Computers in Biology and Medicine 192 (2025) 110237

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A deployment safety case for AI-assisted prostate cancer diagnosis

Yan Jia [a,*], Clare Verrill [b,c,d], Kieron White [b], Monica Dolton [c], Margaret Horton [e], Mufaddal Jafferji [e], Ibrahim Habli [a]

[a] Department of Computer Science, University of York, York, YO10 5GH, UK
[b] Department of Cellular Pathology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK
[c] Nuffield Department of Surgical Sciences, University of Oxford, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK
[d] NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK
[e] Paige, 11 Times Sq, Fl 37, New York, NY 10036, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Prostate cancer diagnosis
Histopathology
AI safety
Safety case

ABSTRACT

Deep learning (DL) has the potential to deliver significant clinical benefits. In recent years, an increasing number of DL-based systems have been approved by the relevant regulators, e.g. FDA. Although obtaining regulatory approvals is a prerequisite to deploy such systems for real world use, it may not be sufficient. Regulatory approvals give confidence in the development process for such systems, but new hazardous events can arise depending on how the systems have been deployed in the intended clinical pathways or how they have been used with other systems in complex healthcare settings. These kinds of events can be difficult to predict during the development process. Indeed, most health systems and hospitals require self-verification before deploying a diagnostic medical device, which could be viewed as an additional safety measure. This shows that it is important to carry on assuring the safety of such systems in deployment. In this work, we address this urgent need based on the experience of a prospective study in UK hospitals as part of the ARTICULATE PRO project. In particular, the system considered in this work is developed by Paige for prostate cancer diagnosis, which has obtained FDA approval in the US and UKCA marks in the UK. The methodology presented in this work starts by mapping out the clinical workflow within which the system has been deployed, then carries out hazard and risk analysis based on the clinical workflow, and finally presents a deployment safety case, which provides a basis for deployment and continual monitoring of the safety of this system in use. In this work we systematically address the emergence of new hazardous events from the deployment and to present a way to continually assure the safety of a regulatorily approved system in use.

1. Introduction

Prostate cancer is one of the most common cancers among men worldwide, with over 52,000 cases diagnosed each year on average in the UK [1], more than 290,000 estimated cases in the US [2] and over one million estimated cases worldwide [3]. A prostate biopsy is typically involved for confirming most prostate cancer diagnoses, and the utilisation of the Gleason grading system is essential for appropriate stratification and clinical management [4]. The presence or absence of cancer can typically be classified in most cases by the reporting pathologist or in difficult cases with additional testing or further opinion by pathologists. There are occasional instances where pathologists may overlook small cancerous areas (resulting in false negatives), or incorrectly diagnose cancer (resulting in false positives) with the risk influenced by factors such as experience, training and degree of specialism. In contrast, Gleason grading, which impacts treatment recommendations for patients, is inherently subjective, posing challenges in establishing clear and unequivocal classification boundaries for humans, despite international efforts to standardise [5]. For example, Flach et al. has shown that there are substantial variations in prostate Gleason grading between and within Dutch pathology laboratories [6]. This highlights opportunities for deploying DL technologies to assist in the assessment of complex cases and to improve the reproducibility of diagnosis. Indeed, recent years have witnessed significant promise in the integration of such technologies into medical diagnostics, e.g. the development of DL-based systems for prostate cancer detection and grading [7]. As these systems transition from research laboratories to real-world applications, it becomes imperative to ensure not only their efficacy but also their safety in their actual clinical settings.

* Corresponding author.
E-mail addresses: yan.jia@york.ac.uk (Y. Jia), Clare.Verrill@ouh.nhs.uk (C. Verrill), Kieron.White@ouh.nhs.uk (K. White), monica.dolton@nds.ox.ac.uk (M. Dolton), margaret.horton@paige.ai (M. Horton), mufaddal.jafferji@paige.ai (M. Jafferji), Ibrahim.Habli@york.ac.uk (I. Habli).

https://doi.org/10.1016/j.compbiomed.2025.110237
Received 19 August 2024; Received in revised form 3 March 2025; Accepted 17 April 2025
Available online 8 May 2025
0010-4825/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

This paper explores the critical aspects of assuring the safety of a DL-based prostate cancer detection and grading systems produced by Paige [8], known as Paige Prostate Suite, as part of the ARTICULATE PRO project [9], which is a prospective study aiming to examine the impact of introducing artificial intelligence (AI) software into the prostate cancer diagnostic pathway on clinical care across 3 National Health Service (NHS) trusts in the UK. We illustrate this work based on insights from one of the NHS sites as the deployment differs across the three sites.

