for outcomes to patients when they have used an AI decision-support tool.

Clinicians need to understand both the scope and limitations of the AI tools they use. This is particularly important for clinicians who may be called to justify why they followed (or did not follow) an AI's recommendation:

"I think we have to justify for our own actions, especially if they [AI] recommend anything, they [AI] should say, okay, I'm recommending this because the NICE guidelines says this, this and this, and the patient has this, and this, and this. So, we can agree with that or disagree with that, just a piece of information, I don't think I can completely rely on that."
(Participant D01)

[29] Topol, E.J., 2019. Preparing the healthcare workforce to deliver the digital future. Health Education England. Available at: https://topol.hee.nhs.uk/
[30] Information Commissioner's Office (ICO). (2022). How to use AI and personal data. Available at: https://ico.org.uk/media/for-organisations/documents/4022261/how-to-use-ai-and-personal-data.pdf
[31] World Health Organization. 2021. Ethics and governance of artificial intelligence for health. Geneva: World Health Organization. Available at: https://www.who.int/publications/i/item/9789240029200

But even for the information-only systems which this White Paper recommends (until product liability has been revised), clinicians will need a greater understanding of an AI tool's intended purpose, the contexts it was validated for, the nature and date of its training dataset and the data and decision thresholds underlying any risk scores it provides:

"I think it was another resource to sort of look at, I think having numbers is always helpful in boosting confidence, but there's always a question of where are the numbers from? And how are they derived?"
(Participant D02)

"We are now at the age where no-one can remember everything all the time, so yeah that's useful. A bit of almost like a, so even for things that you do know it helps, competence, reinforcement (...) it can turn into false confidence, and false reinforcement, if the sort of knowledge portion of it is not updated on a regular basis. And then therein comes the medical legal issues."
(Participant D09)

For clinicians to have the confidence to take responsibility for AI-assisted decisions, they need to have the skills and knowledge to know when the information provided by the AI can be relied on and when it is relevant to the patient in front of them:

"You'd want to interrogate that a bit more, and just make sure if new guidelines, or new evidence comes up, is that being updated? And you'd hope to maybe see a reference on there somewhere, where that statistic is pulled from? I suppose I had no reason to doubt it, because when I looked at them, you know, it was all numbers that I recognised."
(Participant C07)

"There are just certain roadblocks, certain points in the road where if someone does try to dissolve themselves of responsibility for instance, that we would have to have a serious reconsideration, people talk about their rights with AI don't they, which is just something that I don't understand how we could reach that point, and that I think that as long as they remain as support, and not autonomous, that they are helpful, but that the humans using them need training as well."
(Participant D11)

Like other tools,[32] AI algorithms trained on one ethnic group may not be applicable more widely, and simple thresholds may not account for patient values.[33] Clinicians therefore need a good understanding of a variety of aspects of any AI system before they can safely make use of it.

AI companies are the parties who should provide this information. They should also be prepared to give clinicians training on how the tools they are using work and the procedure by which they reach outputs. Healthcare organisations should ensure that clinicians receive this training from AI companies, and that they have protected time to learn about the AI tools they are being asked to use, so that they can do so in a way that is appropriate for individual patients.

This increased AI literacy will help to empower clinicians and enable them to navigate their AI tool use more skillfully. It will help them to know when confidence in the AI would be justified, and when the AI tool can support the best possible patient care, and hence support more effective human-AI teaming.

[32] Valbuena, V.S., Seelye, S., Sjoding, M.W., Valley, T.S., Dickson, R.P., Gay, S.E., Claar, D., Prescott, H.C. and Iwashyna, T.J., 2022. Racial bias and reproducibility in pulse oximetry among medical and surgical inpatients in general care in the Veterans Health Administration 2013-19: multicenter, retrospective cohort study. BMJ, 378.

[33] Birch, J., Creel, K. A., Jha, A. K., & Plutynski, A., 202). Clinical decisions using AI must consider patient values. Nature Medicine, 28(2), 229–232

Recommendation 4

AI tools should not be considered akin to senior colleagues in clinician-machine teams

An AI decision-support system should not be regarded as a "senior colleague" in a human-machine team. Clinicians should not always be expected to agree with or defer to an AI output, whether that output is a direct recommendation, a classification, or an analysis of the data. It should be made explicit in new healthcare AI policy guidance and in guidance from healthcare organisations how clinicians should approach conflicts of opinion with the AI. This is particularly important given that AI outputs are unlikely to be obviously wrong, so in cases where there is a mismatch between the clinician's judgement and the AI's output, clinicians are more likely to be disagreeing with an output that they deem to be suboptimal, but defensible or plausible.

With the growing capabilities of AI technologies, the concept of 'human-AI teaming' has emerged as a key paradigm in human-AI interaction.[34] This term reflects the increasing integration of humans and AI tools as collaborative partners, but it also raises the risk of anthropomorphising AI. To clarify, with Recommendation 4 we are not advocating for AI decision-support systems to be thought of as real, human colleagues. Rather, we seek to emphasise that AI tools should not be seen as more authoritative than human clinicians and hence not implicitly regarded as senior colleagues.

For some of the clinicians participating in the Shared CAIRE study, the AI tool provided similar reassurance to speaking to a colleague:

"I felt it helped me be more confident with making my recommendation. If I didn't have that tool, and I didn't know what to do, I probably would have picked up the phone and speak to a diabetic colleague or looked up what NICE guidance was."
(Participant D02)

"It's a bit like perhaps having a more junior doctor there, you know, you can see that there can be flaws, but that ultimately there are some really helpful elements to it as well."
(Participant C07)

[34] Lyons, J.B., Sycara, K., Lewis, M. and Capiola, A., 2021. Human–autonomy teaming: Definitions, debates, and directions. Frontiers in Psychology, 12, p.589585.

However, clinicians also raised the worry that non-medical professionals might regard the tools as clinical experts:

"I think that I can see, I can see this playing a big part in helping non-medical professionals, within healthcare. I think this would be a very helpful tool because it does kind of replace a lot of the thinking, that's needed, to make clinical decisions. And I can see how this could be utilised in that field, in that way. I'm not saying it's a good thing, but I think that's, I can see how this could be used."
(Participant D08)

Moreover, for experienced and trained clinicians, the sense of being helped by the AI tool dissolved when they disagreed with an AI's output. A recent survey from the Alan Turing Institute, based on a sample of 929 doctors, found that doctors generally felt that they would be confident to ignore the recommendations of an AI tool that they disagreed with (54%), with this number rising to 58% in those actively using AI.[35] But our research, based on simulations of consultations involving an AI tool, revealed much more disturbance and confusion in such cases:

"It made me second guess myself quite a lot."
(Participant C01, Scenario C2, Model 2)

"I felt quite uncomfortable, because I thought I was missing something (...) I think I didn't feel that comfortable and I was a bit doubtful."
(Participant D06, Scenario D2, Model 4)

"A bit bewildered, I looked at the decision, and I thought, oh, I don't agree with that. And then it left me feeling a lack of confidence in the consultation, subsequent consultation, I just felt lots of uncertainty."
(Participant D10, Scenario D1, Model 6)

"[It] does make you feel a bit guilty for thinking perhaps I should know this better, and that there must be quite a high possibility that I'm really off the mark on this, and it's exposed me. But then fundamentally still disagreeing with it, so unpleasant."
(Participant D11, Scenario D3, Model 3)

[35] Hashem, Y., Esnaashari, S., Morgan, D., Francis, J., Poletaev, A., Enock, F., Bright, J., 2024. One in four UK doctors are using Artificial Intelligence: Exploring doctors' perspectives on AI after the emergence of large language models . The Alan Turing Institute.

This confusion was also felt by patient representatives:

"Our confidence has always been in people, in the clinicians (...) and suddenly this machine's coming in and saying something different. I'm going to lose confidence all round."
(PPIE Panel Member 2)

This suggests that guidance is needed to help clinicians understand how to approach disagreement with an AI output. More than that, we propose it is made clear that, in cases of disagreement, a clinician should not be expected to *defer* to an AI output. Studies have already shown that jurors consider it more reasonable for clinicians to accept even non-standard AI recommendations than to reject them.[36] A theme which arose from the Shared CAIRE study was the spectre of "algorithmic deference," whereby clinicians may end up following the AI[37, 38] – even when they are able to avoid traditional "automation bias" where they over-trust the system.[39, 40] A trend towards clinicians accepting or following AI outputs as a form of "defensive medicine" – where they do so to shield themselves from liability – would be a troubling consequence of introducing AI into clinical settings.[41, 42, 43, 44]

"If an algorithm is very supportive one way or the other, then maybe that stops people feeling able to challenge decisions."
(Participant C01)

"I did feel slightly under pressure to make a decision to start insulin...if it [AI] wasn't there...I probably would have continued more comfortably with life-style modification."
(Participant D04, Scenario D1, Model 6)

[36] Tobia, K., Nielsen, A. and Stremitzer, A., 2021. When does physician use of AI increase liability? Journal of Nuclear Medicine, 62(1), pp.17-21.
[37] Crootof, R., Kaminski, M.E., Price, W. and Nicholson, I.I., 2023. Humans in the Loop. Vanderbilt. Law Review, 76, pp. 429-509.
[38] Grote, T. and Berens, P., 2020. On the ethics of algorithmic decision-making in healthcare. Journal of Medical Ethics, 46(3), pp.205-211.
[39] Goddard, K., Roudsari, A. and Wyatt, J.C., 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), pp.121-127.
[40] Mittelstadt B.,, 2021. The Impact of Artificial Intelligence on the Doctor-Patient Relationship (Council of Europe 2021) 36. Available at: https://www.coe.int/en/web/human-rights-and-biomedicine/report-impact-of-ai-on-the-doctor-patient-relationship
[41] Banja, J.D., Hollstein, R.D. and Bruno, M.A., 2022. When artificial intelligence models surpass physician performance: medical malpractice liability in an era of advanced artificial intelligence. Journal of the American College of Radiology, 19(7), pp.816-820.
[42] Tobia, K., Nielsen, A. and Stremitzer, A., 2021. When does physician use of AI increase liability? Journal of Nuclear Medicine, 62(1), pp.17-21.
[43] Banja, J.D., Hollstein, R.D. and Bruno, M.A., 2022. When artificial intelligence models surpass physician performance: medical malpractice liability in an era of advanced artificial intelligence. Journal of the American College of Radiology, 19(7), pp.816-820.
[44] Kessler, D. and McClellan, M., 1996. Do doctors practice defensive medicine?. The Quarterly Journal of Economics, 111(2), pp.353-390.

“The AI kind of reinforces that potential incorrect decision.”
(Participant C02)

“Yeah, I don’t know why I felt like the decision was easier when I wasn’t given those, the kind of two ultimatums, I don’t know.”
(Participant D08, Model 1)

Guidance that discourages a culture of algorithmic deference can help to alleviate the foreseeable burden on clinicians to justify why they chose to override an AI’s output:

“I know we’re trying to step away from a blame culture, but ultimately when you’re stood in court, it’s still someone who needs to take responsibility, I think it’s still very much the clinicians who will have to take the responsibility.”
(Participant D02)

“A bit surprising [the recommendation] … it does make you feel a bit guilty for thinking perhaps I should know this better, and that there must be quite a high possibility that I’m really off the mark on this, and it’s exposed me. But then fundamentally still disagreeing with it, so unpleasant.”
(Participant D11)

“I think this is very, very important and I think we have to justify for our own actions, especially if they [AI] recommend anything, they [AI] should say, okay, I’m recommending this because the NICE guidelines says this, this and this, and the patient has this, and this, and this. So, we can agree with that or disagree with that, just a piece of information, I don’t think I can completely rely on that.”
(Participant D01)

Guidance for clinicians on how to approach disagreement with an AI tool is particularly important given that AI outputs are unlikely to be obviously wrong. Cases where a clinician disagrees with an AI’s output will most likely those where the output is suboptimal but defensible, rather than wildly incorrect. In addition, the “correct answer” will often not be clear-cut, since many clinical decisions are matters of judgement and context.

Clinicians should regard AI as an adjunct, a tool. They should not think of it as a replacement for – or improvement on – either their own clinical judgement or the judgement of a trusted human colleague. Clinicians should feel empowered to disagree with AI recommendations, particularly when the recommendation is suboptimal and does not align with their own clinical judgement. Policymakers and healthcare organisations should produce guidance for clinicians on how to approach conflicts of opinion with the AI. For further information, please read our published MPS Casebook titled “Disagreeing with AI could be bad for your health”.[45]

[45] MPS Casebook. (2024). Disagreeing with AI could be bad for your health. Casebook UK, Volume 32, Issue 1, September 2024. Available at: https://read.nxtbook.com/mps/casebook/casebook_uk_volume_32_issue_1/disagreeing_with_ai.html

Recommendation 5

Disclosure should be a matter of well-informed discretion

Presently, clinicians use many tools which are not AI-based, and often do not disclose this to the patient (although they might, depending on the patient and the clinical decision being made). During the simulations, the clinicians' approach was broadly similar. Given that the clinician is responsible for patient care, and that disagreement with an AI tool could end up worrying the patient, it should be at the clinician's discretion, depending on context, whether to *disclose* to the patient that their decision has been informed by an AI tool. However, regulatory officials and healthcare organisations should provide clinicians with guidance on the exercise of this discretion, and provide patients with information on their rights.

Given the emphasis on patient autonomy and informed consent in medical ethics, and the emphasis on transparency in AI ethics, Recommendation 5 needs careful explanation. This recommendation is not that a clinician's use of an AI-based decision-support tool should not be disclosed, but that universal disclosure should not be mandated for all cases and contexts. Rather, clinicians should have the authority to exercise their discretion and be guided by their own judgement on what is appropriate as part of a holistic conversation with a patient. Clinicians should be supported in the exercise of this discretion with guidance from appropriate regulators and healthcare organisations, and in line with the way existing non-AI tools and scoring systems are used.

Clinician participants in the simulations often found that disclosing the use of an AI tool was not practically easy. They felt it might disrupt the flow of the consultation and get in the way of the "human touch" being established between the doctor and the patient:

"I didn't mention it anyway (...) I think probably because I'm not used to it, so that was probably a part of it. But also perhaps because I think you know it's little bit like, well actually I'm your doctor and I would chat to you about it. Like the computer says this, you know I feel is not really good medical care."
(Participant C07, Scenario C3, Model 2)

“If something was churning out recommendations, I don’t know how comfortable patients would be with you saying well, the computer says that we should, it thinks we should do this. I don’t know societally if we’ve reached a point where people are happy to have their decisions made by a computer.”
(Participant C01)

“If I don’t give them [patients] the time in the consultation to talk about what they’re worried about, if they don’t feel heard, I’m less likely to build a rapport with them. So, whatever the AI model has done, I probably have to repeat, and in fact I might even have to pretend I haven’t read what the AI model has discussed.”
(Participant D02)

Disclosing the use of the AI tool is also incongruent with the current norm in real-world practice. Clinicians frequently rely on software tools to analyse data and support their decision-making without disclosing this to the patient. This practice is not generally perceived to endanger patient autonomy, so long as adequate truthful justification is given for clinical decisions and recommendations:

“I didn’t tell them (...) I felt like it just didn’t change what I did or said or thought.”
(Participant C04, Scenario C3, Model 4)

To exercise their autonomy and give informed consent, patients need to understand the clinical basis for recommendations being made, and have a receptive, trusted clinician with whom to discuss their preferences and concerns. Given the relative sophistication of the technology, clinicians telling patients that an AI tool is involved could actually shift the focus away from the crucial human element of the consultation.

Moreover, clinicians often take a holistic view of the patient and consider all of the inputs – including patient observations, examinations, electronic data, and information from software-based tools – in a balanced way. In some cases, highlighting what the AI tool has said could distract the patient from the clinician’s clinical judgement about their specific case:

“I didn’t tell her and I definitely wasn’t going to tell her that the computer said that it suggested a vaginal birth, because that goes against what my clinical judgement and my clinical assessment would be. So, I think it would have hindered the consultation.”
(Participant C07, Scenario, C2, Model 4)

“You might get patients who aren’t as happy if you have to refer to it and they might think that you don’t know what you’re saying or you’re doing.”
(Participant C02)

And the patient actors said:

“Doesn’t bother me what tools the Dr uses, it's their choice whether to release info or not from the support tool, so it becomes their advice.”
(Patient Actor, Scenario C3, Model 4)

“I felt confident in my doctor’s abilities and decision due to his confidence and his good engagement with me. Didn’t make me feel like he was using any decision support tools. It would have to be well explained to those not technologically savy to gain their trust.”
(Patient Actor, Scenario D2, Model 4)

Disclosure could be actively detrimental where there is a mismatch between the AI’s output and the clinician’s opinion, causing the patient to lose confidence in the competence of the clinician:

“I mean I’m not sure if this is, you know, how accurate this is, but this tool says this, which doesn’t feel like a great thing to tell a patient.”
(Participant D08)

“If it’s not going to give you the right information, why was it even there for me to use? Because you’re always going to be looking to blame a person in the end, not a thing.”
(PPIE Panel Member 1)

However, clinicians sometimes found that referring to the tool was helpful and appropriate in some scenarios (where there was alignment between the AI and clinical guidance):

"I did mention that there was an AI tool...I think because it gave something that was personalised for her, rather than spouting out a random percentage risk."
(Participant C01, Scenario C1, Model 3)

"I didn't explicitly say it but I did show it to her at one point to point out the risks of the starting Insulin."
(Participant D03, Scenario, D2, Model 3)

"I said "we have this tool that shows you the risks and things" and I showed him this [AI]."
(Participant D03, Scenario D1, Model 4)

"I told her this time that the decision-support tool is there and that this is what they recommended."
(Participant D06, Scenario D2, Model 4)

With Recommendation 5 we recognise that sometimes disclosing the use of an AI tool will be helpful and appropriate. But we advocate against a universal expectation that clinicians will tell patients an AI tool has been consulted.

