UNIVERSITY *of York*

This is a repository copy of *The case for delegated AI autonomy for Human AI teaming in healthcare*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/226594/

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

The case for delegated AI autonomy for Human AI teaming in healthcare

Yan Jia[1,*], Harriet Evans[2,3,+], Zoe Porter[1,+], Simon Graham[4], John McDermid[1], Tom Lawton[1,5], David Snead[2,6], and Ibrahim Habli[1]

[1]Department of Computer Science, University of York, York, UK
[2]Histopathology Department, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK
[3]Warwick Medical School, University of Warwick, Coventry, UK
[4]Histofy Ltd, Coventry, UK
[5]Improvement Academy, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK
[6]Department of Computer Science, University of Warwick, Coventry, UK
[*]yan.jia@york.ac.uk
[+]these authors contributed equally to this work

ABSTRACT

In this paper we propose an advanced approach to integrating artificial intelligence (AI) into healthcare: autonomous decision support. This approach allows the AI algorithm to act autonomously for a subset of patient cases whilst serving a supportive role in other subsets of patient cases based on defined delegation criteria. By leveraging the complementary strengths of both humans and AI, it aims to deliver greater overall performance than existing human-AI teaming models. It ensures safe handling of patient cases and potentially reduces clinician review time, whilst being mindful of AI tool limitations. After setting the approach within the context of current human-AI teaming models, we outline the delegation criteria and apply them to a specific AI-based tool used in histopathology. The potential impact of the approach and the regulatory requirements for its successful implementation are then discussed.

Introduction

Artificial intelligence (AI) applications in healthcare are advancing rapidly and their successful implementation depends on effective human-AI teaming[1]. Currently, the human-AI teaming in healthcare tends to position AI tools in a purely supportive role, so that all clinical decision-making authority remains with human clinicians. There are two main modalities of this approach[2]: sequential clinical workflow, where after initial review by the clinician, the output from the AI-based tool is revealed to the clinician for further consideration before they make a final decision; and concurrent clinical workflow, where the output from the AI-based tool is available from the beginning of the clinician's decision-making process. These two modalities are shown in Figure 1.

The intent is that retaining a human-in-the-loop for all decisions provides patients with a safeguard against potential adverse effects stemming from AI errors and that clinicians will be aided to make better decisions with the support of AI. But there is a paucity of evidence that using AI tools in this way correlates with overall improved performance[3] and it might even reduce the sensitivity of diagnosis[4]. Some evidence[5] shows that providing AI-based decision support can lead to unconscious bias in clinician's decision making processes, e.g. anchoring clinicians to a particular diagnosis or triggering confirmation bias resulting in reduced diagnostic accuracy. This suggests that the current model of keeping AI in a purely supportive role limits the potential of human-AI teaming. Additionally, it restricts the potential efficiency gains of AI in healthcare, since a human is always required in the decision-making process which places demands on stretched healthcare professionals. Given that the types of error made by AI and those made by humans are often different[6](for example, AI is susceptible to artefacts or noise that humans normally would not notice[7], yet it can also detect tiny signs of cancer which are practically invisible to human eyes[8]), the ideal of human-AI teaming is to leverage the complementary strengths of human and AI to achieve better overall performance. Against this backdrop, we propose a new approach to human-AI teaming, which we call autonomous decision support. This approach reflects recent developments in this field, but also takes a step further by clarifying the clinical workflow associated with it and articulating an extended set of delegation criteria, making it more practical and comprehensive.

(a) Sequential clinical workflow


(b) Concurrent clinical workflow
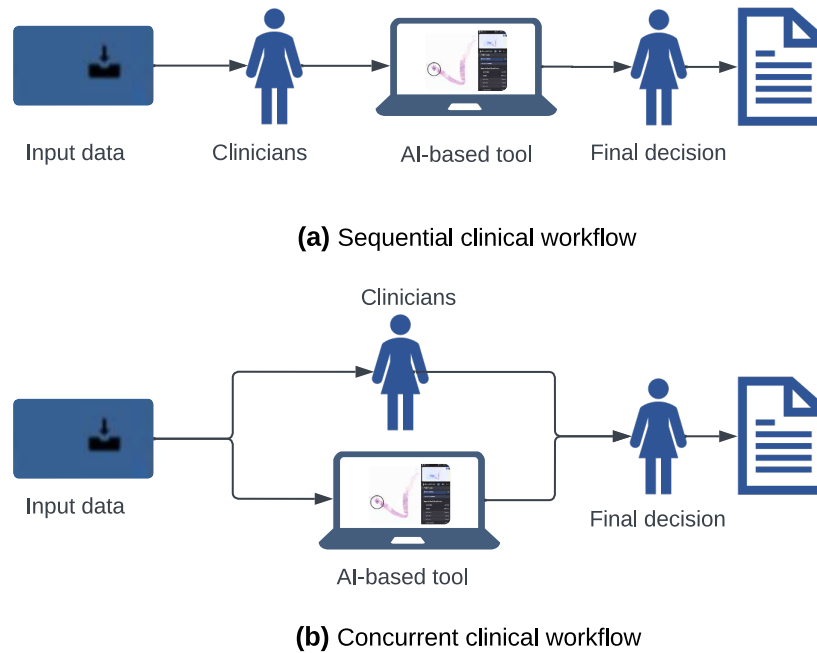
Figure 1. Current Human-AI teaming modalities. This figure shows two clinical workflow models for AI-based decision support tools which represents the current approach to integrating AI into clinical decision making processes. The panel (a) represents a sequential clinical workflow, where input data is first processed by clinicians and then passed to the AI-based tool in a linear, step-by-step manner before arriving at the final decision made by the clinician. The panel (b) shows a concurrent clinical workflow, where clinicians and the AI-based tool simultaneously receive and process input data, then the clinician makes the final decision.