Paige Prostate Suite is designed to support pathologists in diverse facets of prostate needle core biopsy evaluation, which encompasses three separate modules, Paige Prostate Detection (PPD), employed for identifying potentially cancerous areas of tissues on digital whole-slide images (WSIs), Paige Prostate Grade & Quantify (PPGQ), used to grade and quantify tissue samples, and Paige Prostate Perineural Invasion Detection (PP-PNI), tasked with detecting perineural invasion within the tissue. In this study, our primary emphasis is on PPD and PPGQ, as currently they hold the most significance in aiding pathologists in their diagnostic assessments. Although PNI is prevalent and finding PNI can be time-consuming for pathologists, it is a more specialised diagnostic feature and whether PNI could act as an independent prognostic predictor remains controversial for prostate cancer [10–12], therefore it has less direct impact on immediate treatment and it is not always required to be reported [13], e.g. by College of American Pathologists [14]. PPD has received regulatory approval from the US Food and Drug Administration (FDA) as a second-read modality (to check diagnosis after pathologists have made their own judgement) along with UKCA and CE-IVD marks in Europe for both concurrent and second-read modalities. On the other hand, PPGQ and PP-PNI have obtained UKCA and CE-IVD marks, but not FDA approvals. Despite the regulatory approvals, which do provide some confidence in the safe development of such systems, further hazards may still arise in deployment, e.g. due to deviations from the regulatory approval and depending on how the system is deployed in hospitals. There is a key difference as when applying for regulatory approval, it is practically impossible to predict all of the conditions that might be faced in deployment given different hospitals have different ways of working. To an extent, this will be mitigated by various standards and requirements for healthcare systems, e.g. the United Kingdom Accreditation Service (UKAS) accreditation for laboratories and medical testing facilities, and hospital requirements for self-verification before deploying such systems. However, currently the adoption of such technologies is in the early stages, thus many hospitals are developing their own governance structures and processes to evaluate these technologies. To give one example, within the ARTICULATE PRO project, in order to deploy the Paige Prostate Suite in the hospital, it first had to go through internal validation and then departmental and trust level governance processes. Once approved, then the deployment has to be continually monitored by Department Governance Systems with the aim of achieving UKAS accreditation under ISO 15189 for providing confidence in the quality levels of performance and competence in the medical laboratories [15].

Currently, in the UK, the Royal College of Pathologists (RCPath) has issued a position statement where they stated the intention to develop more detailed guidelines for evaluating and deploying AI, including quality assurance and audit requirements [16] and in the US such guidelines have been provisionally developed by the College of American Pathologists [17]. This shows the importance and urgency of developing methods for continually assuring the safety of DL-based medical devices in deployment. Therefore, in this paper we are attempting to“bridge the gap” between regulatory approval and safety in real-world deployment, with the potential to inform relevant stakeholders to develop guidelines and standards for safe deployment of medical AI. We do this by presenting a methodology and apply it to the deployment of the Paige Prostate Suite, covering both the clinical context and integration with other healthcare IT systems.

The rest of the paper is structured as follows. Section 2 presents the background on the Paige Prostate Suite. Section 3 describes the methodology we have used in this work. Section 4 presents our results, based on applying the methodology set out in Section 3. Related work is presented in Section 5. A discussion is presented in Section 6. Section 7 presents conclusions.

2. Background

PPD is intended to assist pathologists in the detection of prostatic acinar adenocarcinoma in digitised core needle biopsy specimens to reduce diagnostic errors. It is a DL-based system which can classify hematoxylin & eosin (H&E) stained WSI from prostate needle biopsies by producing a binary output identifying a given WSI as benign or suspicious for cancer. If the slide is classified as suspicious for cancer, a single Focus of Interest (FOI) showing the location with the highest probability of harbouring cancer will be presented and a tissue map, which is a visual overlay that fogs out the areas of tissue that are considered not to be suspicious, will also be available. This is depicted in Fig. 1. In order to overcome the need for any pixel-level manual annotations, PPD was developed using multiple instance learning [18], which is a type of weakly supervised learning approach where only the reported diagnoses are used for training, i.e. the ground truth for training is solely derived from the binary classification of each WSI as benign or cancerous as indicated in the corresponding pathology report. For example, slides classified as benign could include basal cell hyperplasia, prostatic intraepithelial neoplasia (PIN), atrophy and inflammation. WSIs containing invasive adenocarcinoma are treated as suspicious for cancer.