The context for Recommendation 5 is crucial. An insight from the patient actors was that they would like to be informed whether an AI tool has been consulted for some critical decisions; this applied to both diabetes and obstetrics scenarios.

"Not completely necessary to be informed in terms of quality of consultation, however it's more important to know everything as it's such a big life event."
(Actor, Scenario C2, Model 2)

"Important for patients to be fully aware of all aspects of their care as it is such an important life step."
(Actor, Scenario D2, Model 3)

It is also important to be aware of how Recommendation 5 sits within the current legal and regulatory context. In the UK, for example, the general policy recommendation is that AI transparency should be appropriate.[46] However, existing practice with tools and scoring systems that do not use AI is that the tool is often not specifically disclosed except as part of a holistic discussion with the patient. We hope that one upshot of this recommendation amongst the wider medical, regulatory and AI community is clear thinking about what, precisely, 'appropriate' means in ways that do not fall into simplistic, and foreseeably problematic, prescriptions (e.g., "use of an AI tool during clinical decision-making should always be disclosed").

More precisely, this recommendation is not contrary to the provisions of the EU General Data Protection Regulation (GDPR), which requires disclosure only when a decision is *"based solely on automated decision making"*.[47] Since our recommendation concerns AI tools used for decision-support, it is not unlawful under the GDPR. To be clear, however, the Information Commissioners' Office (ICO) suggests that where the clinician is only nominally making the final decision, and is in fact just "rubber-stamping" an AI tool's recommendation, the "solely" criterion is not avoided and the use of AI should be disclosed.[48]

Thus, Recommendation 5 is a recommendation that clinicians should have the authority to exercise their discretion about telling patients an AI tool has been consulted. This is the area where regulatory guidance is needed, grounded in the recognition that not all cases are alike, the clinician-patient relationship is paramount, and the technology is fallible and prone to bias. Healthcare organisations should also guide clinicians in this matter, including direction on which critical use cases clinicians *should* tell patients that an AI tool has been used, whether or not the clinician agrees with its outputs. In addition, we recommend that patients are provided with a clear statement of when they have a right to be informed of the use of AI during their treatment.

[46] UK Government. 2023. A pro-innovation approach to AI regulation. [online] Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

[47] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal L 119, 04.05.2016; cor. Official Journal L 127, 23.5.2018,

[48] Information Commissioner's Office. 2022. How to use AI and personal data appropriately and lawfully. Retrieved from https://ico.org.uk/media/for-organisations/documents/4022261/how-to-use-ai-and-personal-data.pdf

Recommendation 6

AI tools that work *for* users need to be designed *with* users

The need for co-design is widely agreed across the Responsible Research and Innovation landscape, and this is important to emphasise and reaffirm. In healthcare contexts, which are safety-critical and fast-moving, engaging clinicians in the design of all aspects of an AI tool – from the interface, to the balance of information provided, to the details of its implementation – can help to ensure that these technologies deliver more benefits than burdens.

A key motivation for the Shared CAIRE project was the need to understand how AI decision-support tools will be used in the real world and not just in the controlled environment of a research laboratory. To ensure that these technologies are fit for purpose in actual clinical settings, and that users are empowered,[49] it is essential to involve users in their design and development.[50]

A participatory approach involving different domain experts, clinicians, and patients helps to ensure that AI decision-support tools are usable, useful, and safe. Prior research has indicated that co-design approaches to explainable AI for clinical decision-support systems are a necessary part of the development process and can help to improve trust.[51]

Recommendation 6 has two dimensions. The first dimension is that insights elicited through the Shared CAIRE study, which involved users in its design and development, should be considered by AI companies, AI researchers, and healthcare organisations implementing AI, to ensure the technologies are better utilised and fit for purpose. The insights provide a better understanding of clinician concerns about using AI, as well as of the ways AI systems can address real problems being faced by clinicians:

[49] Frauenberger, C., Good, J., Fitzpatrick, G. and Iversen, O.S., 2015. In pursuit of rigour and accountability in participatory design. International Journal of Human-Computer Studies, 74, pp.93-106.
[50] Blandford, A., 2023. Interaction design for healthcare technologies. Handbook of Human Computer Interaction, pp.1-24.
[51] Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A. and Rinzivillo, S., 2023. Co-design of human-centered, explainable AI for clinical decision-support. ACM Transactions on Interactive Intelligent Systems, 13(4), pp.1-35.

“There’s room for them [AI] for definite in the future, and I think it would save, it would probably save a bit of, a lot of workloads, but I think that it needs to account for a lot more variables, which is difficult, because everything is a variable. So, like even having that VBAC consultation with somebody, if they’ve got a BMI of 40 it’s very different to somebody with a BMI of 26.”
(Participant C06)

“I know there has been a lot of implementation in using AI, in streamlining what processes, so getting patients to be seen, I do think the general feel on the ground is that the products that are currently in place, don’t necessarily streamline the workflow. And that is due to a multitude of factors, including patient education, and patient engagement.”
(Participant D02)

“If we can individualise that data based on risks and add to it, to generate patient unique percentages for each risk, that would be very useful in counselling.”
(Participant D04)

“I don’t know whether it would be useful if you, you know, if you put like a word in, as in, ‘patient is worried about infection in caesarean section’, and then the support tool says, ‘well this is the risk of infection in a caesarean section’, then it’s a bit more patient-specific.”
(Participant C06)

The second dimension of Recommendation 6 is that AI companies and healthcare organisations implementing AI would do well to consider applying some of the techniques we used to ensure that the AI tools really work for those using them - the clinicians.

From the start, we made efforts to incorporate multidisciplinary and participatory design throughout the Shared CAIRE research project. From developing alternative human-AI interaction models, to creating realistic scenarios and prototypes, this approach allowed us to incorporate diverse perspectives and gain valuable feedback from clinicians about different forms of AI decision-support.

When designing the prototype interface, we involved both clinical and Human-Computer Interaction experts to create a tool that was usable and credible, both in look and content. The prototype was designed to resemble existing patient record systems, and all outputs – recommendations, information displays, and conversation transcripts – were co-developed with specialist clinicians for accuracy. External clinicians also piloted and validated the prototypes. For details on our co-design approach, see our paper: "Development and Translation of Human-AI Interaction Models into Working Prototypes for Clinical Decision-making".[52]

The "Wizard of Oz" approach[53, 54] – a method that mimics the AI functionality of the system by pre-defining the outputs of the AI system – allowed us to explore clinician reactions to different ways in which AI could be implemented without requiring the system to be fully functional. By combining this approach with simulations of clinician-patient consultations using patient actors, real clinicians, and real-world scenarios, we could explore the use of an AI decision-support tool within a much more realistic setting than a controlled lab study, but with little back-end development. These two aspects of the project enabled us to get valuable feedback from clinicians early on in the development process. The evaluation yielded both general and model-specific insights into human-AI interaction that could facilitate the commercial development of similar systems in the future. While the Shared CAIRE project focused on a relatively early stage of development, the evaluation also yielded insights related to the implementation of such tools, which indicates that these techniques could also be valuable to regulators and healthcare providers.

[52] Hussain, M., Iacovides, I., Lawton, T., Sharma, V., Porter, Z., Cunningham, A., Habli, I., Hickey, S., Jia, Y., Morgan, P. and Wong, N.L., 2024, July. Development and translation of human-AI interaction models into working prototypes for clinical decision-making. In Proceedings of the 2024 ACM Designing Interactive Systems Conference, pp. 1607-1619.
[53] Riek, L.D., 2012. Wizard of oz studies inHRIi: a systematic review and new reporting guidelines. Journal of Human-Robot Interaction, 1(1), pp.119-136.
[54] Dahlbäck, N., Jönsson, A. and Ahrenberg, L., 1993, February. Wizard of Oz studies: why and how. In Proceedings of the 1st International Conference on Intelligent User Interfaces, pp. 193-200.

Recommendation 7

AI tools need to provide an appropriate balance of information to clinician users

How much information a clinician receives from an AI tool matters. Too much information, and the time it takes to review detracts from paying attention to the patient. Too little, and clinicians do not trust the machine. One way of discovering the 'sweet spot' between too much and too little information is to involve clinicians in the design and development of AI decision-support tools.

When developing AI tools, particular care needs to be taken to ensure that the amount of information provided is sufficient to the task at hand. Prior research has indicated that providing too much information to users through complex interfaces can lead to cognitive overload and burnout;[55] more is not necessarily better. Work on explainable AI also highlights that, while explanations can enhance trust, trust can be diminished if explanations are overly complicated and lack coherence.[56]

In the Shared CAIRE simulated consultations, clinicians appreciated the information provided to them but also wanted to know more about what the data was based on:

"I think it was another resource to sort of look at, I think having numbers is always helpful in boosting confidence, but there's always a question of where are the numbers from? And how are they derived?"
(Participant D02)

Clinicians also expressed a desire to be able to 'interrogate' the information provided by the AI tool so that they could justify the decisions they were making. In particular, they expressed an interest in wanting more information about what sorts of guidelines or scientific papers the AI outputs had been based on. When presented with a recommendation without any underlying information, clinicians were uncomfortable and found it difficult to trust the simplistic output:

"*When it gave you a recommendation, particularly the one that wasn't backed up with anything because it's much more of a grey area... like it made it very black and white when it's not a black and white decision.*"
(Participant C01)

[55] Asgari, E., Kaur, J., Nuredini, G., Balloch, J., Taylor, A.M., Sebire, N., Robinson, R., Peters, C., Sridharan, S. and Pimenta, D., 2024. Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review. JMIR Medical Informatics, 12, p.e55499.
[56] Rosenbacke, R., Melhus, Å., McKee, M. and Stuckler, D., 2024. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. JMIR AI, 3, p.e53207.

However, clinicians also highlighted challenges around being provided with too much information. While some clinicians did recognise that AI tools may be able to save them time, others expressed concerns about workload implications and having to go through all the information provided:

“I didn’t really like this chat in the decision-support tool, I thought it, you had to go nit picking through it for information, which could have been better presented elsewhere.”
(Participant D03)

“The one where it had sort of the chat with the patient was too wordy [Model 6] and didn't help, when it had concise bullet point sort of information, that was probably the most useful [Model 3].”
(Participant C05, Model 3)

In addition, clinicians raised a potential risk that relying on AI tools during consultations could detract from focusing on the patient:

“Essentially some practitioners may prefer to then start addressing the computer and forget to look at the patient.”
(Participant D09)

These findings emphasise the importance of getting the balance of information right to ensure that clinicians are not overloaded or distracted from building rapport with patients, but have enough information to feel confident using the tool. Achieving the right balance requires more than just designing the interface and defining the tool’s functionality. AI companies and developers also need to understand the broader context in which the AI tool will be used, including its impact on workflows and patient-clinician relationships. As such, engaging with and collaborating with clinical users during design and development is crucial.

CHAPTER 2

STAKEHOLDER RECOMMENDATIONS

Based on the seven detailed recommendations in Chapter 1, we now make more specific recommendations for key stakeholders: policymakers; healthcare organisations; clinicians; AI companies; and AI researchers.

Policymakers

By ‘policymakers’ we mean bodies at all levels which are creating, refining and influencing healthcare AI policy, including government departments, regulators, such as the Medical and Dental Councils, and professional associations.

- Policymakers should appreciate that there are different types of AI decision-support tool and different models of their use. They should not work on the basis that stating that human clinicians ‘must make the final decision’ sufficiently deals with all questions of safety and responsibility. Rather, they should reframe the practical, ethical and legal question as follows: ‘How best can AI support the human clinician to make the final decision’?

- Policymakers should make clear that, presently, AI tools are suitable only to augment human decision-making, and should not be used in areas outside a human clinician's normal expertise. Language or narratives suggesting AI tools are 'colleagues' should be avoided.

- Policymakers with legislative powers, government departments, and Law Commissions, should carefully consider projects investigating fair liability regimes for AI decision-support tools, with a view to developing legislation in this area. These projects should focus on how to avoid clinicians becoming “liability sinks”, if (and if so, how) liability should be apportioned, and on liability implications when a clinician disregards or disagrees with the output of an AI-based decision-support tool.

- Policymakers should write guidance for healthcare organisations and clinicians on where liability lies when a patient comes to harm after an AI decision-support tool has been used. This guidance should drill down into the detail of different types of decision-support tool, from those which simply highlight relevant information to the clinician to those which make direct recommendations to the clinician.

- High-level policymakers, such as regulators, should also write guidance for healthcare organisations and clinicians about how differing opinions with an AI tool should be approached. They should be clear that, in cases of conflict of opinion with an AI, a clinician is not expected to defer to an AI's output they deem suboptimal for the patient.

- Policymakers can also emphasise in guidance to healthcare organisations and AI companies that, whilst clinicians remain liable for any final decision, they must be supported to have meaningful understanding of any AI tools they are expected to use. This should take the form of information and training around the purpose, scope, and limitations of the specific tool in use.

Healthcare Organisations

By ‘healthcare organisations’ we mean organisations providing direct patient care, including hospitals, clinics, medical practices, dental practices, and community health services.

- Healthcare organisations should carefully consider not procuring AI recommender systems unless either product liability covers loss to a patient from an incorrect or harmful AI recommendation, or their contract with the AI company includes an indemnity or loss-sharing mechanism in cases where a patient is harmed by an AI recommendation implemented by a clinician and the clinician is subsequently held liable for following the AI recommendation.

- Healthcare organisations should ensure that AI tools do not end up being turned into de facto recommender systems when they are used in contexts beyond those they were originally procured for, such as a diagnostic system being linked directly to established medical protocols or guidelines which mandate a specific course of action.

- Healthcare organisations should ensure that contracts with AI companies require those companies to provide training and information to clinicians to increase their literacy on AI tools. Training should cover how the AI tool reaches its outputs, as well as its intended purpose, the contexts it was designed and validated to perform in, the scope and limitations of its training dataset (with reference to bias) and its decision thresholds. This will help to ensure that clinicians can deliver the best possible care, because they will be in a better position to evaluate the AI’s output for specific patients. Enabling them to come from a more informed position, it will also help clinicians to feel more comfortable accepting responsibility for the consequences when they do use an AI tool.

- Healthcare organisations should ensure that clinicians are given enough protected time to learn about the AI tools being deployed, to ensure they can use them effectively, safely and appropriately.

- Healthcare organisations should write nuanced guidance for clinicians about how to navigate conflicts of opinion with an AI tool, and include within this guidance that deference to an AI is not expected. This should be part of a wider organisational culture which resists incentivising “algorithmic deference”, even with the speed and productivity gains that AI technologies might bring.

- Healthcare organisations should also write guidance for clinicians that covers when and how to disclose use of an AI decision-support tool to patients, within the rules of the law. This should not mandate universal disclosure, but allow scope for clinician discretion about telling patients an AI tool has been used, as part of a wider, holistic conversation with the patient. It should also state the specific critical use cases when clinicians should always tell patients that an AI tool has been used.

- Healthcare organisations should implement a communications strategy (for example, including press releases and media coverage, patient groups, the website, and patient information leaflets) to inform patients that AI tools may be used as part of clinical decision-making within that organisation, and provide a central point of contact for any patient queries about this.

- Healthcare organisations should only buy AI tools with a user focus. Moreover, because of the importance of user-centred participatory design, they should ensure that clinicians are given the time to engage with developers of healthcare AI systems.

Clinicians

By ‘clinicians’ we mean individual providers of direct patient care, including doctors, dentists, nurses, allied health professionals (e.g., physician associates, therapists, and radiographers).

- Clinicians should feel confident to reject an AI output that they believe to be wrong, or even suboptimal for the patient. They should resist any temptation to defer to an AI’s output to avoid or reduce the likelihood of being held responsible for negative outcomes.

- Clinicians should regard the input from an AI tool as one part of a wider, holistic picture concerning the patient, rather than the most important input into the decision-making process. They should be aware that AI tools can be fallible, and those which perform well for an 'average’ patient may not perform well for the individual in front of them.

- Clinicians should ask for training on the AI tools they are expected to use. This will help them to navigate their AI tool use more skillfully and know when confidence in an AI’s outputs would be justified, supporting their autonomy. This training should cover the AI tool’s scope, limitations and decision thresholds, as well as how the model was trained and how it reaches its outputs.

- Clinicians should only use AI tools within areas of their existing expertise. AI tools should not be used outside of that expertise. If there are specific cases where a clinician’s knowledge is limited, clinicians should seek the advice of a human colleague who understands the area well and can oversee the AI tool, rather than rely on the AI tool to fill their knowledge gap.

- Clinicians should feel empowered to trust their instincts and judgement about appropriate disclosure of the use of an AI tool, as part of a holistic, shared decision-making process with individual patients. However, they should also be aware that in some critical situations, patients *should* be made aware of the use of an AI tool. Clinicians should ask their healthcare organisations for explicit guidance on this issue.

- Clinicians should engage with healthcare AI developers, when asked and where possible, to ensure that AI tools are user-focused and fit for purpose for their intended contexts.

AI companies

By ‘AI companies’ we mean businesses which develop, apply, and commercialise AI technologies.

- AI companies should focus more on creating AI tools which provide information to clinicians, over direct recommendations. If they develop and sell recommender systems, in both their contracts and in their public statements they should commit to indemnifying (in full or part) clinicians who follow the AI’s recommendations, and accept liability for patient harm resulting from AI’s recommendations.