Autonomous decision support

Recently, alternative approaches to the sequential and concurrent clinical workflow modalities in human-AI teaming in healthcare have been proposed. These proposals[9–12] have shown that using AI to autonomously handle a subset of patient cases, guided by the AI's confidence score in its prediction or recommendation, can improve overall diagnostic performance, e.g. improvements of 2.6% in sensitivity and 1.0% in specificity have been achieved compared with unaided radiologists[10]. This is particularly significant given that there is usually a trade-off between these desiderata, with increased sensitivity coming at the expense of decreased specificity, and vice versa[13]. However, there are limitations in these approaches. For example, Dvijotham et al[11] have proposed a complementarity-driven deferral to clinical workflow (CoDoc) approach, which either relies solely on AI predictions if the AI is confident or defers entirely to human decisions when the AI is not confident. But this overlooks the fact that not all decisions carry the same criticality, therefore some decisions still require human intervention even if the AI is confident in its predictions. Smith et al[12] have proposed to use AI algorithms autonomously reporting the normal chest radiographs that the AI predicted and has high confidence in, i.e. 'high confidence normal'(HCN), while directing the remaining cases to the normal clinical workflow in the hospital. Compared to the CoDoc approach, this approach suggested that only the normal radiographs should be handled autonomously. However, it fails to leverage the value of AI in the abnormal cases, where it might identify clinically significant lesions that might be missed by humans. Similarly, Leibig et al[10] have proposed a decision-referral approach for autonomously reporting the HCN, and refer the rest to the normal clinical workflow without indicating the AI predictions - with the addition that a warning is displayed if clinicians miss a lesion that the AI is confident in. This approach is the most similar to our proposal, however it fails to recognise that other important factors, beyond AI confidence scores and normal or abnormal predictions, should be considered when directing patient cases. The clinical context is fundamental to all medical decision making, including interpretive tasks required in radiology and cellular pathology and so should be part of determining AI tool deployment, in order to ensure the maximum benefit and minimal risk.

Therefore, we propose an autonomous decision support approach, as shown in Figure 2 where the patient and case data is utilised to determine the most appropriate usage for the AI-based tool. Rather than automatically giving the AI prediction, delegation criteria are utilised to direct the case into one of three distinct pathways:

1. AI only: the AI prediction is taken as the final decision without human intervention;
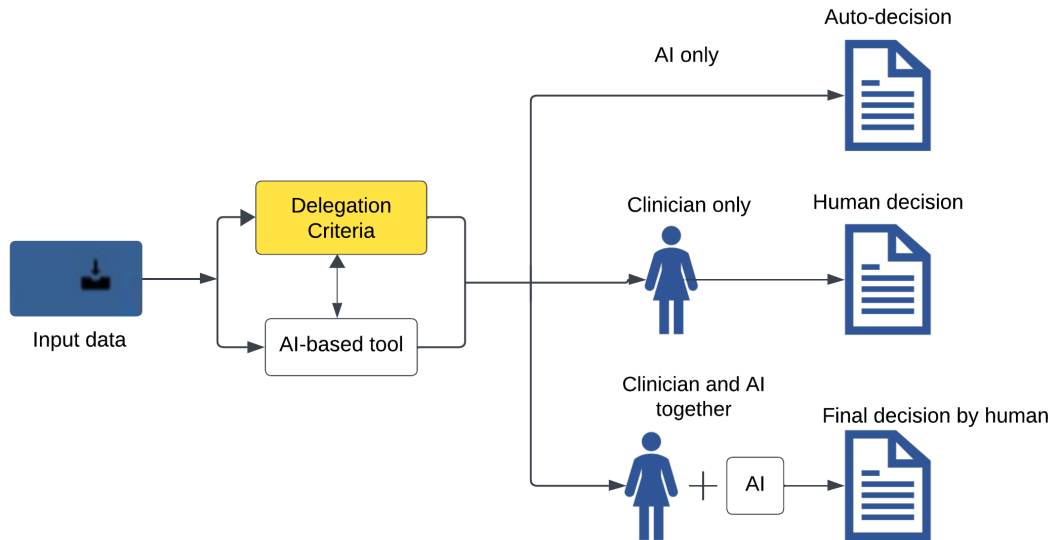
Figure 2. The clinical workflow for an autonomous decision support tool. This figure shows our approach for integrating AI in healthcare, where it has a delegation criteria, which interacts with the AI-based tool, to direct the input patient data to three distinctive pathways, "AI only", "Clinician only", and "Clinician and AI together". "AI only" pathway represents the "autonomous" element in our approach and "Clinician and AI together" pathway represents the "support" element in our approach, hence the name "autonomous decision support" approach.