PPGQ is a DL-based system designed for grading and quantifying cancer on H&E stained WSI obtained from prostate core needle biopsies. If PPGQ detects any foci of cancer, it will produce a primary and secondary Gleason grade prediction, along with a percentage and length of tumour burden, as shown in Fig. 1. Tumour length is measured along the long axis of the core which has the most tumour if more than one core is present on a slide while tumour percentage is measured as the tumour length divided by the length of that core. Further, PPGQ also generates additional slide overlays to highlight the predicted location for each Gleason pattern, therefore the benign tissues will be faded leaving only the highlighted suspicious areas of tissue for review in the overlays. The Gleason score is the most common prostate cancer grading system; it describes the histologic pattern of gland formation and fusion in the prostate [19]. There are 5 different patterns, graded from 1 to 5 where prostate adenocarcinomas with more gland formation and no fusion receive a low score. For each WSI, the primary Gleason grade describes the most predominant pattern and the highest remaining pattern in addition to the primary pattern is represented by the secondary Gleason grade [20]. The two Gleason grades will then be added together to determine the Gleason score, i.e. the most predominant grade and the highest grade should be recorded in the Gleason core [10]. Theoretically, Gleason scores range from 2–10. However, pathologists almost never assign Gleason scores of 5 or lower, i.e. Gleason pattern 1 or 2 are rarely used. Thus, Gleason scores normally range from 6 to 10, with 6 being the lowest grade cancer. In 2013, a new grading system was proposed, which simplified the number of grading categories from Gleason scores 2 to 10 to Grade Groups 1 to 5 with the lowest cancer grade 1 not 6 as in Gleason [21]. A lower-grade cancer typically grows more slowly and is less likely to spread than a high-grade cancer. The aim in using PPGQ is to help the pathologists to standardise the grading approach and to reduce inter- and intra-observer variability.

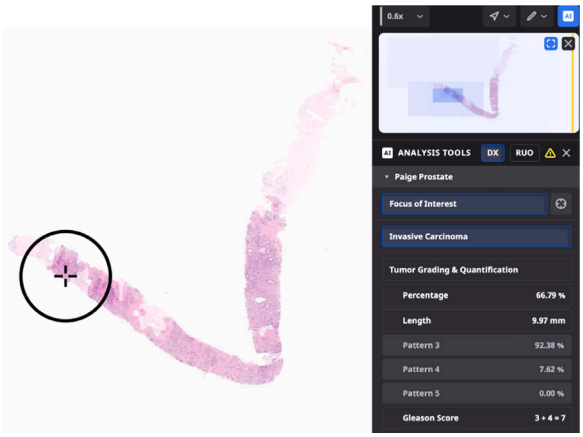


Fig. 1. Paige Prostate Suite sample output.

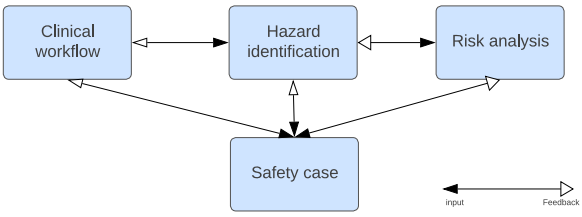


Fig. 2. Overview of the safety analysis methodology.

3. Methodology

There are various safety measures related to medical devices, e.g. intended use statements required for regulatory approval (e.g. by FDA), quality assurance and competence requirements set out by professional accreditation bodies (e.g. UKAS) and evaluation and deployment guidelines set out by professional organisations (e.g. RCPath) if such devices are going to be deployed in hospitals. For this study, we undertake additional activities to examine and analyse clinical context to develop a more holistic understanding of the safety of DL-based medical devices in deployment, i.e. in their usage context, where the guidance is sparse for DL technologies. In order to identify potential new hazards from the deployment of a system, we first endeavour to understand the way in which the tool is integrated into the clinical workflow, and how it is used in conjunction with other tools. Therefore, the first step in our methodology, illustrated in Fig. 2, is to map out the ***clinical workflow*** in which the medical device is intended to be deployed. This will ensure that we not only consider how safe the DL tool is in isolation, but also consider the impact from the wider clinical context which introduces further factors with inherent unpredictability. Hence, the clinical workflow also gives us a basis for safety analysis.

The second step in the methodology is ***hazard identification***, which is critical in safety analysis. A hazard is a potential source of harm [22]. If the hazards were not identified at this step then they would not be subject to the rigour of the risk management process, which could result in a failure to identify risk controls that could be implemented in the system to prevent or mitigate potential harms that may arise during the system's use. There are many different methods for hazard identification, e.g. HAZard and OPerability studies (HAZOP) [23], which typically focus on deviations from intent that could be hazardous in context. Once hazards have been identified, the system or situation is analysed to determine potential causes of the hazards and the potential clinical consequences. Then for each identified hazard, the associated risk is estimated. Therefore, hazard identification also gives us a focus for risk analysis, which is the third step in the methodology.