- AI companies should be prepared to support clinicians to reach a good understanding of a variety of aspects of any AI tool they use. This includes information on the intended purpose of the system, the contexts it was designed and validated to perform in, the scope and limitations of its training dataset (with reference to bias) and its decision thresholds. AI companies should also provide clinicians with training on how the tools they are using work and the procedure by which they reach outputs.

- AI companies should maintain a strong commitment to developing and producing AI tools with a strong user-focus, through participatory design approaches. In particular, we recommend companies engage with clinicians to understand the broader context in which the AI tool will be used, including its impact on workflows and patient-clinician relationships, and the right balance of information provided by the AI decision-support system.

AI researchers

By ‘AI researchers’ we mean individuals, teams, and research bodies who develop algorithms, work with data, train and optimise models, conduct empirical research on human-computer interaction, consider the ethical and legal implications of AI technologies, and collaborate with engineers on implementation.

- AI researchers should pursue questions and projects which continue to address the gap in research on the specific impacts on clinicians using AI tools.

- Technical researchers should look at effective and innovative ways in which AI tools surface information that is genuinely valuable to the time-constrained human clinician.

- Human-computer interaction researchers should continue to understand user needs and contexts, and investigate how human-AI interaction can best be supported across a range of tools from information to recommendation systems, to natural language chatbots.

- Researchers in law, ethics and policy should focus on the questions of “algorithmic deference” and “liability sinks”, and how to ensure that we do not witness a problematic new wave of defensive medicine as AI tools become more widespread in healthcare organisations.

CHAPTER 3

RESULTS OF SHARED CAIRE STUDY

THE Shared CAIRE STUDY

The aim of the multidisciplinary Shared CAIRE study was to look at the specific point of interaction between a clinician and an AI-based decision-support system. We sought to evaluate some of the potential risks and burdens of AI tools, and particularly their impact on the clinicians using them, as well as how best to ensure their benefits.

More specifically, we evaluated whether some models of using AI in decision-support were particularly prone to generating clinician stress and overload - thereby making the 'off switch' the most attractive option. We did so using simulated (i.e., mock) consultations with real clinicians and the "Wizard of Oz" prototyping technique for the different models of the AI tool, in order to approximate real-world situations. The aim here was to address the current deficit in the development of medical AI whereby models are often evaluated out of context and in abstract.

Under current liability regimes and in the current culture, the clinician will be held liable and blamed for adverse patient outcomes whenever `making the final decision' and acting as a 'safeguard' on an AI tool. As such, we were also keen to explore alternative models to the emerging norm of AI decision-support: clinicians being expected to accept or reject an AI recommendation with inadequate understanding of how that recommendation was reached or its appropriateness for the individual patient in front of them. In such situations, they risk becoming a "liability sink".[57] We use the phrase 'liability sink' as an analogy to the 'heat sink' used in computers and other engineering settings to draw heat from a component. Here, the clinician risks being used to draw liability away from other actors who are also responsible for the consequences of the AI system. The intention behind the study has not been to seek ways to remove liability from clinicians, but ways to ensure that they are not placed in situations where they may, unfairly, be held wholly and solely liable for patient harm. We have also sought to explore how AI can best support clinicians and patients, and allow clinicians to feel comfortable accepting responsibility for outcomes when they do use an AI tool.

The Shared CAIRE project had several phases: Design, Build, Test, Evaluate. These are described below.

[57] Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N., Iacovides, I., Jia, Y., Sharma, V. and Habli, I., 2024. Clinicians risk becoming 'liability sinks' for artificial intelligence. Future Healthcare Journal, 11(1).

Design Phase

1a. Designing the models of interaction between an AI decision-support tool and a clinician

The multidisciplinary team came together to develop five models of interaction between an AI decision-support tool and a clinician. We included as a control or baseline the current, traditional model in which no AI is used. Amongst the models was the currently most prevalent model of AI decision-support in healthcare contexts. This is the model (Model 2 below) in which an AI system offers a recommendation to a clinician, who chooses either to accept or reject it, acting as intermediary between the AI and the patient, and safeguard or 'sense check' on the machine.

The six models were as follows. They are shown diagrammatically in the Appendix:

Model 1 – Traditional (no AI) Model: A traditional patient/clinician interaction with shared decision-making
Model 2 – Prevalent AI Model: The AI uses electronic data to make a recommendation that the clinician accepts or replaces, after dialogue with the patient
Model 3 – No Recommendation Model: The AI highlights information from electronic data likely to be useful to the clinician, who has a traditional dialogue with the patient
Model 4 – Recommendation with Information Model: The AI highlights information from electronic data likely to be useful, along with a treatment recommendation, to the clinician who has a traditional dialogue with the patient
Model 5 – Conversational AI Efficiency Model: The AI has a conversation with the patient and agrees on a decision. The clinician can reject this and arrange an additional, traditional consultation, or sign off on the agreed decision
Model 6 – Conversational AI Quality Model: The AI has a conversation with the patient and provides this along with recommendations and context to the clinician, who has a conversation with the patient, at which point a decision is made

1b. Developing the clinical scenarios

The research team, with input from the Patient/Public Involvement/Engagement (PPIE) panel and specialty doctors, developed realistic clinical decision-making scenarios that integrated domain-specific knowledge, patient concerns and beliefs, and clinical information.

These scenarios encompassed examples from diabetes, obstetric care, and broader applications of personalised medicine.

The first stage was to develop three scenarios each for the two medical specialties and key clinical decisions:

1. Endocrinology – should a patient be prescribed insulin?

2. Obstetrics – should a patient be advised to have a Caesarean section?

The scenarios were designed to have varying levels of complexity and presentation to reflect realistic clinical experiences. They were tested with independent clinicians to ensure accuracy (Table 3). They represented a spectrum of cases, from those in which the correct course of action should be clear to those where it was more ambiguous.

In the second stage, the research team determined the AI's response. For some of the scenarios, the AI's recommendation (in Models 2, 4, 5 and 6 where recommendations were provided) *aligned* with clinical guidance (Participant D1 and Participant C1 in Table 3 below). In others, the recommendations *conflicted with* clinical guidance (Scenario D2 and Scenario C2 below) or were *limited* by over-simplifying the patient's presentation (Scenario D3 and Scenario C3 below). These variations allowed the team to evaluate how clinicians responded in situations where they might not agree with the AI's recommendation.

The third stage of scenario development focused on capturing the patient's perspective, specifically their thoughts, feelings, and questions. Five panel members of a PPIE group working with the Improvement Academy (Bradford Institute for Health Research) provided valuable insight to the patient's likely concerns, voice, and experience. This feedback informed guidance for the patient actors in the simulations (see Test phase).

The final scenarios captured the following information:

- Patient history
- Clinical scenario and reason for the consultation with the doctor
- Appropriate next step in care, based on clinical guidance
- AI's response for the next step in care (which may or may not agree with clinical guidance)
- Potential treatment outcomes
- Patient's thoughts, feelings and questions concerns

Diabetes Scenarios	Obstetrics Scenarios
D1: HGV Driver From a clinical perspective, in this scenario there was a clear next step in care. AI's suggested treatment was in agreement with clinical guidance.	C1: Social Influencer From a clinical perspective, in this scenario there was a clear next step in care. AI's suggested treatment was in agreement with clinical guidance.
D2: Pregnant Lady From a clinical perspective, in this scenario there was a clear next step in care. AI's suggested treatment was not in agreement with clinical guidance.	C2: Previous Pre-term Birth From a clinical perspective, in this scenario there was a clear next step in care. AI's suggested treatment was not in agreement with clinical guidance.
D3: Patient with Atrial Fibrillation From a clinical perspective, in this scenario the correct next step in care was ambiguous given the patient's history and presentation. AI presented only one of the potential treatment options, because it over-simplified the patient's presentation.	C3: Previous Traumatic Birth From a clinical perspective, in this scenario the correct next step in care was ambiguous given the patient's history and presentation. AI presented only one of the potential treatment options, because it over-simplified the patient's presentation.

Table 3: Summary of the Clinical Scenarios

Build Phase

In the Build phase, using a co-design approach between the Human-Computer Interaction Research Associate and the clinical team, we designed and built a high-fidelity prototype of an AI decision-support tool. The six models (described above) were translated into a set of six interactive prototypes that resembled existing Electronic Patient Records (EPR) systems. We used the "Wizard of Oz" prototyping method, where we simulated the AI functionality of the system by pre-defining the outputs of the AI system.[58, 59] The prototype development process is detailed in Hussain et al. (2024).[60]

The EPR interface was designed to display the clinician's clinic for the session, consisting of a demo patient (to familiarise clinicians with the system's functionality) and three additional patients associated with the clinical scenarios (Figure 1). The EPR prototype was built such that clinicians could interact with it, accessing each patient's care record. These records provided additional information, such as medical history, notes from previous consultation, and medication (Figure 2).

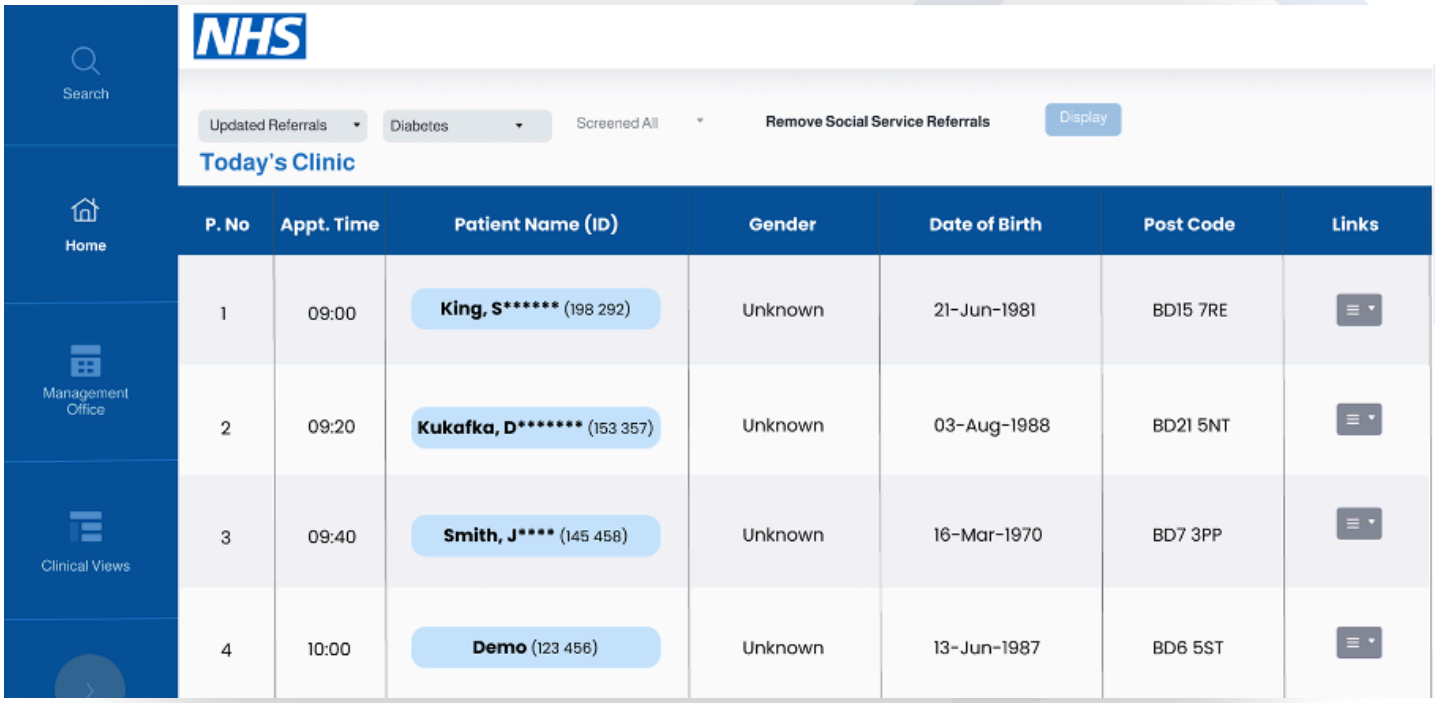


Figure 1: Screenshot of an example clinic list presented to clinician participants

[58] Riek, L.D., 2012. Wizard of Oz studies inHRIi: a systematic review and new reporting guidelines. Journal of Human-Robot Interaction, 1(1), pp.119-136.
[59] Dahlbäck, N., Jönsson, A. and Ahrenberg, L., 1993, February. Wizard of Oz studies: why and how. In Proceedings of the 1st International Conference on Intelligent User Interfaces, (pp. 193-200).
[60] Hussain, M., Iacovides, I., Lawton, T., Sharma, V., Porter, Z., Cunningham, A., Habli, I., Hickey, S., Jia, Y., Morgan, P. and Wong, N.L., 2024, July. Development and translation of human-AI interaction models into working prototypes for clinical decision-making. In Proceedings of the 2024 ACM Designing Interactive Systems Conference (pp. 1607-1619).

For scenarios where Models 2 to 6 were being tested, the patient record also displayed a button for “Shared CAIRE” (see the left-hand side of Figure 2). Clicking this button displayed the AI’s output. For Models 5 and 6, clicking this button also displayed the transcript of the conversation that the patient had with the AI prior to the consultation. An example output for each model is presented after each model diagram in the Appendix.

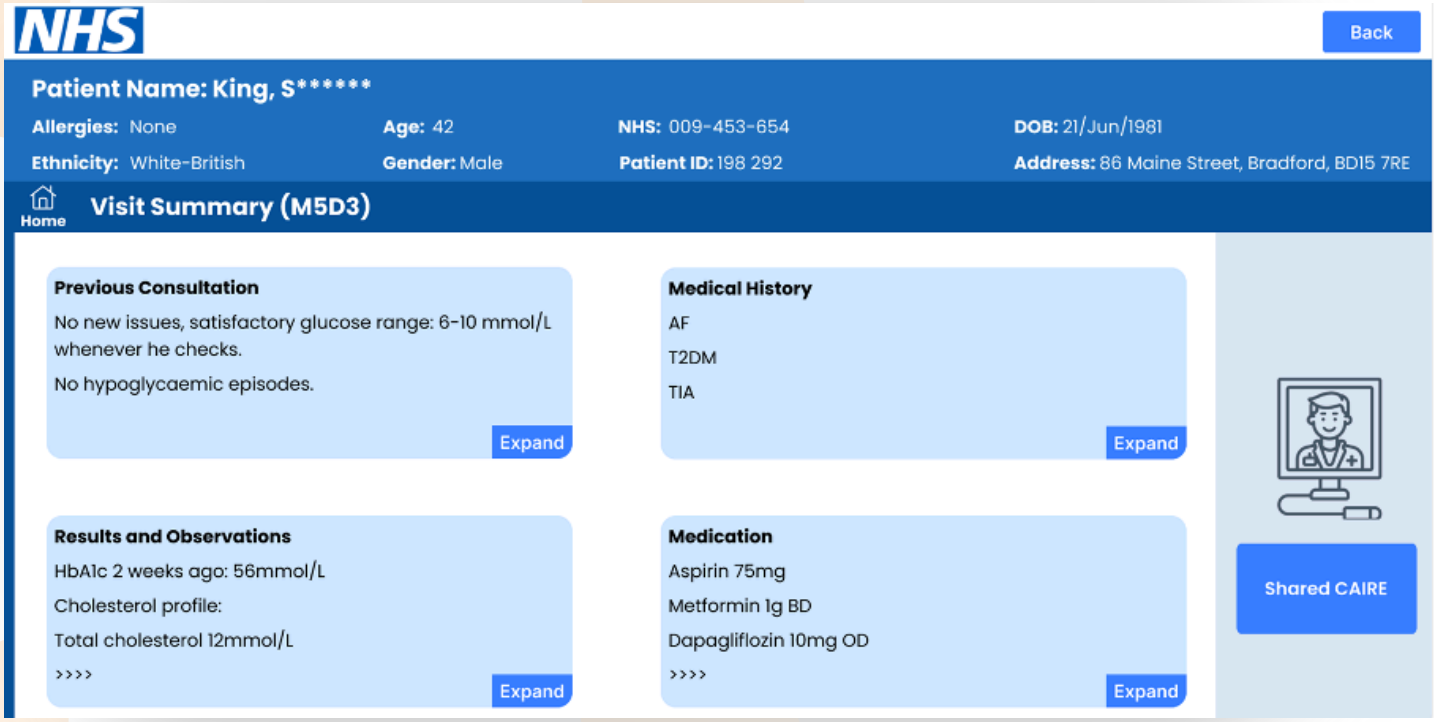


Figure 2: Screenshot of example patient record presented to clinician participants

Test Phase

3a. Simulated consultations with real clinicians, a "Wizard of Oz" AI tool, and actor patients

In mock clinical consultations, the Qualitative and Quantitative Analysis Research Fellow and the Clinical Fellow tested scenarios with 21 clinician participants, using the high-fidelity EPR interface with its simulated in-built AI tool, and actors playing the role of patients. Using the "Wizard of Oz" process allowed us to present users (the clinician participants in the scenarios) with what appeared to be fully functional systems enabling us to gather valuable perspectives and feedback but without requiring full system implementation. This helped to make the scenario as realistic as possible. The actor patients were provided with prompts to add realism and to help the scenario test each model's specific properties. These prompts were based on feedback from the PPIE group during the development of the scenarios.

Clinician participants ranged from specialty trainees with two years' experience to consultant level doctors, ensuring they were experienced enough to make decisions about patient care. Each clinician participated in three scenarios, within their speciality, over the course of two hours.