2. Clinician only: the AI does not make a prediction and refers the case to a human;

3. Clinician and AI together: the AI prediction is revealed to support the clinician, who then makes the final decision, akin to existing human-AI teaming modalities (see Figure 1).

Compared to existing human-AI teaming models in Figure 1, our approach enhances the third pathway by informing the clinician of the AI's prediction rationale, thereby providing an efficient transfer of information which the clinician can accept, reject or modify as they deem appropriate. For instance, instead of displaying AI predictions for all patient cases, the system could only present confident abnormal predictions to the clinician, prompting a more careful review. Overall, our autonomous decision support approach leverages the benefits of AI in both supportive and autonomous roles, enhancing overall decision-making and patient safety. It ensures that automation is only delivered for cases for which it is safe to rely on AI (e.g. high confidence recognition of normal cases) and flags cases for clinician review when the prediction carries more criticality, e.g. abnormal predictions, even when the AI is confident. This allows safe handling of the patient cases whilst allowing clinicians to potentially reduce review time and be mindful of AI limitations. Additionally, our approach also allows certain cases to be handled solely by clinicians when the input is beyond the AI's capability, e.g. out of scope cases, or when the AI is less reliable, e.g. if a patient's clinical context indicates AI prediction of normal is not to be relied upon.

The design of the delegation criteria, which directs patient cases into these three distinct pathways, is the core of the autonomous decision support approach. In the work cited above, the delegation criteria mainly rely on the utilisation of AI confidence scores and its predictions. In this paper, we present a more comprehensive set of factors to inform the delegation criteria, including the *types of tasks*, *clinical context*, *criticality of the decision*, *AI confidence score*, and the *failure modes of both humans and AI*, as shown in Table 1. The factors presented here are intended to be general, so that they can be applied widely. Therefore, it is necessary to operationalise the delegation criteria for a specific application. This necessitates collaboration within a multidisciplinary team involving AI developers and clinicians, who between them have the respective expertise to support the development and implementation of the delegation criteria. This multidisciplinary collaboration is becoming more widely recognised with medical devices companies employing clinicians working alongside AI developers, but is not yet common.

Delegation Criteria
In table 1, we have shown the detailed descriptions for each factor, accompanied by a set of design questions tailored for AI tool developers, clinical team and other relevant experts to discuss in order to facilitate the design of the delegation criteria. Overall, there are three important steps to consider when developing delegation criteria for the AI tool.

Factors	Description	Design questions
Types of tasks	The type of task that the AI is going to be applied to, e.g. image classification, treatment recommendation, etc.	Is the task to which the AI will be applied suitable for implementing the delegation criteria? If so, define the exact type of input data, the output of the AI tool and the patient group for which the AI tool will be used.
Clinical context	Factors encompassing the patient's medical history, current symptoms, results of other investigations and treatment history.	What clinical context is the AI tool going to be used in including the current clinical pathway? Are there any explicit factors in the clinical context that cannot be accommodated by the AI? If so, consider incorporating these factors into the delegation criteria.
Criticality of the decision	The importance or significance of a decision in a given clinical context.	How much impact does each output of the AI tool have on patient care? If an output of the AI tool carries greater significance, then involving human would be necessary.
AI confidence score	Quantification of the uncertainty in an AI output	Can a confidence score be utilised to indicate the reliability of the AI prediction, e.g. through a validation study? If so, consider combining the use of the confidence score and AI prediction in the delegation criteria.
Failure modes of AI	The scenarios where the AI often gives erroneous output even when the input is within scope	Are there any known failure modes of the AI tool, e.g. through validation study, which are not easy to resolve by further algorithm training? If so, consider using the failure modes of AI to direct these types of patient cases to humans.
Failure modes of humans	The scenarios where humans often makes erroneous judgement	Are there any known failure modes of humans in this clinical context? If so, consider using the AI tool to support the human in these types of cases.