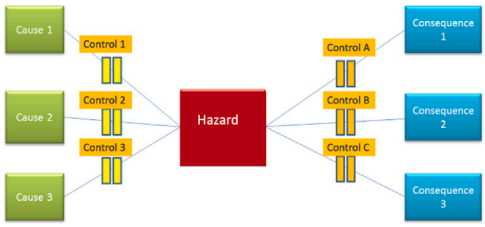


Fig. 3. An example of Bow tie diagram.

In safety engineering, it is common to organise ***risk analysis*** and controls around the notion of a hazard, however risk analysis needs to be interpreted in the context of a particular system or situation [24]. Typically, the risk is expressed as the combination of the probability of occurrence and severity of the hazard's consequences [25]. Sometimes the probability can be quantified; other times estimates are qualitative based on domain knowledge. In addition, the risk of the hazards will determine the priority for the introduction of risk controls (means of preventing the causes of hazards or mitigating the impact of hazards if they do arise). Once risk controls have been identified and introduced, then the risk associated with the hazards can be re-evaluated. This type of risk aversion approach is central to the development of medical devices for market release, ensuring that risk controls are implemented as appropriate.

In this study, we focus on the deployment phase, therefore it potentially could uncover the need for new risk controls, which in turn could impact and inform further development efforts. During risk analysis, it is also useful to employ ***Bow tie diagrams***, as illustrated in Fig. 3, to provide a clear and visual presentation of the relationship between the hazards, causes and consequences of the hazards, and the risk controls. By mapping out the potential causes of the hazards (left side of the bow tie) and their consequences (right side of the bow tie), with the risk controls in the middle, it can help to understand how the hazards are controlled, and systematically assess the risk landscape. The use of the Bow tie diagrams can help to expose the weak points in the system and identify the need for new controls if necessary. There will usually be iteration within risk analysis and bow tie diagrams can help to support the iteration. The value of the visualisation is that it gives a basis for discussion of risk controls with a wide range of stakeholders and supports the development of a safety case.

Finally, the use of ***safety cases*** is a long-established practice in many safety critical domains. Particularly in the UK, the development of a safety case is a mandatory requirement in key sectors such as defence, nuclear and railways [26]. For the NHS in England, compliance with the clinical safety standards, e.g. DCB0160 requires a safety case [27]. This might be legally mandated in addition to the requirements of the medical device regulations if the medical device needs to be implemented within a Health IT system. A safety case for clinical risk management is "a structured argument which is supported by a body of relevant evidence that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given operating environment" [27]. In our methodology, the safety case draws evidence from all the other phases in our analysis (see Fig. 2) and documents the safety rationale, or argument, and any available supporting evidence for the deployment of the DL tools in their clinical context.

By comparison with other approaches, our methodology has the following distinctive merits. Firstly we put the DL-based medical device into its clinical context where we not only consider how the tool is going to be used to support clinicians, e.g. second read or concurrent read, as has been stated in the intended use statements requirement during regulatory approval, but also consider other tools used in conjunction with it. We then go through hazard identification and risk analysis, as is required for regulatory approval. Risk analysis is a fundamental component for regulatory approval, guided by standards such as ISO

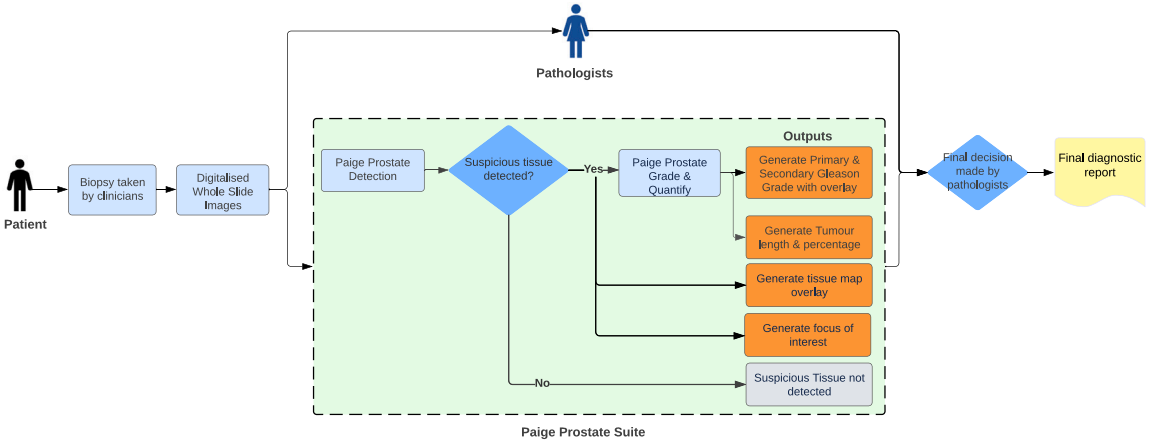


Fig. 4. Paige Prostate Suite clinical workflow in deployment (Concurrent read).