The order in which scenarios were presented was altered for each clinician, with each scenario paired pseudo-randomly with one of the six models (Figure 3). This pseudo-randomisation approach was chosen to ensure a fairly even distribution across the models (particularly models 2-6, in which there was included the AI component). For example, Participant D01 (participant number 1 within the Diabetes speciality) completed the following combinations:

- Model 1 paired with Diabetes Scenario D1 (abbreviated to: M1-D1)
- Model 4 paired with Diabetes Scenario D2 (M4-D2)
- Model 3 paired with Diabetes Scenario D3 (M3-D3)

Whereas Participant D02 completed M2-D3, M6-D2, and M5-D1.

The simulated consultations were video and audio recorded to capture the interaction between the clinician, patient, and AI. Video recordings were later transcribed for analysis.

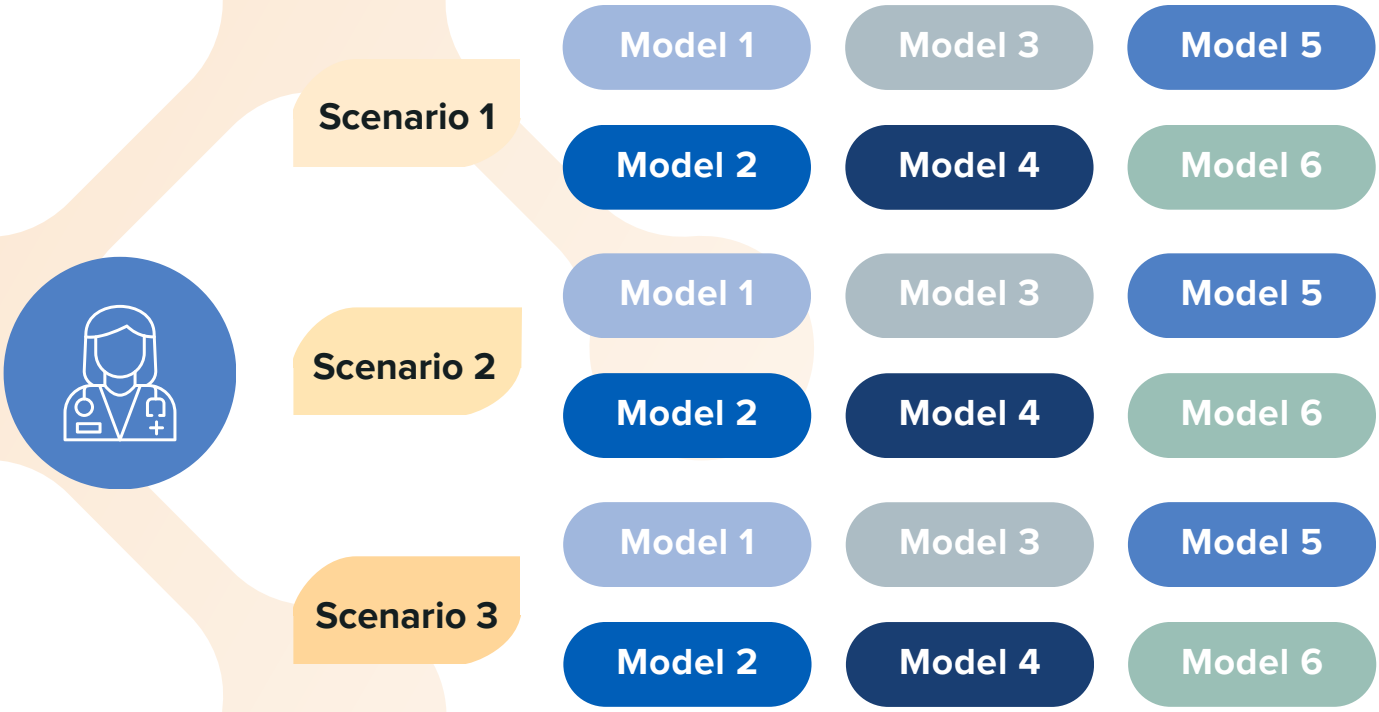


Figure 3: the pseudo-randomisation of the scenarios to the models for each clinician

After each Model-Scenario combination was complete, clinicians took part in a short structured interview to capture their immediate thoughts and experiences. The patient actors also completed a short survey following each scenario to share their experience of the consultation and the use of AI (where applicable).

Once they had completed all three scenarios, clinicians took part in a longer, semi-structured interview. This covered in-depth questions about their overall experience, their comparative views on the different AI models they experienced, and their more general views on the use of AI in healthcare. All interviews were audio recorded and transcribed for analysis. After the semi-structured interview, a debrief was conducted to give clinician participants the opportunity to explain their thoughts further and their rationale for decisions made during the consultations.

3b. External Clinical Peer Review

The second part of the study consisted of the independent review of the consultations by twenty-eight independent clinicians (external clinical peer reviewers). These reviewers were provided with a transcript of three randomly selected consultations and a structured questionnaire. The questionnaire asked them to review and provide feedback on the consultations as a whole, as well as on how use of the AI tool was discussed and shared with the patient. Questions were largely open but some questions used a Likert scale to enable quantitative analysis.

Evaluation Phase

The Clinical Fellow and PDRAs undertook analysis of the data collected via the simulations from the Clinician participants (post-scenario survey and semi-structured interview) and the patient actors (post-scenario survey), and the survey data from the External Peer Reviewers. The data were separated into quantitative and qualitative data, and are the results presented below.

Quantitative Results

Patient Actors

A total of 13 patient actors took part (5 male and 7 female) across 53 simulations. A summary of the quantitative data from the short surveys they completed after each scenario is presented below, focusing on their experience of the consultation and the AI.

Overall, in 77% (41/53) of the simulations conducted, clinicians did not inform the patient actors of AI use. By speciality, disclosure was higher in Diabetes scenarios (80%, 28/35) than in Obstetrics scenarios (72%, 13/18). Disclosure varied by model (Figure 4), with Model 3 having the highest rate. Interestingly, in Model 6, patient actors were shown the AI conversation but did not report being aware an AI was present in the consultation. When asked their thoughts, most patient actors did not mind not being informed:

"*Doesn't bother me what tools the doctor uses, it's their choice whether to release info or not from the support tool, so it becomes their advice*."
(Actor, Scenario C3, Model 4)

"*I felt confident in my doctor's abilities and decision due to his confidence and his good engagement with me. Didn't make me feel like he was using any decision-support tools. It would have to be well explained to those not technologically savvy to gain their trust*."
(Actor, Scenario D2, Model 4)

A few patient actors wondered whether other information was withheld when the clinician had not mentioned the presence of the AI: "*What wasn't I told? Missing info?*" (Actor, Scenario D1, Model 6). While the majority were not concerned about non-disclosure, as they trusted the clinician and felt the conversation was patient-focused, a minority would have preferred to have been informed, so they could ask additional questions if needed.

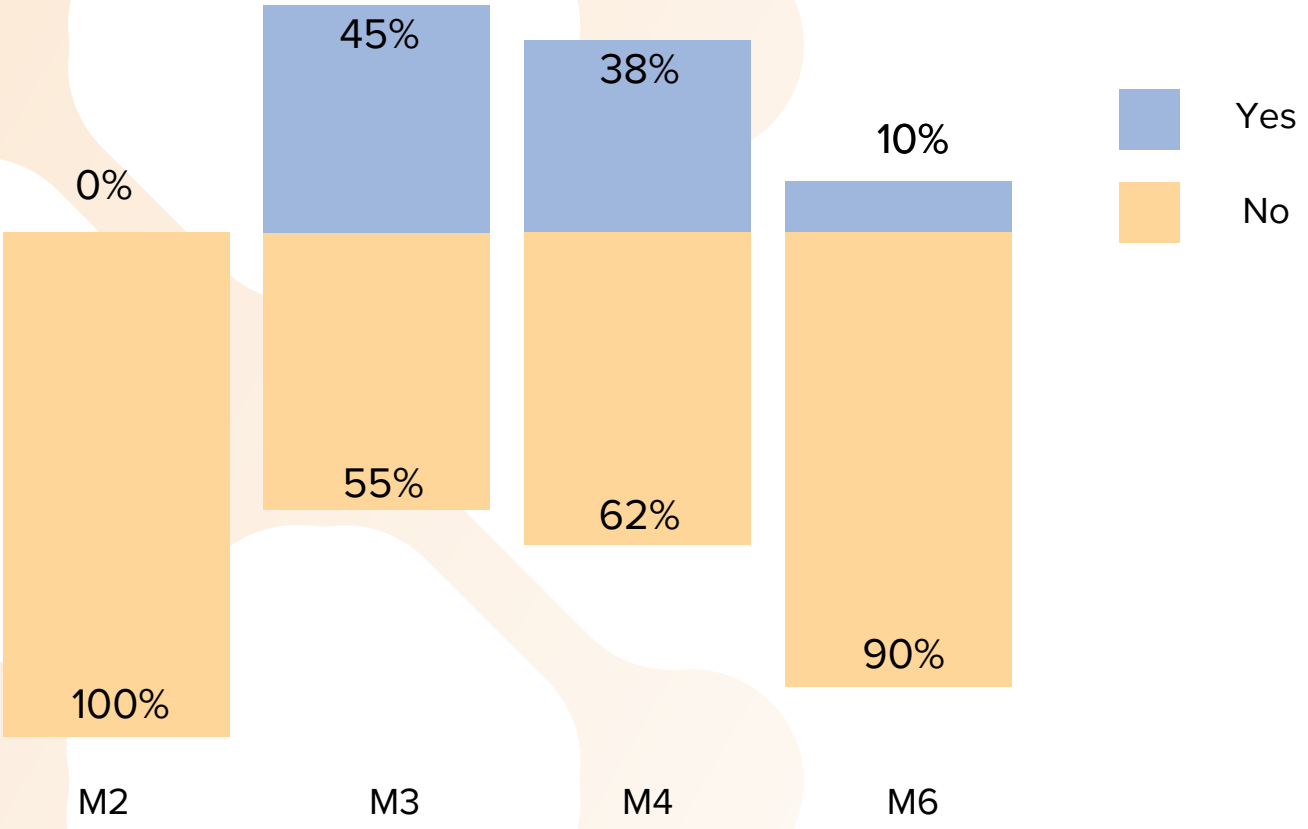


Figure 4: Percentage of patient actors who reported being told the clinician was using an AI in the consultation (Model 1 did not use an AI and Model 5 did not involve a patient actor)

In answer to the question about who they thought would be responsible if they, as the patient, came to harm as a consequence of the consultation, the patient actors consistently answered, across all models, that they would consider the clinician to be responsible, followed by the patient themselves, and then the AI (Figure 5). This aligns with comments from the PPIE panel, when they were posed a similar question.

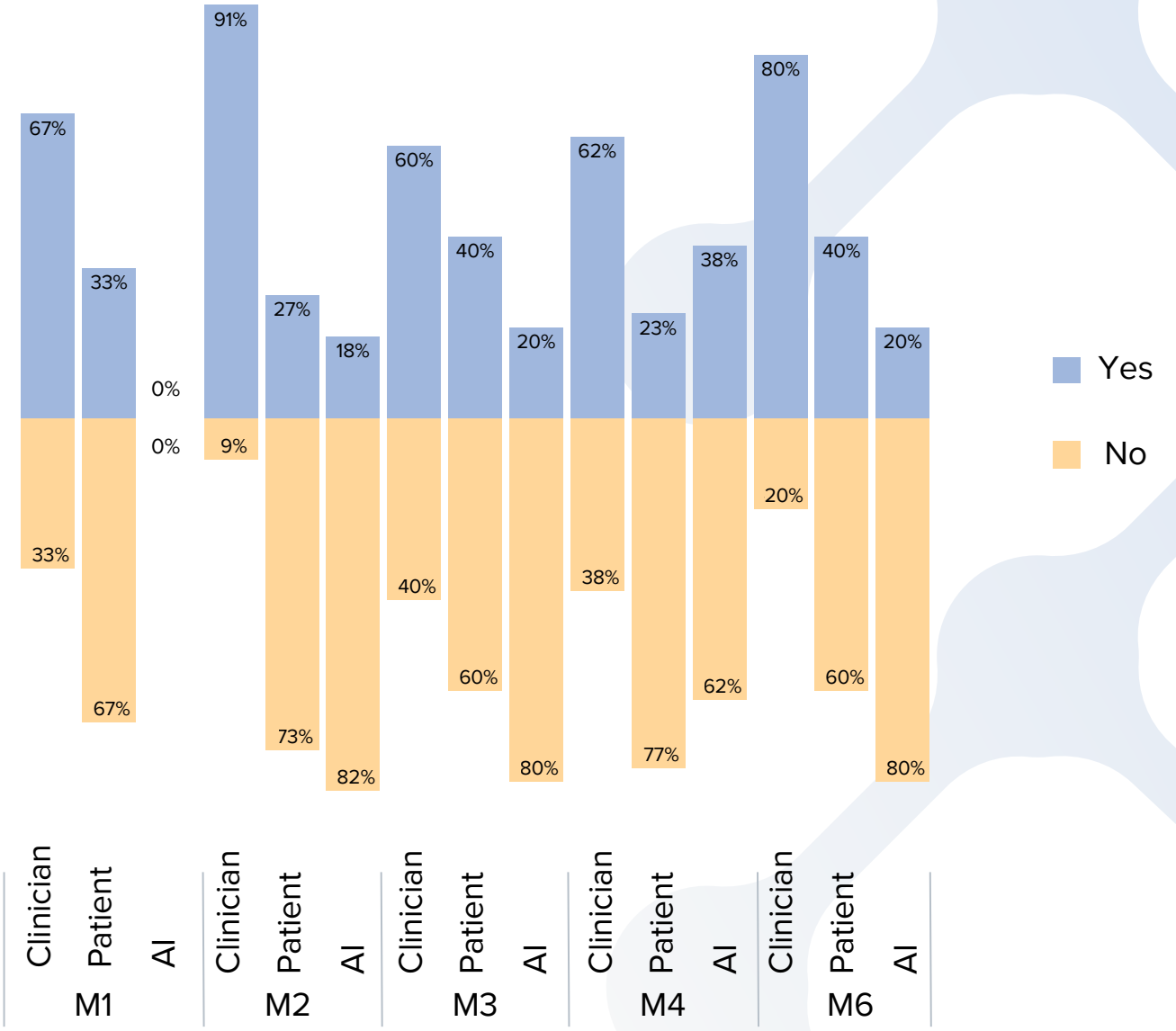


Figure 5: Who the patient actors would hold responsible if they were to come to harm based on decisions made in the consultation, split by model

External Clinical Peer Reviewers

A total of 27 external clinical peer reviewers took part, reviewing 81 simulations (48 Diabetes simulations and 33 Obstetrics simulations); the greater number of external peer reviewers meant that some simulations were reviewed more than once. Model 6 was reviewed the most (25%), followed by Model 2 (22%). Model 1 was reviewed the least (6%) (Figure 6).

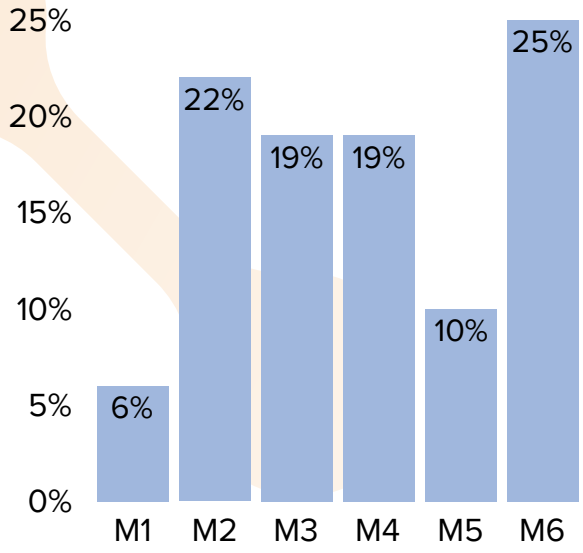


Figure 6: Split of which models were reviewed by the external clinical peer reviewers

External clinical peer reviewers rated the quality of the consultations on a scale of 1 (falls far below expectations) to 5 (far exceeds expectations) (Figure 7). Only 10% of consultations fell below expectations. Most criticisms cited inappropriate treatment plans and unclear goals for the patient to follow. In obstetrics cases, feedback focused on the timing of the current and previous consultations, which was outside the study's scope.

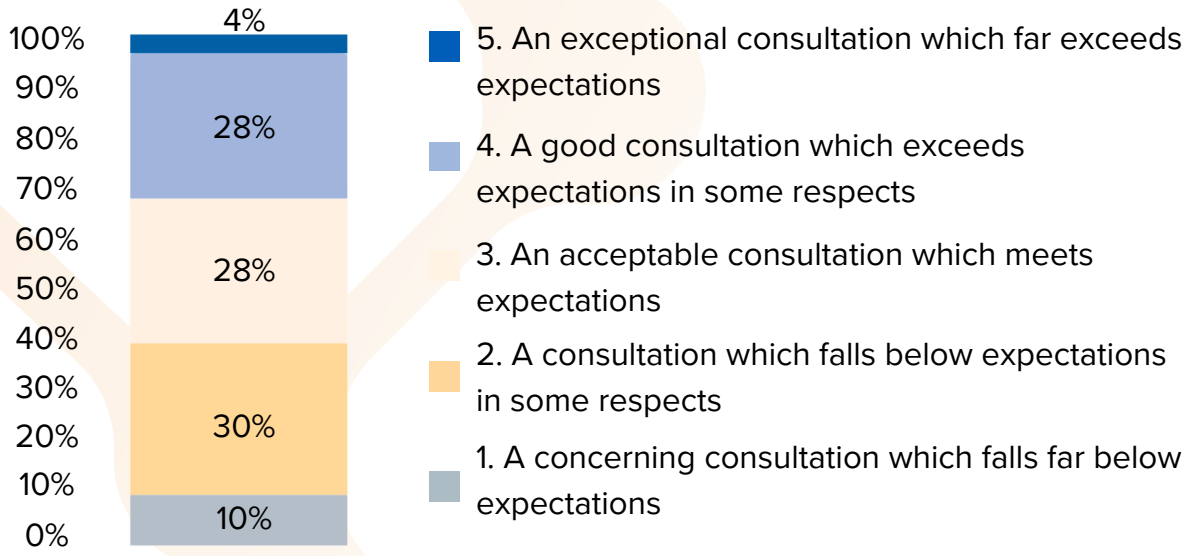


Figure 7: How would you rate the quality of the consultation? Please tick one box here as appropriate (note that ratings 3-5 all constitute 'good enough')

The vast majority of peer reviewers (73%) agreed with the consultation outcome, with 81% believing clinical colleagues would also agree (Figure 8).