Table 1. Factors that influence the design of Delegation Criteria

1. The first crucial step is to evaluate the ***type of task***. This initial assessment determines whether a clinical task is suitable for implementing the delegation criteria or not. Tasks with less context dependency, such as radiology and pathology, which primarily involve image analysis and laboratory results, tend to be more suitable to implement the delegation criteria than tasks in an ICU setting, where patient conditions can be highly variable and require real time adjustments.

2. If the task is determined to be suitable for implementing the delegation criteria, then it is important to clearly delineate the ***type of task*** an AI tool is performing to ensure correct use and safety. This involves precise specification of: (i) the exact type of input data the AI system can process, for example, within pathology this involves defining the anatomical site, type of specimen, type of tissue stain. (ii) what output the AI system can produce, e.g. binary output or multi-class classification (iii) which patient group for which the AI tool can safely operate, e.g. adult. The precise definition of the type of input data and the task to be performed ensures that wherever possible AI is utilised in decision making, but out of scope scenarios will be identified for human intervention.

3. The third step involves considering the other factors, i.e. ***Clinical context, Criticality of the decision, AI confidence score, Failure modes of AI and human*** to further define how the patient cases are routed to the three pathway outlined in Figure 2. These five factors are not hierarchical; instead, they should be considered simultaneously to weave into the design for the delegation criteria. Further, each factor may need to revisited multiple times during the process. In the following, we illustrate these five factors in detail.

   Clinical context is crucial when making decisions in healthcare. It encompasses the patient's medical history, current symptoms, results of other investigations and treatment history. This contextual view allows clinicians to make informed decisions that are tailored to the individual patient. For instance, when a pathologist or radiologist is reviewing a patient's specimen or imaging, they consider whether their diagnosis would match the clinical context, or if the patient is more at risk of certain conditions due to their medical history. Despite advances in AI technology, current systems lack the ability to integrate and interpret this nuanced clinical context. Therefore, consideration of the clinical context in which the AI tool is used should be reflected in the design of the delegation criteria. Clinical context not only helps to determine whether a patient case is suitable for passing to the AI tool, but also provides insight into the reliability of the AI predictions.

When considering the **criticality of decisions**, it is important to recognise that even within the same task, different outputs can have vastly different patient implications. For instance, when examining colonic (large bowel) biopsies, abnormalities cover an enormous range of conditions from minor changes requiring no further intervention through to life threatening disease such as cancer. This range of likely patient impact illustrates the need for careful consideration of decision criticality in designing delegation criteria. Ideally, the most critical decisions should combine AI and clinician perspectives.

For the **AI confidence score**, it is important to quantify uncertainty from both known and unknown data distributions (also known as out-of-distribution where the input data is from a significantly different distribution compared to the training data). Ideally, an AI confidence score should express high uncertainty in cases of incorrect predictions and out-of-distribution inputs. However, in reality, neural networks are known to produce overly confident predictions[14, 15]. There are various methods that can be employed to calculate a confidence score, ranging from basic techniques such as utilising probabilities from softmax distributions to more advanced approaches such as Bayesian methods[16, 17] or ensemble methods[18]. Despite extensive research, there is no consensus on the optimal method for determining confidence scores[19]. Studies[20] have shown that while the AI tends to produce over confident predictions, correctly classified examples tend to exhibit higher maximum softmax probabilities compared to erroneously classified samples and out-of-distribution samples, which could serve as a baseline for confidence scoring. Nonetheless, solely relying on the AI confidence in the delegation criteria can be problematic as these scores may not accurately reflect the real world conditions. Thus, it's crucial to consider additional factors to improve the reliability of the delegation criteria.

In terms of **failure modes of AI**, more studies are beginning to investigate the nature of errors made by AI systems. For instance, recent studies on a prostate cancer screening tool reported their false negative errors, highlighting that the tool struggled with variants of adenocarcinoma such as foamy gland carcinoma, adenocarcinoma with hormone deprivation therapy effects, adenocarcinoma with an atrophic appearance, low grade adenocarcinoma or isolated foci of perineural invasion[21–23]. This knowledge of AI error patterns can guide the design of the delegation criteria (if not resolved by further algorithm training), prompting the delegation of cases prone to AI errors to human clinicians. In order to do this, there should be emphasis on reporting AI errors in clinically relevant terms rather than solely technical metrics[24], e.g. false negatives or positives.