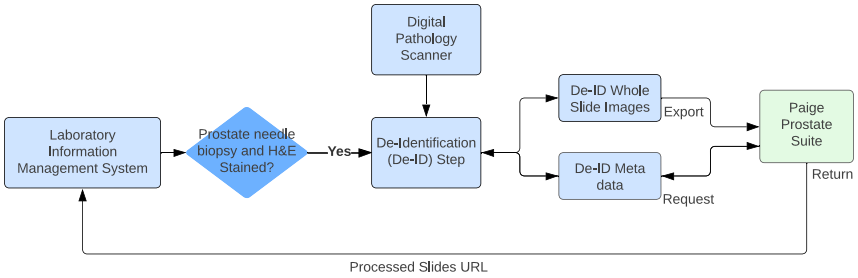


Fig. 5. One simplified example of Paige Prostate Suite technical workflow in deployment.

14971 [22] for medical devices. While many hazards and risks can be identified during the development phases, our approach incorporates the context of the clinical deployment pathway, which adds a depth and richness that cannot be fully achieved during the development phase. Finally we use the safety case to synthesise the evidence obtained during the assessment, using a safety argument to clearly present the links between the different stages. This shows that our methodology goes above and beyond other assessment frameworks in terms of safety assurance.

4. Results

4.1. The clinical workflow

Currently, most of the DL tools that have achieved regulatory approval are for clinical decision-support in the sense that they assist clinicians, and clinicians make the final decisions. The Paige Prostate Suite falls into this category, serving as a decision-support tool assisting pathologists in prostate needle core biopsy evaluation. Fig. 4 shows a depiction of the ***clinical workflow*** for the deployment of the Paige Prostate Suite, which is the same across the 3 NHS sites. Note that, this workflow is downstream of a complex clinical process to identify risk of prostate cancer, and hence a patient's candidacy for biopsy. Therefore, the workflow starts with the prostate biopsies being taken from the patient by clinicians to send to the laboratory for examination. Then, the biopsies are processed in the laboratory and eventually digitised to WSIs, which will be reviewed by the pathologists. If the WSIs are of H&E stained samples, they will be also analysed by the Paige Prostate Suite. Then the pathologists have the option to click the "AI button" on their review screen to see the outputs from the Paige Prostate Suite. If the PPD detects cancerous tissues in the WSI, it will trigger further tumour grading and quantification with PPGQ. Therefore, when there is cancerous tissue detected in the WSIs, the output from the Paige Prostate Suite will include cancerous prediction, a single focus of interest for suspicious cancerous tissue on the WSI, tissue map overlay and also the primary and secondary Gleason pattern along with the length and percentage of the tumour. When there is no suspicious tissue detected, the only output from Paige Prostate Suite is "Suspicious Tissue not detected" as shown in Fig. 4. Then, the pathologist will make the final decision with the support from the Paige Prostate Suite, render the final diagnosis and send it back to the clinicians.

Implementing the clinical workflow involves integrating tools and providing appropriate user interfaces for healthcare staff. This can be quite challenging, as hospitals use many systems from different suppliers, and these may vary significantly between hospitals. For example, for the Paige Prostate Suite to work optimally, it has to be interfaced with the laboratory information management system (LIMS) and/or the digital pathology system in the hospital to obtain essential metadata associated with the relevant H&E stained WSIs and WSIs themselves. However, there may not be sufficient or appropriate access to or collaboration with the suppliers of other software tools deployed in the hospital to assist the integration with the DL tools. Indeed, this was one of the main issues that was highlighted, and overcome during deployment.[1] In Fig. 5, we present one simplified example of the ***technical workflow*** that has been implemented during deployment (note that there are 3 different technical workflows for the 3 NHS sites). The exact technical workflow includes a tool to translate the messages between the different systems. For example, when the Paige system sends a request for de-identified (De-ID) Meta data for the slide it received, it has to pass the message to another tool, which will translate the message, then request the De-ID data from the LIMS, and then return it to the Paige system. In reality these interfaces will be different for every LIMS and AI vendor interaction, which reflects the inherent heterogeneity across hospital systems, so we kept this high level to

[1] At a meeting between the York, Paige, and Oxford teams in John Radcliffe Hospital, Oxford, June 2023.