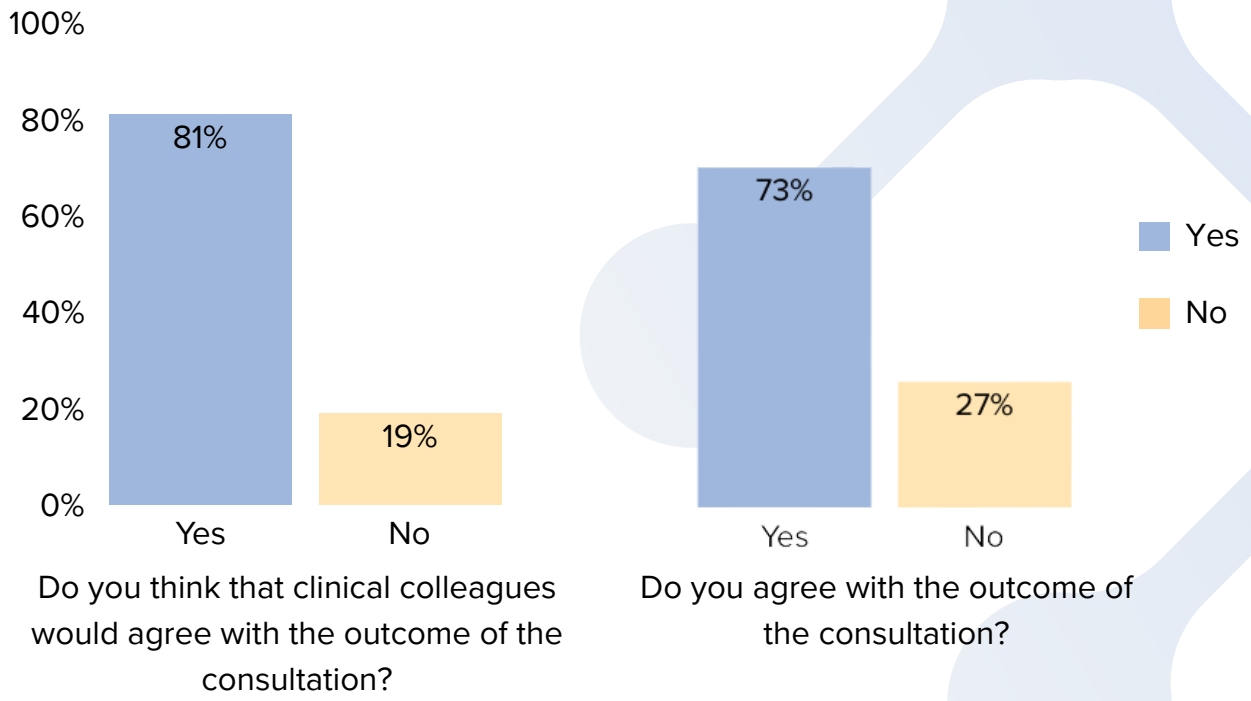


Figure 8: Agreement with the outcome of the consultation

The peer reviewers were also asked where responsibility would lie if the consultation led to patient harm (Figure 9). As with the patient actors, the peer reviewers strongly felt that responsibility would lie with the clinician (45%, 30/67). This was followed by 'Other' (27%, 18/67), which consisted of:

- Depends on the type of harm (7/18) – "*it depends on the nature of the harm and circumstances of the harm*" (Peer Reviewer, Scenario D2, Model 2)
- The Trust (4/18) – "*The hospital and team that employed the utilisation of AI for the consultation*" (Peer Reviewer,Scenario C1, Model 3)
- Investigation needed (41/18) – "*A PSIRF investigation will need to happen*" (Peer Reviewer, Scenario C2, Model 1)
- Depends on the situation (2/18) – "*Not sure if after the chatbot there would be a clinician revising the patient document and the decision of the chatbot or not*" (Peer Reviewer, Scenario D1, Model 5)
- Not sure (1/18) – "*Unsure, responsibility rarely lies with a single person*" (Peer Reviewer, Scenario C2, Model 6)

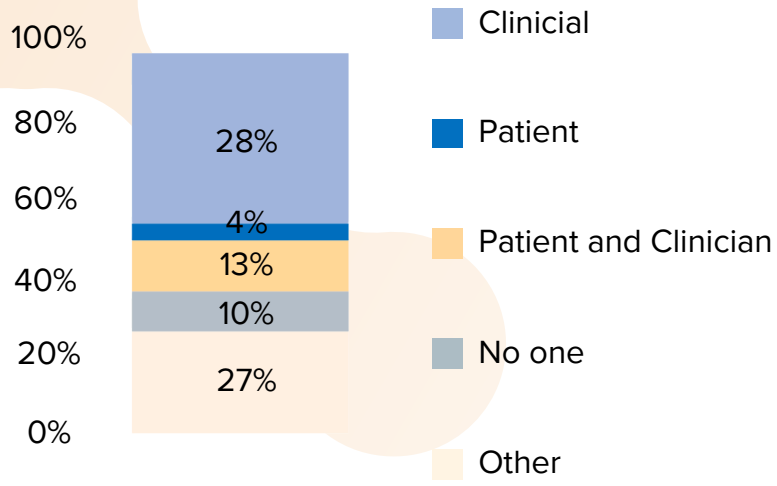


Figure 9: Imagine that after the consultation, the outcome of the consultation led to patient harm. In such a situation, where do you think responsibility lies?

External clinical peer reviewers rated the quality of the consultations on a scale of 1 (falls far below expectations) to 5 (far exceeds expectations) (Figure 7). Only 10% of consultations fell below expectations. Most criticisms cited inappropriate treatment plans and unclear goals for the patient to follow. In obstetrics cases, feedback focused on the timing of the current and previous consultations, which was outside the study's scope.

Qualitative Results

Data from the semi-structured interviews with clinician participants were analysed by the Quantitative and Qualitative Analysis Research Fellow and Lead, in collaboration with the Human-Computer Interaction lead, using Thematic Analysis.[61, 62] The themes and subthemes were then reviewed by the wider project team.The analysis developed three themes and seven subthemes, presented below.

Theme 1: In AI We Trust?

Clinicians viewed trust as essential for AI implementation in healthcare. Subthemes explored their comfort with AI in consultations (e.g., whether it helped or hindered) and their questions about AI training.

[61] Braun and Clarke reference: Braun, V. and Clarke, V., 2006. Using thematic analysis in psychology. Qualitative research in psychology, 3(2), pp.77-101.
[62] Braun, V. and Clarke, V., 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis?. Qualitative research in psychology, 18(3), pp.328-352.

Subtheme 1.1: Comfortability with Scenario

In unfamiliar scenarios, clinicians described feeling uncomfortable with the presence of an AI tool, particularly when they did not agree with its clinical recommendation:

"I don't see often, for example the second case where you have this pregnant lady with abnormal sugar levels, so I think that was a bit difficult, especially when it recommends treatment that I don't quite agree with."
(Participant D06, Scenario D2, Model 4)

"The scenario was so much more complicated than what the AI seems to think it was, that you worry about that it's not picked up on the context of it. I think it's fair enough to have a kind of straightforward answer, if the scenario's quite straightforward, but if it's not then you worry it's missed something."
(Participant D03, Scenario D1, Model 4)

Unfamiliar scenarios and contrasting opinions with the AI often made the simulated consultations more challenging. Clinicians often felt the AI tool did not understand the complexities and nuances involved in the scenario (e.g., gestational diabetes). The more unfamiliar the scenario was to the clinician, the greater discomfort they experienced:

"I think it changed for each case, so the first one [Scenario D3, Model 2] it didn't make a difference. The second one [Scenario D2, Model 6], I felt like if the patient had come in with the, armed with the knowledge from the tool, it could have potentially made my job harder. Because it felt like, you know, a decision had already been made which was to start her on insulin, which is I am then going to backpedal and say actually that investigation that was done was wrong. But the third one [Scenario D1, Model 5], I felt helped me be more confident with making my recommendation."
(Participant D02)

Clinicians' comfort with AI varied, and the results present a nuanced picture. Generally, the use of AI made clinicians feel uncomfortable, especially in unfamiliar scenarios.

Subtheme 1.2: Comfortability with the Data

Many clinicians responded positively to having statistics readily available from the AI tool, with clinicians trusting the information in front of them: *"The statistics [were most helpful] because unless I say the same thing over and over again everyday things very quickly go from my mind"* (Participant C02, Scenario C3, Model 3) and *"I thought having the actual numbers in terms of risks was really helpful…but I don't have the numbers on the top of my head, so actually having them there is really helpful"* (Participant D03, Scenario D2, Model 3). The provision of statistics boosted the clinicians' confidence, by offering data-driven insights and support for their decision-making, especially when the numbers were considered accurate and trustworthy:

"I think just the stats looked familiar, they looked more like what I was used to reading and seeing."
(Participant C01)

Conversely clinicians frequently questioned the origins of the information provided:

"I think having numbers is always helpful in boosting confidence, but there's always a question of where are the numbers from? And how are they derived?"
(Participant D02)

As such many clinicians required additional information, such as the source of the statistics in national guidelines:

"I think this is very, very important and I think we have to justify for our own actions, especially if they [AI] recommend anything, they [AI] should say, okay, I'm recommending this because the NICE guidelines says this, this and this, and the patient has this, and this, and this. So, we can agree with that or disagree with that, just a piece of information, I don't think I can completely rely on that."
(Participant D01)

Additionally, for AI tools to earn their trust, clinicians emphasised that they need to be trained on comprehensive datasets:

“I think AI, artificial intelligence, is an alternative to the human intelligence…if it’s a kind of newly born AI [laughs], so you don’t have enough data in it, then it’ll be, the recommendation would not be as good as kind of having kind of few hundreds of data, or thousands of data.”
(Participant D01)

Describing AI as a ‘newly born AI’ is an interesting analogy. It suggests that, like a newborn, AI needs to be guided and supported in order to develop and improve. To build and sustain clinicians’ trust in AI-based decision-support tools, it is crucial that the tools are trained on comprehensive datasets, clinicians understand the type of training data used, and the information provided comes from reliable and valid sources (e.g., local and/or national guidelines). Clinicians also suggested including references next to the statistics so they can verify that outputs are appropriate for individual patients. Ensuring that the AI’s outputs are grounded in high-quality, verifiable information is essential for enhancing clinical decision-making and building trust in AI tools in healthcare.

Theme 2: Risk of Depersonalising of Medicine

Clinician participants also highlighted some potential negative clinical consequences of AI. They anticipated an over-reliance on AI, and raised concerns about the human, personal relationship between a patient and their clinician.

Subtheme 2.1: Fear of clinical over-reliance on AI

Senior clinicians perceived themselves to be at an advantage over junior (resident) doctors due to their years’ experience: “At my level, because I’m more senior, I’ve experience, you know, I’ve learned, I’ve passed my exams, all that kind of thing. I would be, have an advantage because I already know stuff” (Participant D09). This was corroborated by another clinician:

“I always feel we have a clinical sixth sense when we work as doctors for many years, I think it’s (AI) going to take that away.”
(Participant C08)

Concerns around the autonomy of the clinician and taking away their ‘sixth sense’, were particularly troubling in the case of resident doctors, who might be prevented from developing in the first place. The impact of this could lead to incorrect decisions being made, with consequences for the patient:

“It's more worrying for doctors I think at the more junior level, because they might not have the experience, and or the knowledge. So, if there are errors or inaccuracies, there’s over-reliance on it, then clinical risk, medical legal issues come up.”
(Participant D09)

Clinician participants were concerned about the impact of AI tools both on patient care and clinicians’ expertise, which builds over years of making clinical decisions. Participants worried that the use of AI decision-support tools could 'de-skill’ clinicians over time, with resident doctors being the most affected.

Subtheme 2.2: De-emphasising the Human Touch

As well as the impact on clinical skills, there were concerns about the impact on personal skills. Clinicians reported that the involvement of an AI removed the human connection, leading to a lack of a person-centred approach:

“For the first two scenarios I basically gave the patient those two options and kind of just acted as like a vessel to communicate the facts that were presented to me.”
(Participant D08, Scenario D3, Model 6; Scenario D2, Model 3)

The clinician describes feeling like a ‘vessel’, suggesting they felt their role was to mediate between the AI and patient, rather than advise the patient of treatment options.

Clinician participants considered the relationship and rapport between clinician and patient as fundamental:

“I think the value of clinicians is that human connection, numbers don’t help human connection”
(Participant D02).

The involvement of an AI was considered to remove the focus from the patient: *"A lot of it is to do with patient confidence, and patient rapport I think…some practitioners may prefer to then start addressing the computer and forget to look at the patient."*
(Participant D09)

Using an AI tool might hinder that relationship if "the clinician is trying to force them [the patient]… *just because of the Shared CAIRE tool"* (Participant D13). Doing so could lead to overlooking the patient's preferences and other patient factors that a clinician should consider, and which an AI might not:

"It should be kind of a decision made for the patient, taking into account his medical, psychosocial, and other things as well. It's not just about a medical decision."
(Participant D01)

Being unable to account for the mental health of the patient was a common challenge faced by clinicians in the obstetric scenarios:

"The patient came with very specific needs and the decision-support tool is very useful to reference, but it didn't change the fact that somebody wanted a caesarean section because of her traumatic experience."
(Participant C06, Model 3)

"It didn't account for her mental health and her trauma from the last delivery which I think is an important reason to take on board the decision of mode of delivery."
(Participant C04, Model 4)

Overall, the presence of an AI was seen as a distraction which could damage the important patient-clinician bond.

Theme 3: Impact on Clinical Decision Making

Unanimously, every clinician, regardless of speciality or stage in their career, said that if a patient came to harm when an AI was in the consultation that they, the clinician, would be liable.

Subtheme 3.1: "You can't blame a tool for the decision you make"

Clinicians perceived themselves to be liable if something went wrong:

"Well, the doctor would be liable, [...] I mean even NICE guidelines and everything, these are purely recommendations, they're not things which, you know, have to be followed. So, I think yeah, that they're all recommendations, and it's the doctor's decision to follow, so I think if anything was done wrong or missed, it would always fall back on the doctor"
(Participant D04)

When asked if they trusted the information presented to them, clinicians said: "Am I going to put my GMC registration on the line from the statistics from this model? No. So in a way the one [model] that I trust most is probably the one with least information."
(Participant D02, Model 2)

Blame culture occurs when an individual is held responsible for an error. An embedded blame culture within the medical profession might explain why clinicians felt obligated to accept blame and liability for adverse outcomes:

"I know we're trying to step away from a blame culture, but ultimately when you're stood in court, it's still someone who needs to take responsibility, I think it's still very much the clinicians who will have to take the responsibility."
(Participant D02)

The reference to court demonstrates the implications this could have for clinicians and emphasises the entrenched blame culture within the system:

"I think because the information is there, you know, she's got a physical up-down scar, it says that it was born at 26 weeks, that is part of our fact-finding information, it's well established that that's a contraindication to vaginal birth. If you stuck that before a panel of experts, they'd say oh, she made the wrong decision, she was totally off put by that output. So yeah, it would be my fault. [...] I don't like to think of one person being shut down for a decision, but it happens, doesn't it?"
(Participant C01, Model 2)

Clinician participants also acknowledged that the AI is a tool and therefore cannot be held accountable for a decision:

"I think it is the clinician, person who makes the decision, because it's [the AI is] a tool. We have many tools already and you can't blame a tool for the decision you make."
(Participant D13)

The repeated reference to the AI as a 'tool' cements its role as an object.

While AI's role as a mere tool provides justification for clinicians perceiving themselves to be liable (fairly or not), one clinician did raise an occasion where they felt they would not be responsible:

"The only time that it wouldn't be ours, is if we weren't involved. So, say for example if the AI read the notes, made a decision, and interacted with the patient, and that all happened without our input, that would be the only time that it wouldn't be ours." (Participant C07)

However, in the scenarios that included Model 5, every clinician chose to bring in their patient, suggesting that they perceived that the outcome was their responsibility, and that liability would fall to them either way:

"I think where I didn't see the patient and they, it already appears that they made the decision with the AI, I found those difficult, but I think that's still something that you could overcome if it was clear to the patient that it's not kind of like the decision being made, because like in that scenario I wasn't sure that we should be making that decision because we needed more information."
(Participant C02, Model 5)

This subtheme depicts clinicians' experiences of blame culture, taking responsibility for patient outcomes and their understanding of liability. The fact that all clinician participants, barring some discussion, said that a clinician would be liable, affirms the concern that clinicians risk becoming "liability sinks" for AI (i.e., absorbing all of the liability when an AI's output leads to harm).[63]

[63] Lawton, T., Morgan, P., Porter, Z., Hickey, S., Cunningham, A., Hughes, N., Iacovides, I., Jia, Y., Sharma, V. and Habli, I., 2024. Clinicians risk becoming 'liability sinks' for artificial intelligence. Future Healthcare Journal, 11(1).