Categorisation of the **failure modes of humans** may be difficult due to limited published work and the inherent variability and complexity involved in categorising human errors. Nonetheless, where available, studies should be utilised to provide insight into existing human errors. For instance, in the examination of colon biopsies, one study[25] identified that the two most frequent areas prone to errors were the identification of the type of serrated polyp and the grading of dysplasia within a polyp. By pinpointing these specific areas of vulnerability, AI predictions can potentially complement human expertise and ultimately improve diagnostic accuracy and patient care.

Identifying these factors represents the initial step in designing the delegation criteria. Implementing these criteria requires careful consideration both in the design of the AI tool (such as ensuring the tool can filter out data that is not part of its defined task) and how it is utilised in practice. This requires definition of the intended use case for which the healthcare system allows a tool to operate and how data flow is managed in the pathway (such as ensuring that cases with a low AI tool confidence are automatically sent to a clinician to review). Importantly, the delegation criteria must evolve in tandem with the AI system itself; any updates to the AI system should prompt a review and corresponding update of the delegation criteria. In the following section, we illustrate how our approach can be applied to an AI tool in histopathology.

Applying autonomous decision support in histopathology: COBIx

The colon can be affected by a wide range of inflammatory and neoplastic diseases, diagnosed primarily by endoscopic examination and histological sampling. The increased volume of endoscopic biopsies, as well as an increasing complexity of reporting has led to an extra demand on already struggling pathology services. Currently, large bowel biopsies account for 10% of all tissue sample requests in laboratories with approximately 30-40% of these samples being normal. Therefore, implementing an autonomous decision support approach that can automatically report some normal biopsies and support pathologists with other cases would be highly beneficial. To this end, an AI-based tool has been developed to assist in the analysis of Colon and Rectal Endoscopic Biopsy (COBIx). Details on the development of the COBIx algorithm can be found in this paper[26]. In the following, we illustrate how application of the delegation criteria could determine the most appropriate pathway for COBIx use: AI only, pathologist only, and pathologist and AI (matching Figure 2).

Figure 3 starts with the **type of task**. COBIx has been trained to classify haematoxylin and eosin (H&E) histopathology whole slide images (WSI) from colon and rectal biopsy for adult patients, which defines the correct WSI to be processed by COBIx.

The COBIx assessment of colonic WSI is a multi-step process. First, a quality control step ensures the slides are suitable to be processed by the algorithm. It excludes samples containing non-colonic tissue, e.g. adjacent small bowel (ileal) or squamous (anal) mucosa in the gastrointestinal tract, and those with inadequate colonic tissue to make a diagnosis. Further, poor quality
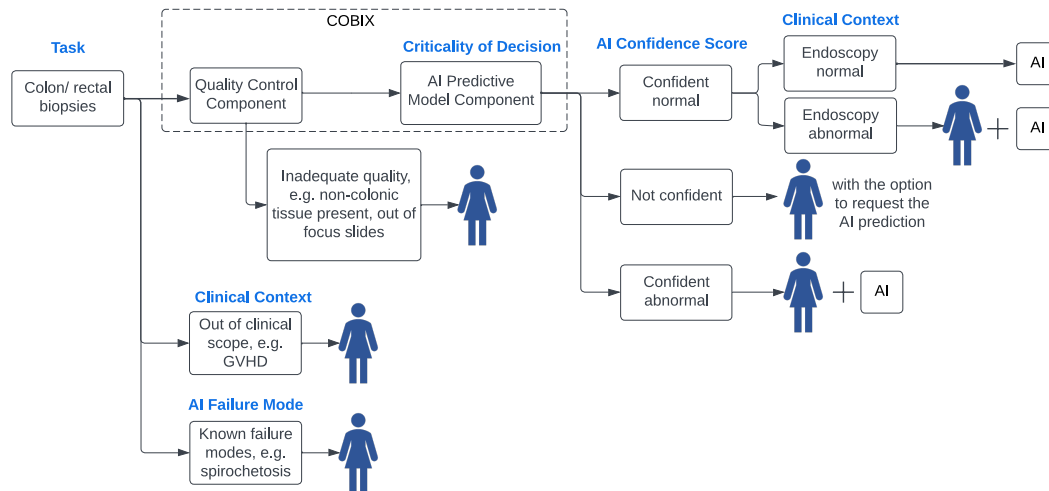
Figure 3. The potential clinical workflow for using COBIx as an autonomous decision support tool. COBIx is a published multi-class AI-based tool for analysing colon and rectal (large bowel) endoscopic biopsies. It produces five possible outputs with one normal category and four abnormal categories, i.e. Neoplastic urgent, Neoplastic non-urgent, Non-neoplastic urgent, Non-neoplastic non-urgent. This workflow shows how Clinical context, AI failure mode, Criticality of decision, and AI confidence score can be incorporated into designing the delegation criteria to safely direct the biopsies into three pathways: "AI only", "Clinician only" and "Clinician and AI together".