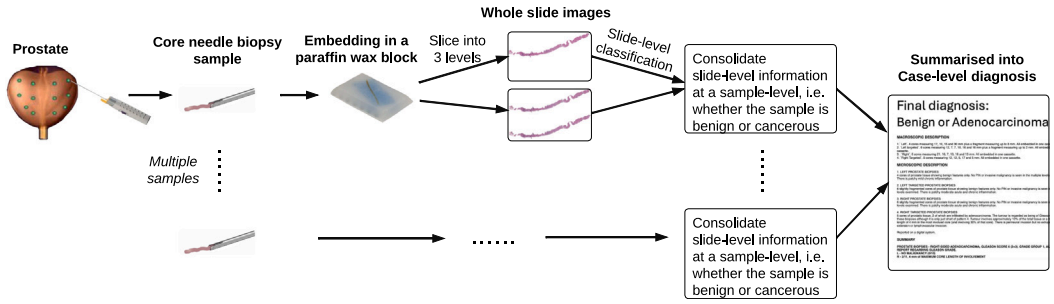


Fig. 6. One example of Reporting Prostate Cancer.

be more representative. In this case, one of the specific requirements from the local hospital was that the Paige Prostate Suite should not have direct access to any patient ID data and pathologists should still be able to review cases from the LIMS environment. Therefore, the relevant WSIs (i.e. H&E stained WSIs from prostate core needle biopsy) are flagged in the LIMS automatically using specimen code and stain code and de-identified before they can be sent to the Paige Prostate Suite. The metadata stored in the hospital systems can include patient ID, case ID, specimen ID, Block ID and slide ID. This allows the result from the Paige Prostate Suite to be re-associated with the patient in the LIMS when the results are returned through the slides' URL. The Paige Prostate Suite has met HIPAA [28] and ISO/IEC 27001 compliance standards [29] and systematically addresses the confidential handling of patient data. However, this case shows how complex and variable the deployment environments and individual requirements can be. The requirements may also evolve: taking a longer term view, if the hospital decides to fully deploy the system after the initial trial, this specific de-identification requirement for this NHS site may be lifted after further discussions with the relevant governance teams.

4.2. Hazard identification

In this section, we first use a variant of HAZOP for computer-based systems, i.e. SHARD [30], for hazard identification. Then we illustrate the clinical impact of the identified hazards. Finally, we applied SHARD again to identify the causes of the Hazards. SHARD, which considers information flows through systems, is suitable for identifying both hazards and causes of hazards. It provides a structured approach to the identification of deviations from intent by systematically applying the guidewords (*omission*, *commission*, *early*, *late* and *incorrect*) to each flow. Here we apply the SHARD method to the clinical workflow in Fig. 4 with a specific focus on the Paige Prostate Suite to identify potential clinical hazards that can arise from the use of the DL tools. The analysis was carried out by a multidisciplinary team with clinical, ML and safety backgrounds.

4.2.1. Identified hazards

The clinical hazards identified by applying the SHARD method on the clinical workflow are as follows:

- H1: False negative classification (slide-level).
- H2: False positive classification (slide-level).
- H3: True positive classification but incorrect localisation as displayed by the single FOI.
- H4: True positive classification but under-annotated tissue map.
- H5: True positive classification but over-annotated tissue map.
- H6: True positive classification but incorrectly annotated tissue map.
- H7: True positive classification but Gleason score is discrepant with reporting pathologists.
- H8: Delayed output.

The guideword *early* is not applicable, as early output does not have a practical meaning here. The guideword *omission* is interpreted as H1: False negative classification on the slide-level. In this context, *commission* means doing something that was not intended. Therefore, it is interpreted as H2: False positive classification on the slide-level. In these kinds of false positive cases, the Gleason score should be discarded. The guideword *late* is interpreted as H8: Depalyed output from the Paige system.

The guideword *incorrect* is interpreted as part of the output is incorrect meaning right slide-level classification, but other parts of the output are not correct. This results in five further potential hazards H3 to H7. The first *incorrect* hazard concerns the location of the single FOI, i.e. "H3: True positive classification but incorrect localisation as displayed by the single FOI". In this context we consider it is a hazard only when the area highlighted by the focus is benign. If the area highlighted by the focus is indeed cancerous, but not the most aggressive tumour, then it is not considered as a hazard as the FOI function is designed to highlight the single point that is predicted to most definitively contain cancer, which is different to a prediction for the highest grade of cancer. The second to fourth *incorrect* hazard concerns the border and coverage of the tissue map, resulting in three sub-hazards, i.e. H4, H5, H6: (i) under-annotated tissue map (i.e. parts of tumour are labelled but not all); (ii) over-annotated tissue map (i.e. all of tumour labelled as well as benign areas); (iii) incorrectly annotated tissue map (i.e. only benign tissue labelled, adenocarcinoma tissue is not).