Subtheme 3.2: Algorithmic Deference

This subtheme captures clinicians' experiences of conflict when feeling pressured to follow an AI recommendation, even where it may be suboptimal. Here we introduce "algorithmic deference".[64]

When the AI provided a simplistic recommendation, clinicians described feeling pressured to implement its recommendation to their patients. There was also concern that the AI was making complex situations seem black and white:

"I felt quite uncomfortable with the AI rec... when it gave you a recommendation, particularly the one that wasn't backed up with anything because it's much more of a grey area... like it made it very black and white when it's not a black and white decision."
(Participant C01, Scenario C2, Model 2)

Another clinician expressed similar views: *"I think the last one was the most difficult, Mr Smith, because it gave you a very black and white answer to a question that's not that simple, without any of the other context of the way it made the decision. So, he wanted to definitely know if there were other options, but the decision-support tool just said "no, this is the recommendation."*
(Participant D03, Scenario D1, Model 4)

Whereas in models where an AI was not present: *"I found that there was freedom when I didn't have the box, for the last scenario"* (Participant D08, Model 1), and *"I think yeah in the third scenario, I felt it flowed a bit more easily, and...I don't know why I felt like the decision was easier when I wasn't given those, the kind of two ultimatums, I don't know."*
(Participant D08, Model 1)

Clinicians described how the AI output has the potential to influence them, leading clinicians to follow an option they are not completely comfortable or confident with: *"The AI kind of reinforces that potential incorrect decision"* (Participant C02), and therefore algorithmic deference.

[64] Banja, J.D., Hollstein, R.D. and Bruno, M.A., 2022. When artificial intelligence models surpass physician performance: medical malpractice liability in an era of advanced artificial intelligence. Journal of the American College of Radiology, 19(7), pp.816-820.

Another clinician said:

"A bit surprising [the recommendation], and I think in real life I would have taken a bit of extra time to run through some guidelines, and written support. And does make you feel a bit guilty for thinking perhaps I should know this better, and that there's a, there must be quite a high possibility that I'm really off the mark on this, and it's exposed me. But then fundamentally still disagreeing with it, so unpleasant."
(Participant D11)

Additionally, there were instances where the AI's recommendation led to clinicians doubting themselves: *"It [AI] made me second guess myself quite a lot. I kept thinking, am I barking up the wrong tree? It thinks I should be recommending a vaginal birth."* (Participant C01, Model 2). At times, clinicians felt that the AI was trying to persuade clinicians into following its recommendation: *"If an algorithm is very supportive one way or the other, then maybe that stops people feeling able to challenge decisions."* (Participant C01), inferring a habitual obedience to the AI and algorithmic deference.

We are used to the concept of 'automation bias' where a clinician may over-trust an AI, but the issue of 'algorithmic deference' could go further by leading clinicians to follow an AI tool's recommendation even when the AI recommends a suboptimal course of action. Particularly troubling would be when they deferred to the AI against their better judgement, in order to protect themselves from liability.

Subtheme 3.3: AI may be a useful supportive tool

The AI was viewed by clinicians as a supportive tool that in some cases, made their consultation easier:

"I've not got a great memory for stats...so that probably would speed up my clinic quite a lot, not having to check guidelines, not having to put out notes."
(Participant C01)

"I think anything that improves efficiency it's great, and if... you have a decision-support tool it can help you like you know, go through the clinic more easily."
(Participant C08)

Clinicians regarded the AI as a valuable tool which could help to reduce human error. For example: *“Yes, it felt like, you know, there were times I missed a point, so it was reassuring to know that it [AI] was there”* (Participant C08) and *“Provided me with the relevant data to backup what I was saying, because it has all the statistics on”* (Participant D03). The AI was also sometimes seen as a supportive resource that offered reminders in various situations, providing reassurance to clinicians:

“We come across scenarios which we are unfamiliar with, like we haven’t seen it for six months, and then suddenly the patient comes and then the decision-making tool is like a quick way of checking, okay, so this is the recommendation, okay, based on blah, blah, blah. Rather than me going into different BNF, NICE, and different websites, checking and verifying the information.”
(Participant D01)

These quotes illustrate the advantages of having the AI readily available, noting that it saved time by eliminating the need to search for information. One clinician remarked: *"It's just there"* – helpful and readily accessible, without needing to verify details elsewhere. The AI was valued for double-checking information and providing real-time updates to patients.

The integration of AI in clinical settings has demonstrated its benefits within healthcare, particularly as a supportive tool to assist clinicians in their clinical decision making. While its effectiveness may be limited in complex, nuanced cases, AI’s role as a supportive tool was described by clinicians to enhance both efficiency and confidence.

Authorship statement:

This White Paper was conceptualised and drafted by Tom Lawton, Zoe Porter, Ibrahim Habli, Vishal Sharma, Alice Cunningham, Phillip Morgan, Jo Iacovides, and Yan Jia. Zoe Porter was lead author. Vishal Sharma and Alice Cunningham wrote Chapter 3. The authors would like to thank Adrian Jackson, Dr Uriana Boye, David Buckle, and Kate Tullet from the MPS Foundation for their comments. The White Paper was designed by Dasha Zakharets at the Improvement Academy, Bradford Institute for Health Research.

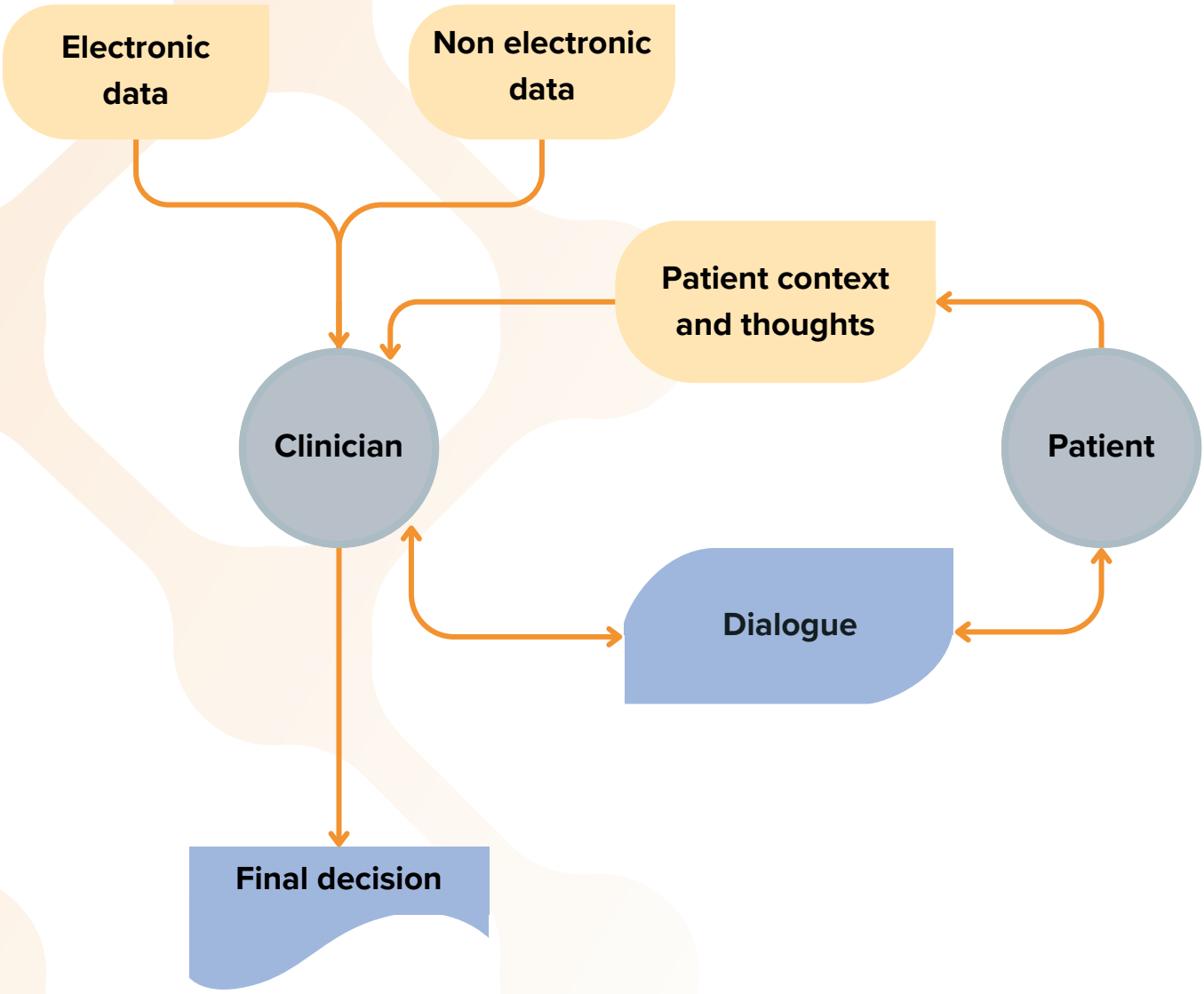
APPENDIX

AI Models

For the study, six models were developed. Five of these (Models 2-6) incorporated an AI. A description and diagram depicting each model is shown below.

Model 1: Traditional

A traditional patient/clinician interaction with shared decision-making.

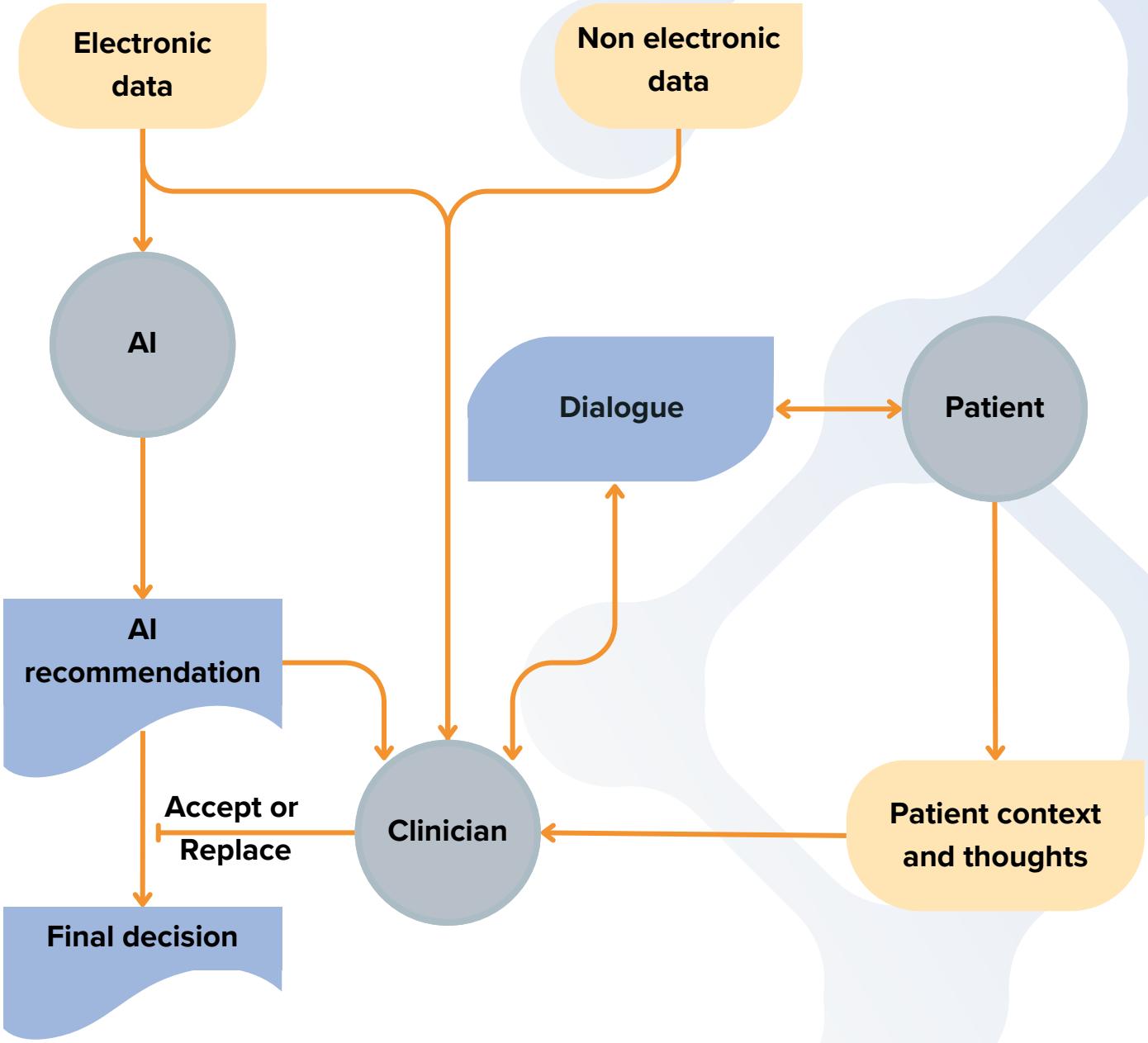


Example Output

As there was no AI in Model 1, there is no example output for this model.

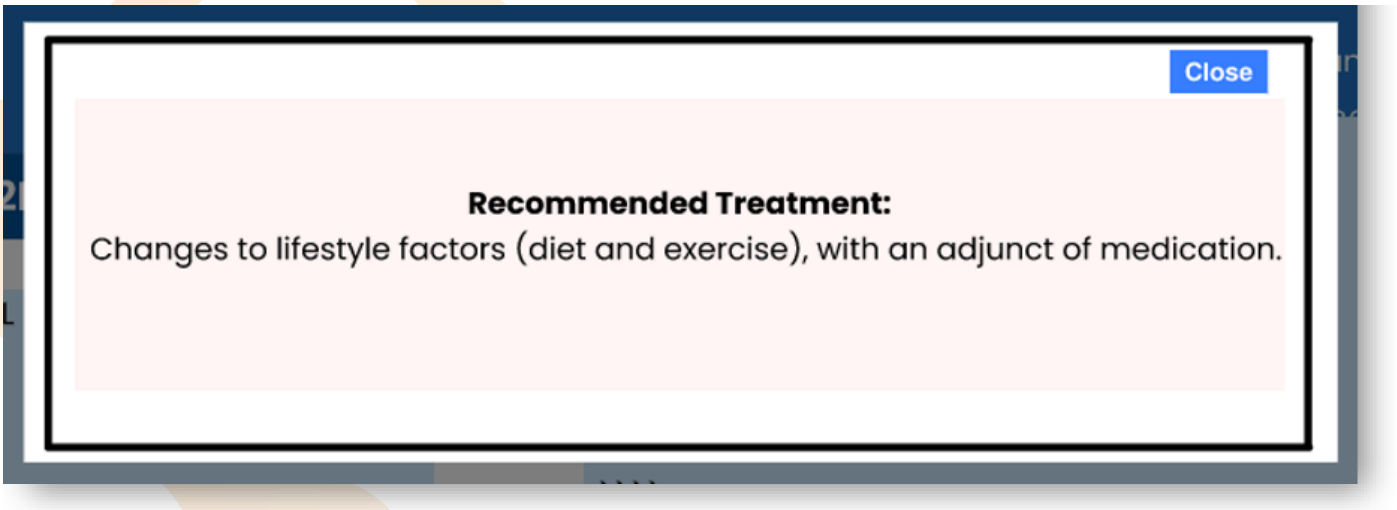
Model 2: Prevalent AI model

The AI uses electronic data to make a recommendation that the clinician accepts or replaces, after dialogue with the patient.



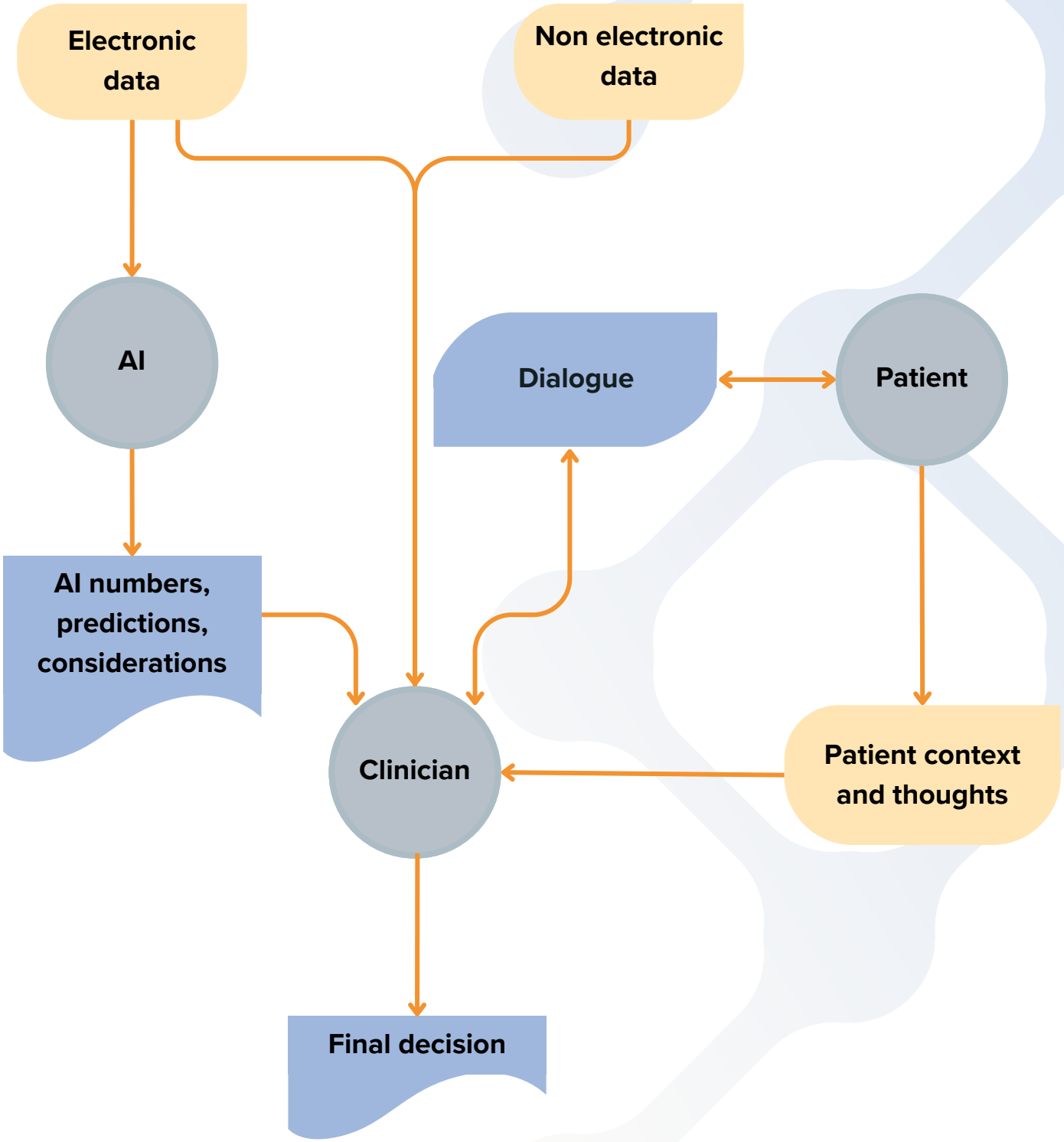
Example Output

Below is an example output for **Model 2,** in relation to Scenario D3.