WSIs, such as those that are out of focus or folded, where quality hinders the ability of COBIx to perform its task, will be excluded and directed to a pathologist.

Next, a segmentation step, trained on large annotated datasets, detects various important structural and cellular regions in the WSI, such as glands, lumen and different nuclei. This segmentation facilitates the identification of histological features used by human pathologists to recognise disease or normality. Then, these histological features identified on a WSI are fed into a neural network to predict the classification. COBIx classifies biopsies into normal or one of four abnormal categories based on the features present, i.e. Normal, Neoplastic urgent, Neoplastic non-urgent, Non-neoplastic urgent, Non-neoplastic non-urgent, thus **Criticality of decision** is built into COBIx. It is vital that malignant conditions such as adenocarcinoma, and pre-malignant entities such as high-grade dysplasia, are not missed. Therefore even a small area of these on a WSI is classified as abnormal and must be seen by a pathologist. This use of harsh cut-off criteria for certain high-risk features helps ensure critical decisions are made by pathologists.

Further, COBIx's output needs to be calibrated, so that the output score accurately reflects the **model's confidence**. For example, a score of 0.5 suggests the model is correct 50% of the time, indicating low confidence, while a score of 0.9 indicates the model is correct 90% of the time, reflecting high confidence. To reduce the likelihood of incorrectly screened cases, we can set a threshold to consider only slides with confident predictions. This threshold can be determined by analysing the reduction in the screening error rate. For confident normal cases, an automated report could potentially be issued, without pathologist review, but confident abnormal cases would be reviewed by a pathologist with the support from AI (with the categorisation of different abnormal cases by COBIx allowing cases to be triaged based on urgency). Not confident cases would be reviewed by pathologists with the option to request the AI prediction.

The **clinical context** is considered at multiple stages in the COBIx processing pipeline. First, despite extensive training, there will be entities COBIx has not been trained to detect due to the rarity in the population. In such scenarios it is inappropriate to use COBIx and instead a pathologist should make the diagnosis. One such example is graft versus host disease (GVHD), a condition only seen following allogenic stem cell transplant or occasionally following solid organ transplant, so knowledge of the patient's medical history is vital. A patient history of transplantations and/or a clinical suspicion of this entity can thus ensure this specimen is reviewed by a pathologist.

A second, and more fundamental, way in which the **clinical context** is relevant is the endoscopic information. All colon biopsies are taken during endoscopy, during which the endoscopist can visualise abnormalities in the colon, providing highly relevant insight into the likely disease process. This endoscopic information forms part of a key delegation step in the use of COBIx. If COBIx classifies a specimen as confident normal, and the endoscopy is also normal, this provides strong support that the colon is actually normal and so AI alone can make this diagnosis. However, if COBIx reports the biopsy as confident normal but the endoscopy report is abnormal, this contradiction prompts suspicion that the COBIx classification could be

incorrect therefore a pathologist must review the case. The use of endoscopic information specifically for this filtration of slides classified as confident normal by COBIx also reflects **criticality of decision**, because false negative errors - potentially resulting in disease progression and morbidity - would have the greatest clinical significance for patients, making it critical to minimise false negatives.

Knowledge of COBIx **AI failure modes** could further enable tailored use of the tool. For example, if COBIx struggled to detect an entity such as spirochetosis (which is associated with very little other changes in the colonic tissue so would be otherwise undetected), then clinical suspicion of this entity can direct the case towards a pathologist.

Human failure modes are less relevant as delegation criteria in the current proposed use of COBIx because the scenarios where pathologists struggle are variable and poorly defined. Additionally, COBIx may encounter similar challenges in identifying these cases.

This illustrates how delegation criteria can help ensure the safe and efficient use of an autonomous decision support tool. They enable the use of COBIx for independent diagnosis when appropriate, reducing pathologist workload and minimising delays in reporting. At the same time, they ensure that colon biopsies with abnormalities or out of clinical scope, or in patients with endoscopic abnormalities, are safely reviewed by pathologists.

Potential impact of the autonomous decision support approach

Rather than reduce the burden on an already overworked clinician, the purely supportive approaches place an additional task on clinicians, requiring them to act as a sense check on the AI at all times. By contrast, our autonomous decision support approach only imposes this additional task on the clinician in a subset of cases, in accordance with the delegation criteria. Therefore, if implemented, it will allow clinicians to shift their focus to complex cases, meaning all patients receive their results sooner, alleviating any stress from awaiting a diagnosis and allowing quicker initiation of treatment where required.