The fifth *incorrect* hazard concerns Gleason score, i.e. H7: "True positive classification but Gleason score is discrepant with reporting pathologists". Note that we use "discrepant" over "wrong" here. There are two reasons for this. First, when AI's Gleason score differs from the reporting pathologists' opinion or final authorised report, it is not necessarily that the AI is incorrect. Determining the "correct" Gleason score is challenging and in reality more than one score may be reasonable especially in ambiguous or borderline grading cases. The use of discrepancy indicates when the Gleason score suggested by the AI tool differs from the reporting pathologists. Second, although the use of DL technology for predicting Gleason score is intended to improve reproducibility of diagnosis since the Gleason score is inherently subjective as we mentioned in the introduction, the ground truth used for training such DL tools is still human opinion based. Therefore, having a more consistent DL tool does not mean that such subjectivity will disappear. Future AI developments may consider this by identifying robust prognostic indicators such as distinct phenotypic signatures that can be detected by AI, rather than training AI to replicate a subjective human classification system.

Finally, although one of the outputs from Paige Prostate Suite is tumour percentage & length as shown in Fig. 4, we have not included "True positive classification but wrong tumour percentage & length" as one of the hazards because the tumour percentage & length is calculated based on the annotation on the tissue map, similar to the Gleason pattern percentages which are also calculated based on the annotated tissue map. An example of such an annotated tissue map

Table 1
Paige Prostate Suite hazard and risk analysis.

Guide word	Clinical Hazards	Possible Causes	Risk controls	Probability	Severity	Risk Rating	Comparison to development
Omission	H1: False negative classification (slide-level) *Clinically at a case level significant if no cancer is flagged at all by Paige when cancer is present on any slide(s) and the pathologist making a case-level false negative diagnosis (see Fig. 6). In a case containing cancer, if slides within the case are already flagged correctly as cancer and reported as so by the pathologists, then a false negative read on one or more other slides in the case are of lesser significance.*	1. Information mismatch 2. PPD DL model produced false negative prediction (e.g. when out of focus slides are analysed and given a false negative classification)	1. Contextual launch 2. Model testing/tuning, analytical performance and clinical validation 3. User Training: extensive training is provided to pathologists with examples of potential failure mode 4. Instructions for Use: pathologists should make the final decision with the support from the tool 5. The use of immunohisto-chemical (IHC) testing	Low	Significant	Moderate	Extra risk controls are identified, i.e. contextual launch, but overall risk is deemed unchanged
Commission	H2: False positive classification (slide-level) *Clinically at a case level significant if a case is entirely benign/PIN but Paige has flagged an area or areas suspicious of cancer and the pathologist also calls this cancer leading to a case-level false positive diagnosis. In a case containing cancer, if slides within the case are already flagged correctly as cancer and reported as so by the pathologists, then a false positive read on one or more other slides in the case are of lesser significance.*	1. Information mismatch 2. PPD DL model produced false positive prediction	1. Contextual launch 2. Model testing/tuning, analytical performance and clinical validation 3. User Training: extensive training is provided to end users with examples of potential failure mode 4. Instructions for Use: pathologists should make the final decision with the support from the tool 5. The foci generated by the tool will further act as a risk control 6. The use of immunohisto-chemical (IHC) testing	Low	Significant	Moderate	Extra risk controls are identified, i.e. contextual launch, but overall risk is deemed unchanged

(*continued on next page*)

can be seen in Fig. 1. Therefore wrong tumour percentage & length is a result of the tissue map being under-annotated, over-annotated or incorrectly annotated rather than a stand-alone hazard. The identified hazards are presented in Table 1. Examples of the potential AI failure modes that could contribute to these hazards include: prostate biopsy cores may be spaced very closely together in one slide so that the algorithm cannot distinguish them as separate cores leading to over-annotated tissue map; over-annotation where a benign area is marked as suspicious in one core due to sensitivity; core is fragmented or not complete leading to under-annotated tissue map. In the case of false negatives, a previous investigation [31] has reported this could arise in instances of glandular atypia.

Table 1 (*continued*).