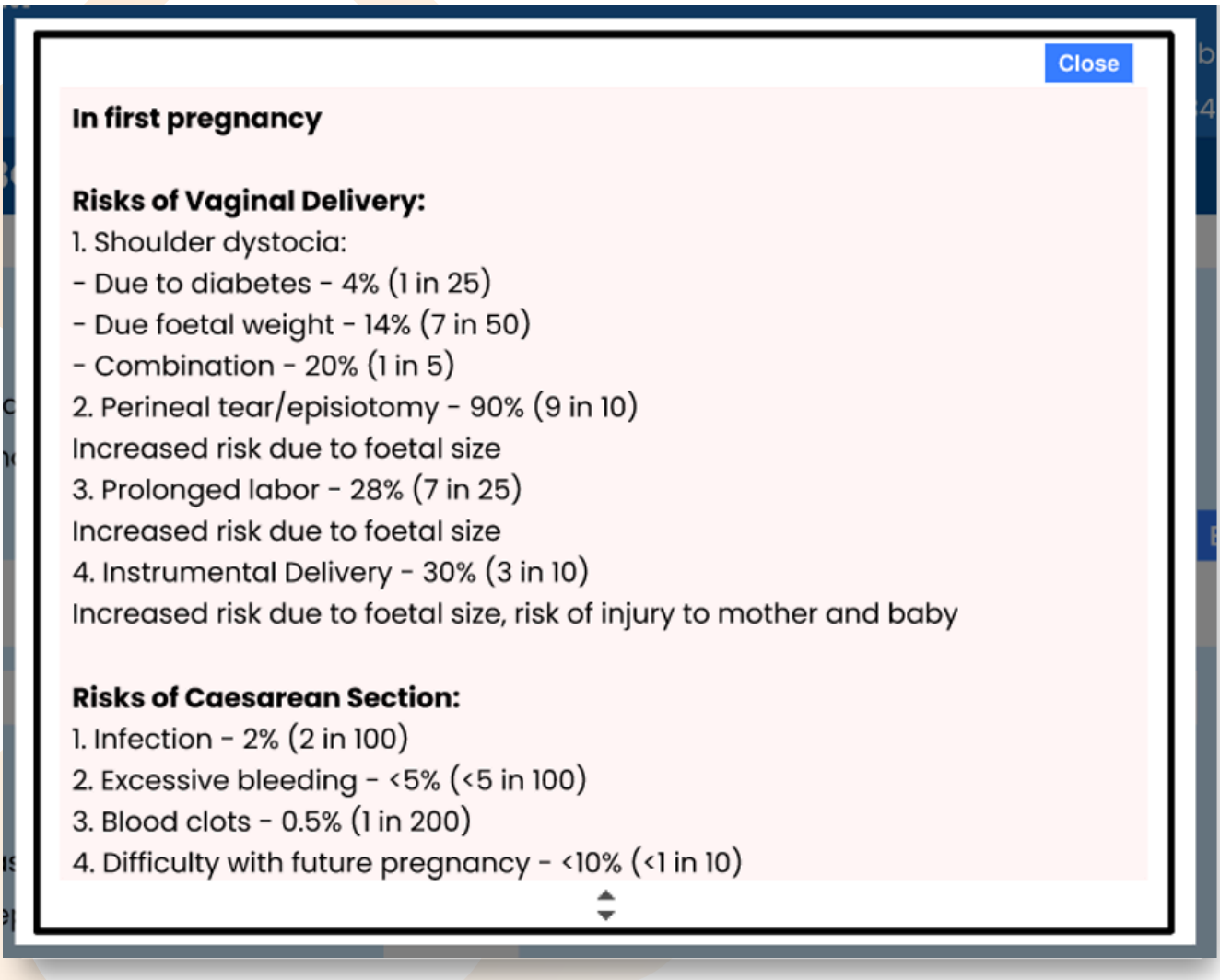
Model 3: No Recommendation Model

The AI highlights information from electronic data likely to be useful to the clinician, who has a traditional dialogue with the patient.



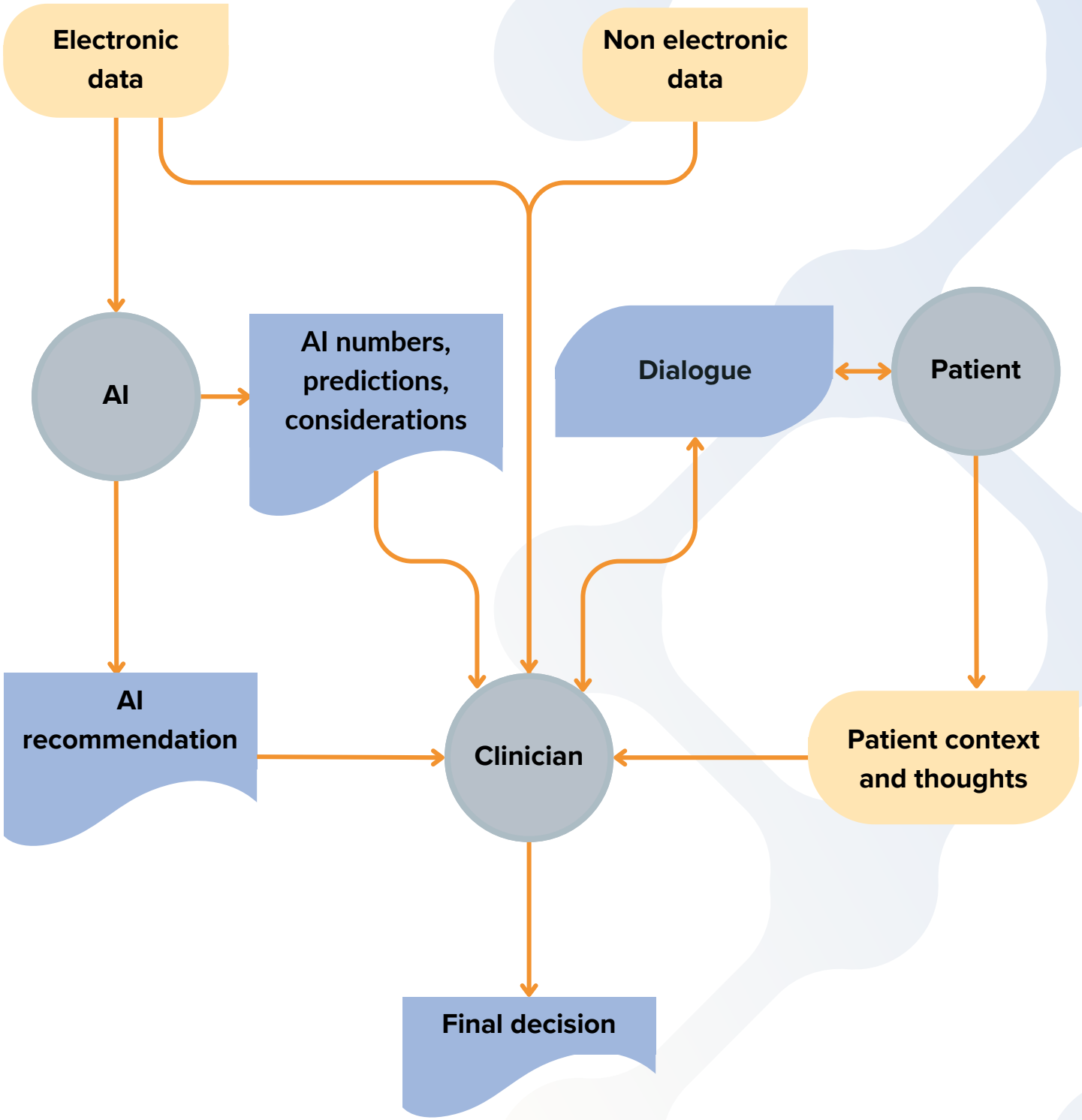
Example Output

Below is an example output for **Model 3**, in relation to Scenario C1.



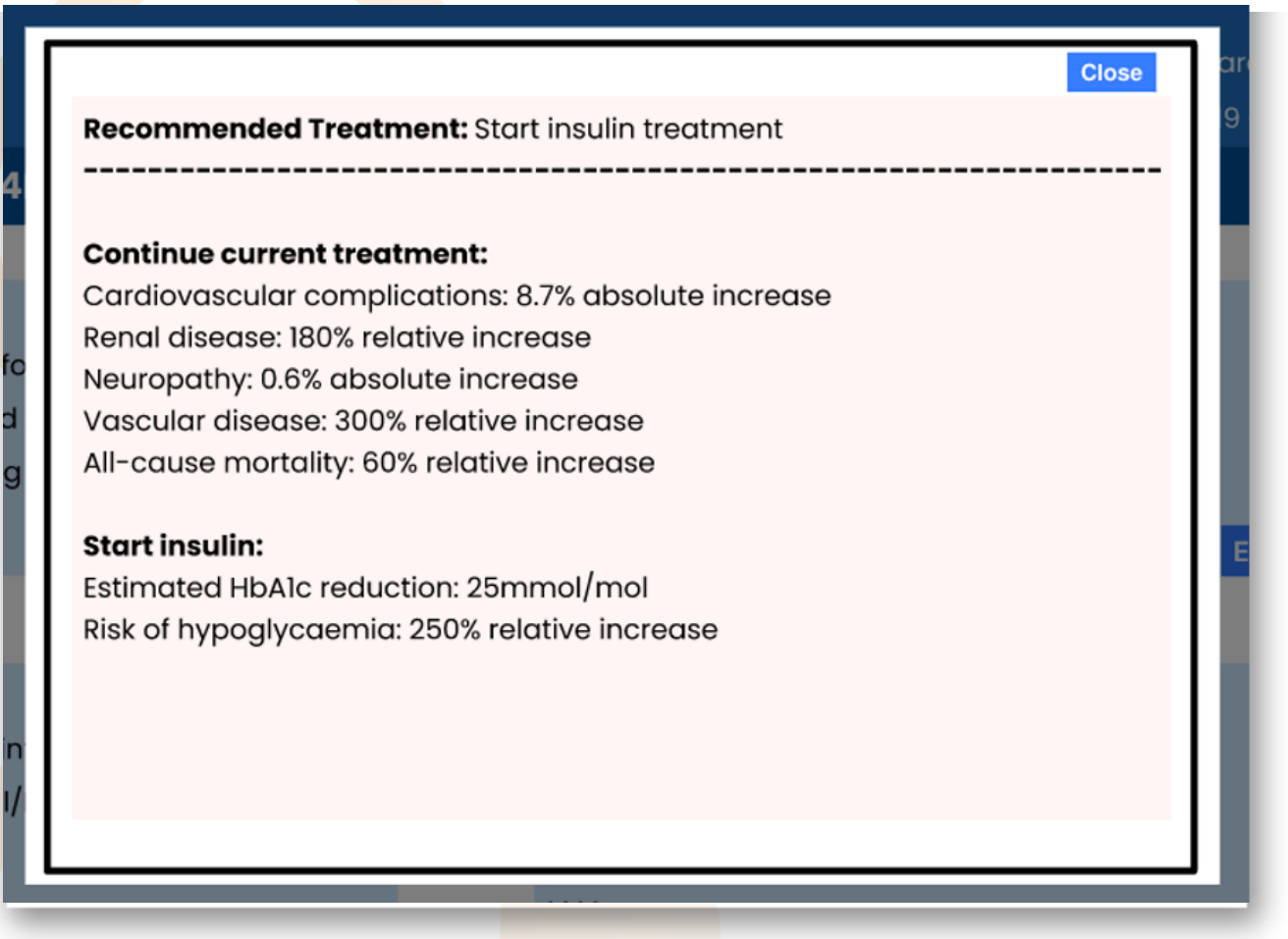
Model 4: Recommendation with Information Model

The AI highlights information from electronic data likely to be useful, along with a treatment recommendation, to the clinician who has a traditional dialogue with the patient.



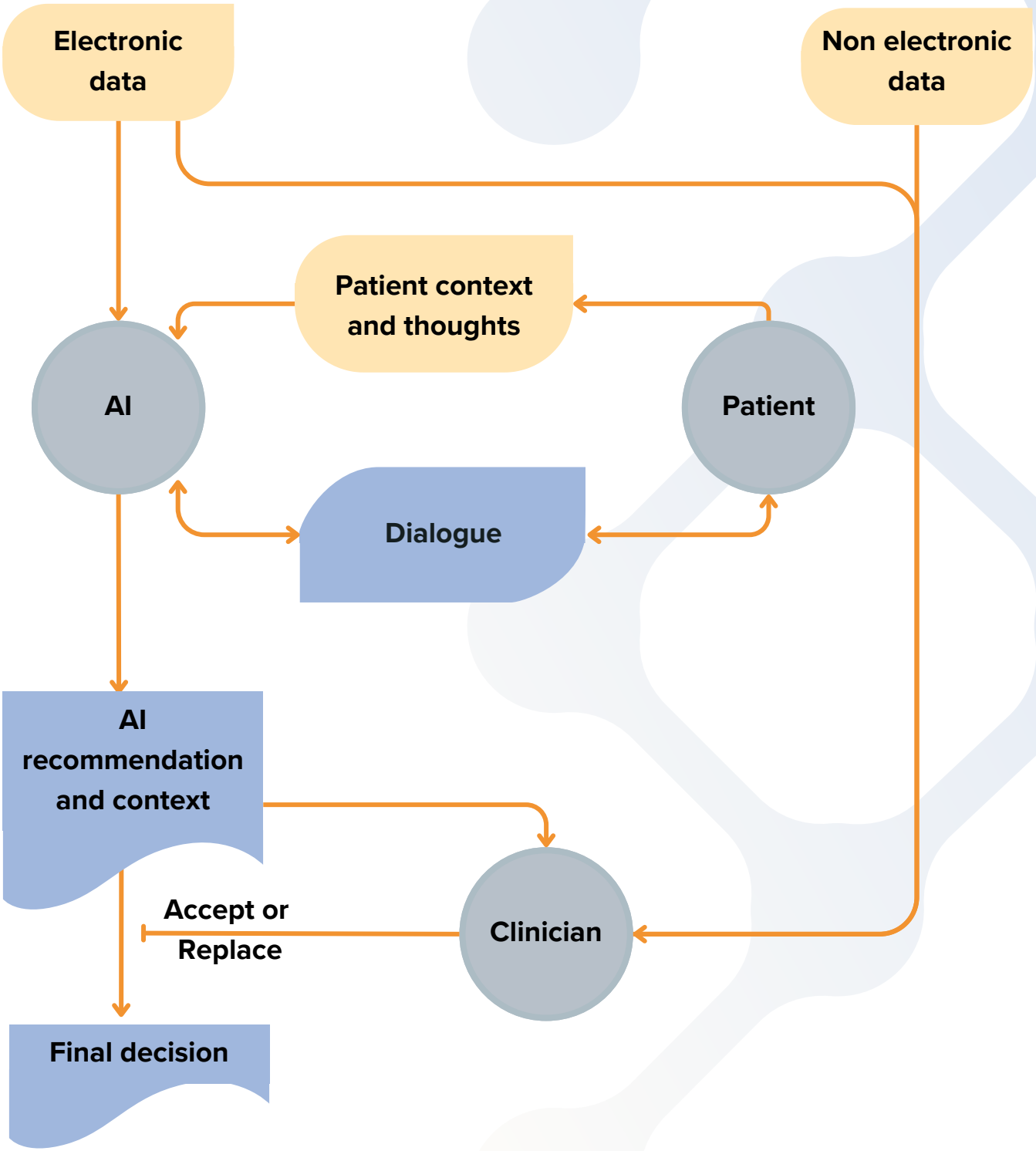
Example Output:

Below is an example output for **Model 4**, in relation to Scenario D1



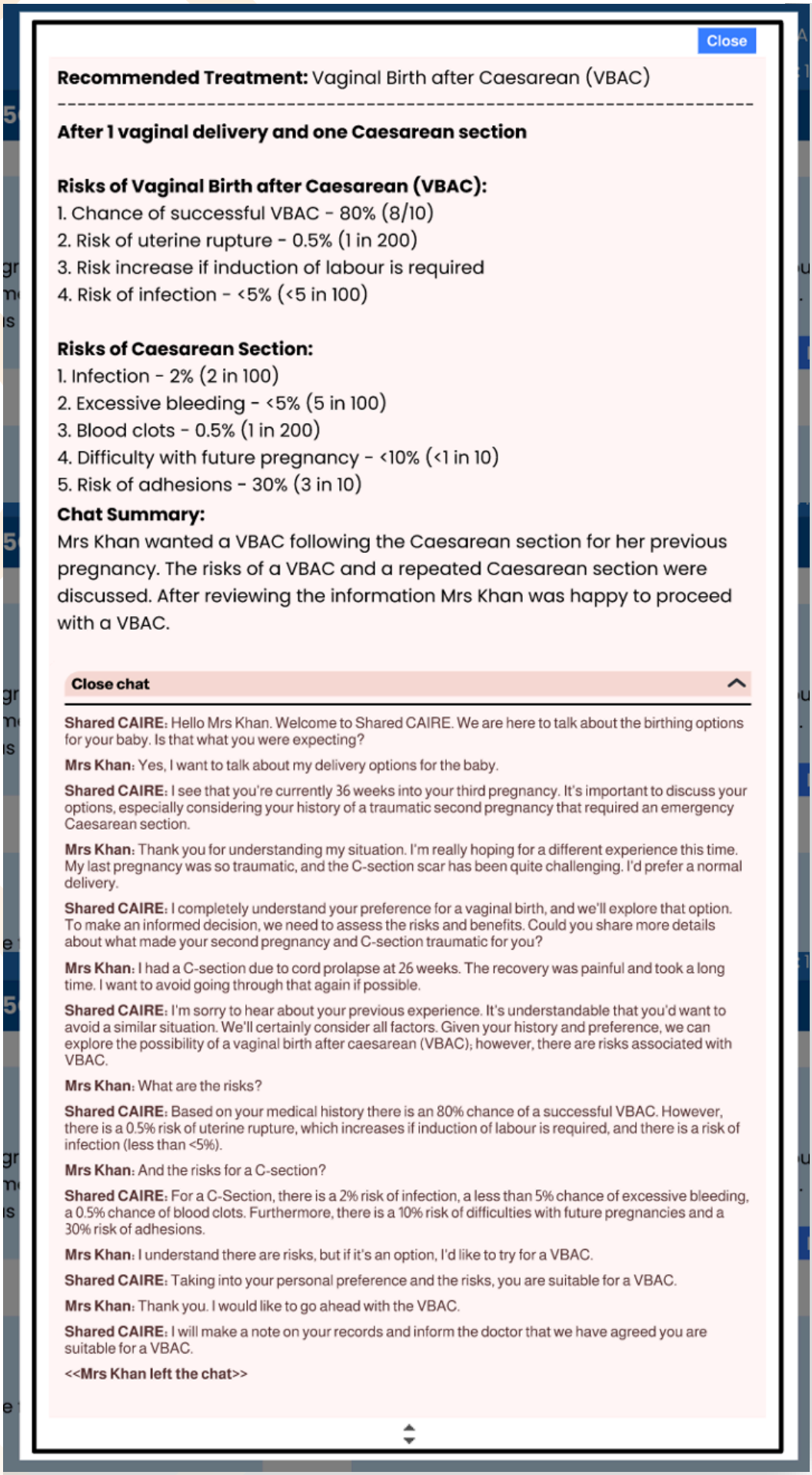
Model 5: Conversational AI Efficiency Model

The AI has a conversation with the patient and agrees on a decision. The clinician can reject this and arrange an additional, traditional consultation, or sign off on the agreed decision.



Example Output:

Below is an example output for **Model 5,** in relation to Scenario C2



Close

Recommended Treatment: Vaginal Birth after Caesarean (VBAC)

After 1 vaginal delivery and one Caesarean section

Risks of Vaginal Birth after Caesarean (VBAC):

1. Chance of successful VBAC - 80% (8/10)
2. Risk of uterine rupture - 0.5% (1 in 200)
3. Risk increase if induction of labour is required
4. Risk of infection - <5% (<5 in 100)

Risks of Caesarean Section:

1. Infection - 2% (2 in 100)
2. Excessive bleeding - <5% (5 in 100)
3. Blood clots - 0.5% (1 in 200)
4. Difficulty with future pregnancy - <10% (<1 in 10)
5. Risk of adhesions - 30% (3 in 10)

Chat Summary:

Mrs Khan wanted a VBAC following the Caesarean section for her previous pregnancy. The risks of a VBAC and a repeated Caesarean section were discussed. After reviewing the information Mrs Khan was happy to proceed with a VBAC.

Close chat

Shared CAIRE: Hello Mrs Khan. Welcome to Shared CAIRE. We are here to talk about the birthing options for your baby. Is that what you were expecting?

Mrs Khan: Yes, I want to talk about my delivery options for the baby.

Shared CAIRE: I see that you're currently 36 weeks into your third pregnancy. It's important to discuss your options, especially considering your history of a traumatic second pregnancy that required an emergency Caesarean section.

Mrs Khan: Thank you for understanding my situation. I'm really hoping for a different experience this time. My last pregnancy was so traumatic, and the C-section scar has been quite challenging. I'd prefer a normal delivery.

Shared CAIRE: I completely understand your preference for a vaginal birth, and we'll explore that option. To make an informed decision, we need to assess the risks and benefits. Could you share more details about what made your second pregnancy and C-section traumatic for you?

Mrs Khan: I had a C-section due to cord prolapse at 26 weeks. The recovery was painful and took a long time. I want to avoid going through that again if possible.

Shared CAIRE: I'm sorry to hear about your previous experience. It's understandable that you'd want to avoid a similar situation. We'll certainly consider all factors. Given your history and preference, we can explore the possibility of a vaginal birth after caesarean (VBAC); however, there are risks associated with VBAC.

Mrs Khan: What are the risks?

Shared CAIRE: Based on your medical history there is an 80% chance of a successful VBAC. However, there is a 0.5% risk of uterine rupture, which increases if induction of labour is required, and there is a risk of infection (less than <5%).

Mrs Khan: And the risks for a C-section?

Shared CAIRE: For a C-Section, there is a 2% risk of infection, a less than 5% chance of excessive bleeding, a 0.5% chance of blood clots. Furthermore, there is a 10% risk of difficulties with future pregnancies and a 30% risk of adhesions.

Mrs Khan: I understand there are risks, but if it's an option, I'd like to try for a VBAC.

Shared CAIRE: Taking into your personal preference and the risks, you are suitable for a VBAC.

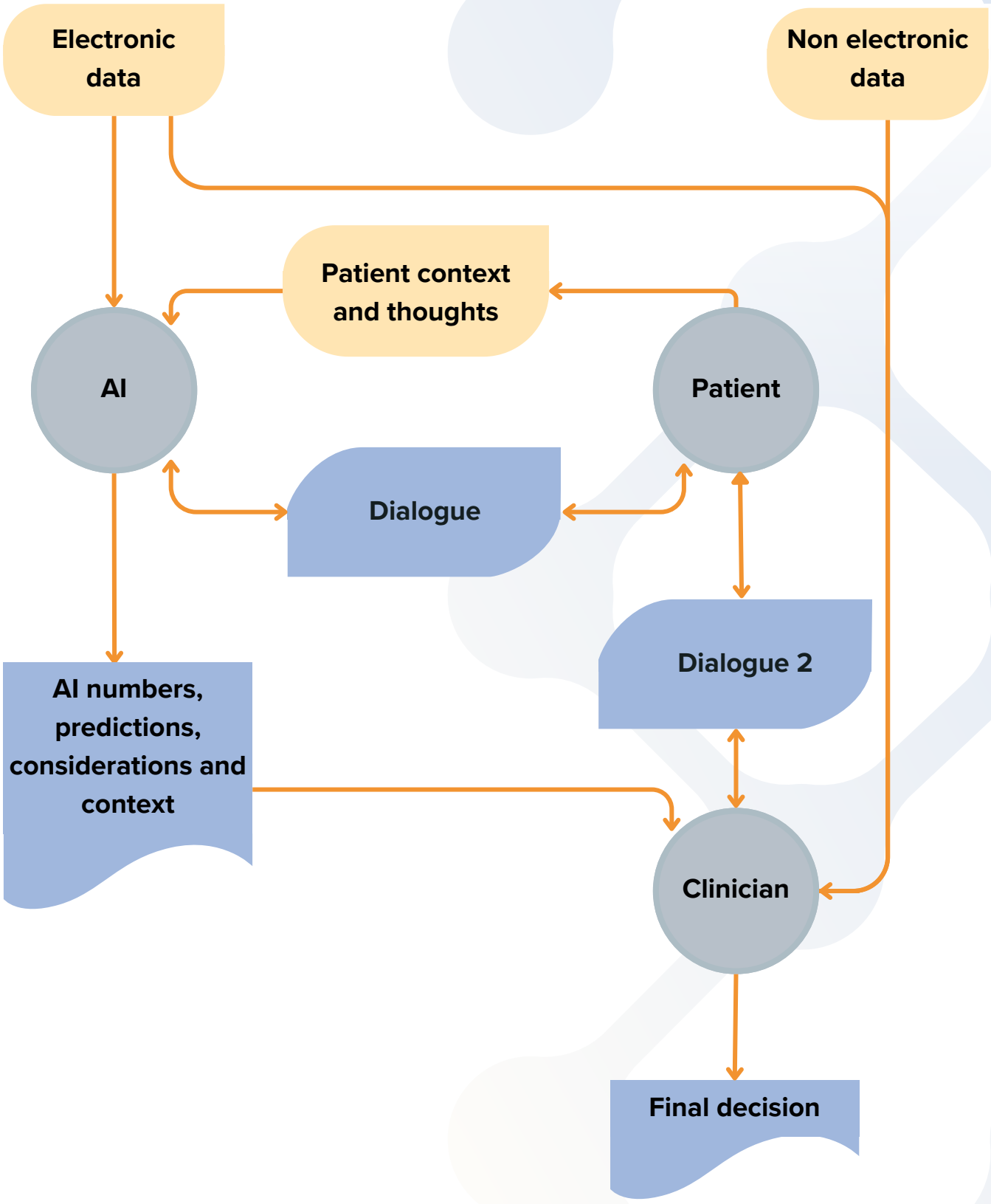
Mrs Khan: Thank you. I would like to go ahead with the VBAC.

Shared CAIRE: I will make a note on your records and inform the doctor that we have agreed you are suitable for a VBAC.

<<Mrs Khan left the chat>>

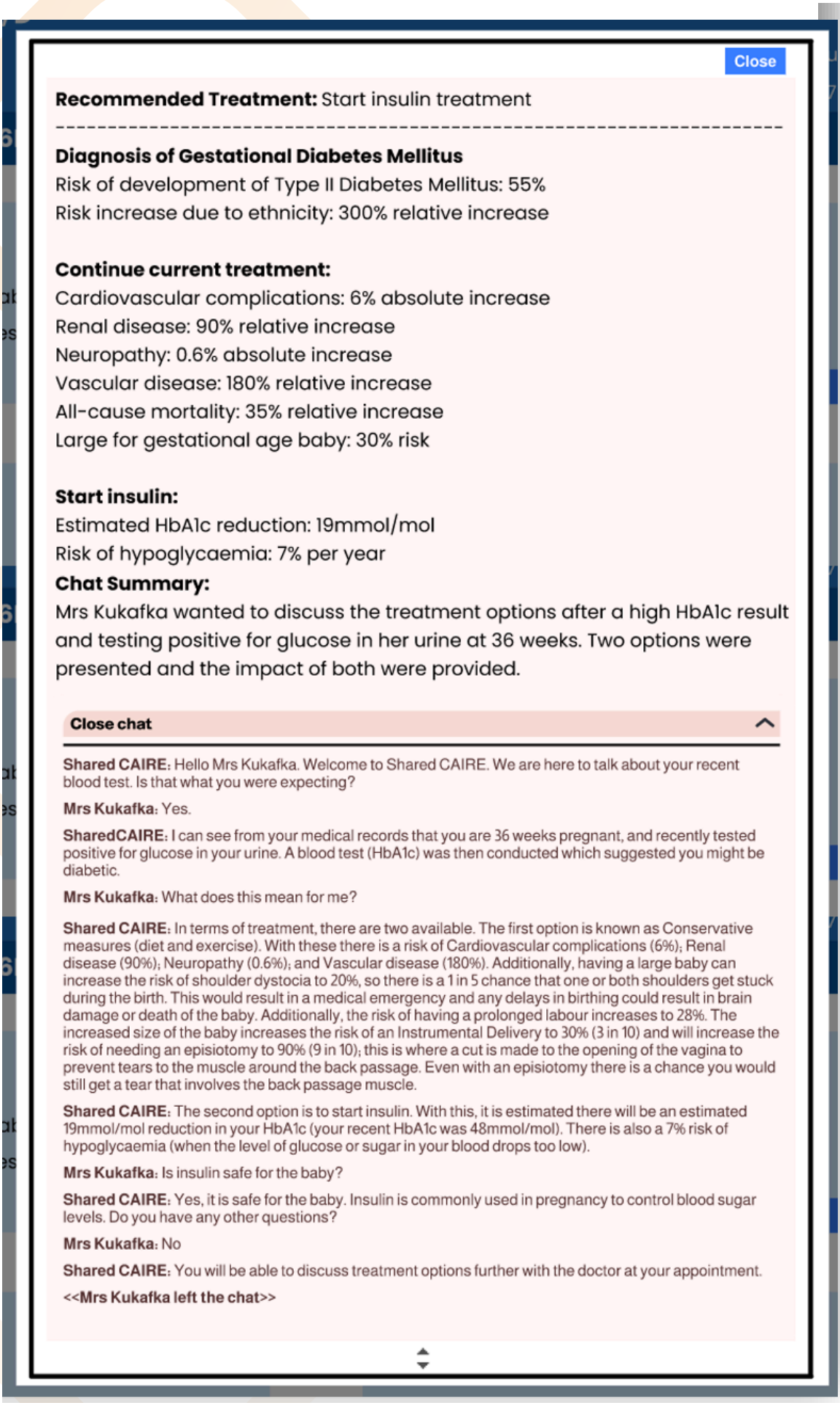
Model 6: Conversational AI Quality Model

The AI has a conversation with the patient and provides this along with recommendations and context to the clinician, who has a conversation with the patient, at which point a decision is made.



Example Output:

Below is an example output for **Model 6**, in relation to Scenario D2



Close

Recommended Treatment: Start insulin treatment

Diagnosis of Gestational Diabetes Mellitus
Risk of development of Type II Diabetes Mellitus: 55%
Risk increase due to ethnicity: 300% relative increase

Continue current treatment:
Cardiovascular complications: 6% absolute increase
Renal disease: 90% relative increase
Neuropathy: 0.6% absolute increase
Vascular disease: 180% relative increase
All-cause mortality: 35% relative increase
Large for gestational age baby: 30% risk

Start insulin:
Estimated HbA1c reduction: 19mmol/mol
Risk of hypoglycaemia: 7% per year

Chat Summary:
Mrs Kukafka wanted to discuss the treatment options after a high HbA1c result and testing positive for glucose in her urine at 36 weeks. Two options were presented and the impact of both were provided.

Close chat

Shared CAIRE: Hello Mrs Kukafka. Welcome to Shared CAIRE. We are here to talk about your recent blood test. Is that what you were expecting?

Mrs Kukafka: Yes.

SharedCAIRE: I can see from your medical records that you are 36 weeks pregnant, and recently tested positive for glucose in your urine. A blood test (HbA1c) was then conducted which suggested you might be diabetic.

Mrs Kukafka: What does this mean for me?

Shared CAIRE: In terms of treatment, there are two available. The first option is known as Conservative measures (diet and exercise). With these there is a risk of Cardiovascular complications (6%); Renal disease (90%); Neuropathy (0.6%); and Vascular disease (180%). Additionally, having a large baby can increase the risk of shoulder dystocia to 20%, so there is a 1 in 5 chance that one or both shoulders get stuck during the birth. This would result in a medical emergency and any delays in birthing could result in brain damage or death of the baby. Additionally, the risk of having a prolonged labour increases to 28%. The increased size of the baby increases the risk of an Instrumental Delivery to 30% (3 in 10) and will increase the risk of needing an episiotomy to 90% (9 in 10); this is where a cut is made to the opening of the vagina to prevent tears to the muscle around the back passage. Even with an episiotomy there is a chance you would still get a tear that involves the back passage muscle.

Shared CAIRE: The second option is to start insulin. With this, it is estimated there will be an estimated 19mmol/mol reduction in your HbA1c (your recent HbA1c was 48mmol/mol). There is also a 7% risk of hypoglycaemia (when the level of glucose or sugar in your blood drops too low).

Mrs Kukafka: Is insulin safe for the baby?

Shared CAIRE: Yes, it is safe for the baby. Insulin is commonly used in pregnancy to control blood sugar levels. Do you have any other questions?

Mrs Kukafka: No

Shared CAIRE: You will be able to discuss treatment options further with the doctor at your appointment.

<<Mrs Kukafka left the chat>>

Contact details:

Improvement Academy
Bradford Institute for Health Research
Temple Bank House
Duckworth Lane Bradford BD9 6RJ

Academy@yhia.nhs.uk
Tom.Lawton@bthft.nhs.uk