This holds great promise for a safer and more efficient healthcare system. Even so, careful evaluation is required. The cases which are autonomously reported by the AI are likely to be those which are normally dealt with most quickly by clinicians. Reporting a reduction in the proportion of cases needing review as if it were an equal reduction in workload is misleading, as the more complex cases left behind are likely to be the ones taking the majority of the clinicians' time[27]. Further, we should also consider the impact of autonomous decision support on clinicians over time. Clinicians traditionally develop their expertise starting from simpler cases, so if this approach becomes established, there may be too few simple or normal cases from which to learn and develop their judgement. This is likely both to affect their perception of what is normal and alter their beliefs about the base-rate of normality in the population, with the possible consequence of over-diagnosis. Ensuring that clinician workload, training and understanding is not adversely impacted by the autonomous decision support approach should be at the forefront of its wider evaluation.

An ethical concern arises when an AI tool is in a purely supportive role (whether in the sequential or concurrent clinical workflow model) and where the clinician is expected to assess the validity or appropriateness of the AI's output in order to make the final decision. Thus, there is a concern that clinicians may end up being unfairly held liable for acting on an AI prediction or implementing an AI recommendation that turns out to be wrong. This has been referred to as clinicians being 'liability sinks' for AI[28]. The term 'liability sink' refers to when one frontline individual (in this context, the clinician) absorbs all of the liability for a harmful outcome in which an AI system was a major cause. The problem is not clinicians being held liable, but that clinicians are being held liable when they have insufficient understanding of how the AI tool reaches its output to be able to evaluate those outputs effectively.

The autonomous decision support approach proposed here helps to address this liability sink issue because the clinician is no longer placed in a position of acting as a safeguard on the machine in all cases. In the first pathway, referred to as 'AI only' in Figure 2, the cases are autonomously processed. By removing the clinician as intermediary, this pathway could shield the frontline clinician from being a liability sink. However, it should also be recognised that clinicians in the medical device company, auditing the patient cases reported by the AI, could face the liability sink problem. In the second pathway, referred to as 'clinician only', the clinician decides without any AI recommendation at all, which is the same as the traditional practice. As such, they would not be absorbing liability for faults or errors of the AI tool. And in the third pathway, referred to as 'clinician and AI together', which is the same as the existing human-AI teaming modality, but the clinician at least has more information about the AI tool's output, e.g. only confident abnormal predictions from the AI are shown to the clinician prompting a more careful review, therefore the clinician is able to make more informed decisions, which should limit the degree to which they absorb liability unfairly due to a lack of understanding. However, there is still a need to evaluate and update how liability should be distributed across the wider healthcare system[28, 29].

Establishing new regulatory requirements

Autonomous decision support marks an advance in the design and use of AI systems that healthcare regulators, e.g. the U.S. Food and Drug Administration (FDA)[30], currently approve and new regulatory requirements need to be established to unlock its benefits. Currently approved AI/ML based medical devices are intended to be used in a purely supportive role, either in a sequential or concurrent clinical workflow. No devices have been approved for autonomous use and there is a need to establish new regulatory requirements specifically for AI tools intended for such use. Further, the design of the delegation criteria, which requires a multidisciplinary team to cover the different areas of expertise, might be possible within the medical device company in the development phases, but it is more likely to require collaboration between the medical device company and hospitals that use the AI tool, during the deployment phases. Therefore any new regulations need to allow such flexibility when clearing the medical devices. This potentially can be facilitated by the recent draft guidance from the FDA for pre-determined change control plans for medical devices[31], which describes what modifications can be made to medical devices and how they can be assessed after approval.Further, this will also likely affect both technical and clinical validity.

In terms of technical validity, one key issue is to evaluate the calibration between AI confidence scores and the accuracy of the AI model. This calibration should demonstrate that as AI confidence increases, the accuracy of predictions also improves, ideally showcasing a linear relationship between confidence level and prediction accuracy, ensuring a consistent and proportional increase in accuracy with higher confidence score.

In terms of clinical validity, while prospective and retrospective studies are commonly used, evaluating the longer-term impact of this approach necessitates a novel methodology. This involves finding ways to clinically follow up patients, ensuring that the approach does not systematically disadvantage any particular demographic group, while preserving patient privacy. Further, there also needs to be ways to evaluate the impact on clinicians learning ability without turning this into unconstructive criticism of individual clinician performance. By addressing both technical and clinical validity through comprehensive evaluation, we can foster greater confidence in this new approach to integrating AI-driven technologies in healthcare systems.

Conclusion

Introducing AI in healthcare necessitates investment in training and infrastructure, but our autonomous decision support approach can minimise extra demands by leveraging available clinical information for the design of the delegation criteria. Further, the transparency and adaptability of the delegation criteria would be necessary to support the adoption and regulatory approval of this approach. Finally, our approach further emphasises the importance of a multidisciplinary team in AI tool development, including clinicians, AI developers, ethicists and other relevant stakeholders.

Acknowledgement

Author contributions statement

Y.J. conceived the idea and contributed to the design and structure of the manuscript. J.M., T.L., and I.H contributed to the conception of the idea. Z.P. contributed to the analysis of the potential impact of the approach. H.E., S.G., and D.S. contributed to the histopathology COBIx example. All authors contributed to the writing and reviewing of the manuscript.

Competing interest

Prof Tom Lawton reports he is the Head of clinical AI for Bradford Teaching Hospitals NHS Foundation Trust. Prof David Snead reports he is co-founder, shareholder and director of Histofy Ltd and has received honoraria from Astra Zeneca and Oliver Wyman LLC. Simon Graham reports he is co-founder and shareholder of Histofy Ltd. Harriet Evans reports working ad hoc sessions for Histofy Ltd. The remaining authors declare no competing interests.

References

1. Henry, K. E. *et al.* Human–machine teaming is key to ai adoption: clinicians' experiences with a deployed machine learning system. *NPJ digital medicine* **5**, 97 (2022).

2. Barinov, L. *et al.* Impact of data presentation on physician performance utilizing artificial intelligence-based computer-aided diagnosis and decision support systems. *J. Digit. Imaging* **32**, 408–416 (2019).

3. Vasey, B. *et al.* Association of clinician diagnostic performance with machine learning–based decision support systems: a systematic review. *JAMA network open* **4**, e211276–e211276 (2021).

4. Lehman, C. D. *et al.* Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine* **175**, 1828–1837 (2015).

5. Gaube, S. *et al.* Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* **4**, 31 (2021).

6. Ruamviboonsuk, P. *et al.* Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ digital medicine* **2**, 25 (2019).

7. Goodfellow, I. J. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

8. Kleinman, Z. NHS AI test spots tiny cancers missed by doctors. https://www.bbc.co.uk/news/technology-68607059 (2024). Accessed: 2024-09-17.

9. Dyer, T. *et al.* Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm. *Clin. radiology* **76**, 473–e9 (2021).

10. Leibig, C. *et al.* Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis. *The Lancet Digit. Heal.* **4**, e507–e519 (2022).

11. Dvijotham, K. *et al.* Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nat. Medicine* **29**, 1814–1820 (2023).

12. Smith, J., Morgan, T. N., Williams, P., Malik, Q. & Rasalingham, S. Real-world performance of autonomously reporting normal chest radiographs in nhs trusts using a deep-learning algorithm on the gp pathway. *arXiv preprint arXiv:2306.16115* (2023).

13. Florkowski, C. M. Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clin. Biochem. Rev.* **29**, S83 (2008).

14. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436 (2015).

15. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. neural information processing systems* **30** (2017).

16. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural network. In *International conference on machine learning*, 1613–1622 (PMLR, 2015).

17. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059 (PMLR, 2016).

18. Kompa, B., Snoek, J. & Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit. Medicine* **4**, 4 (2021).

19. Gawlikowski, J. *et al.* A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589 (2023).

20. Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *In Proceedings of International Conference on Learning Representations (ICLR)* (2017).

21. da Silva, L. M. *et al.* Independent real-world application of a clinical-grade automated prostate cancer detection system. *The J. pathology* **254**, 147–158 (2021).

22. Perincheri, S. *et al.* An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod. Pathol.* **34**, 1588–1595 (2021).

23. Raciti, P. *et al.* Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod. Pathol.* **33**, 2058–2066 (2020).

24. Evans, H. & Snead, D. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *NPJ Digit. Medicine* **7**, 89 (2024).

25. Azam, A. S. *et al.* Digital pathology for reporting histopathology samples, including cancer screening samples–definitive evidence from a multisite study. *Histopathology* **84**, 847–862 (2024).

26. Graham, S. *et al.* Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study. *Gut* **72**, 1709–1721 (2023).

27. Park, C. M. AI: Workload Reduction by Autonomous Reporting of Normal Chest Radiographs (2023).

28. Lawton, T. *et al.* Clinicians risk becoming 'liability sinks' for artificial intelligence. *Futur. Healthc. J.* **11** (2024).

29. Smith, H. & Fotheringham, K. Artificial intelligence in clinical decision-making: rethinking liability. *Med. Law Int.* **20**, 131–154 (2020).

30. U.S. Food & Drug. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (2024). Accessed: 2024-09-17.

31. U.S. Food & Drug. Predetermined Change Control Plans for Medical Devices. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-medical-devicess (2024). Accessed: 2025-01-13.