Guide word	Clinical Hazards		Possible Causes	Risk controls	Probability	Severity	Risk Rating	Comparison to development
		H3: True positive classification but incorrect localisation as displayed by the single FOI	1. PPD DL model misidentifies suspicious areas	1. Model testing/tuning, analytical performance and clinical validation 2. Tissue map overlay 3. User Training: pathologists are trained not to locate a single focus 4. Instructions for Use: pathologists should make the final decision with the support from the tool	Low	Negligible	Minor	Extra risk controls are identified, i.e. the tissue map, thus overall risk is reduced
		H4: True positive classification but under-annotated tissue map	1. PPD DL model produced under-annotated tissue map	1. Model testing/tuning, analytical performance and clinical validation 2. User Training: extensive training is provided to pathologists to be aware of such potential failure mode& frequency 3. Instructions for Use: pathologists should make the final decision with the support from the tool	Low	Minor	Minor	More refined Hazard definition
		H5: True positive classification but over-annotated tissue map	1. PPD DL model produced over-annotated tissue map	1. Model testing/tuning, analytical performance and clinical validation 2. User Training: extensive training is provided to pathologists to be aware of such potential failure mode& frequency 3. Instructions for Use: pathologists should make the final decision with the support from the tool	Medium	Minor	Moderate	More refined Hazard definition

(*continued on next page*)

Table 1 (*continued*).

Guide word	Clinical Hazards		Possible Causes	Risk controls	Probability	Severity	Risk Rating	Comparison to development
Incorrect	Part of the output is incorrect	H6: True positive classification but incorrectly annotated tissue map	1. PPD DL model produced incorrectly annotated tissue map	1. Model testing/tuning, analytical performance and clinical validation 2. User Training: extensive training is provided to pathologists to be aware of such potential failure mode& frequency 3. Instructions for Use: pathologists should make the final decision with the support from the tool	Low	Minor	Minor	More refined Hazard definition
		H7: True positive classification but Gleason score is discrepant with reporting pathologists	1. PPGQ DL model produced discrepant Gleason score	1. Model testing/tuning, analytical performance and clinical validation 2. User Training: extensive training is provided to pathologists to be aware of such potential failure mode& frequency 3. Instructions for Use: pathologists should make the final decision with the support from the tool	Medium	Moderate	Moderate	More refined Hazard definition

(*continued on next page*)

4.2.2. Clinical impact of the hazards

In order to understand the clinical impact of the hazards identified above, it is important to understand prostate cancer management. In the clinical context of managing prostate cancer, there are three important histopathology factors to consider [32], i.e. a case-level diagnosis, Gleason score, and tumour burden assessment. Among them, a case-level benign/malignant diagnosis along with other clinical findings, e.g. multiparametric MRI result, have a significant impact on further management, usually whether patients enter a prostate cancer pathway or are discharged. Once malignancy is confirmed, Gleason score then becomes the next important feature in determining management options. To a lesser extent the number of cores involved and tumour burden are used to assess suitability for surveillance or other options. All of the clinical hazards identified above are related to individual slide analysis as the Paige Prostate Suite analyses individual WSIs. The final diagnosis involves consolidating information from individual WSIs into a case-level assessment. There will be small variations between labs in workflows, sample preparation and reporting practices which can influence the clinical impact of the hazards. We illustrate this based on one of the NHS trusts participating in this study, see Fig. 6. In this case, multiple core needle biopsy samples are usually taken from various sites in the prostate by either a transrectal or templated transperineal protocol (LATP), including specific lesions on MRI scan which may be targeted. The biopsies from each sample or area are blocked into one paraffin wax block and 3 sections or levels from the block are cut for analysis. This ensures thorough examination and reduces the risk of missing any cancerous tissues in the prostate. Two levels go on one slide, i.e. L1 and L2 go on one slide (see Fig. 8(a)), and L3 goes on a second slide (see Fig. 9). Pathologists do not report on individual WSIs; instead, they often report on whether each sampled site has been infiltrated by adenocarcinoma, i.e. sample-level assessment. Finally, pathologists will summarise all of the sample-level information to produce an overall diagnosis, i.e. case-level diagnosis.

Therefore, WSI-level false negative classification does not necessarily result in false negative diagnosis for the patient if other slides from that core needle biopsy sample contain cancer tissues and have been analysed correctly. Nor does it result in an overall false negative diagnosis if other core needle biopsy samples in the case contain cancer. Similarly, WSI-level false positive classification does not necessarily result in false positive diagnosis for the patient if other slides from that core needle biopsy sample contain cancer tissues and have been analysed correctly. Nor does it result in an overall false positive diagnosis if other core needle biopsy samples in the case contain cancer.

Table 1 (*continued*).

Guide word	Clinical Hazards	Possible Causes	Risk controls	Probability	Severity	Risk Rating	Comparison to development
Late	H8: Delayed output	1. Paige Prostate Suite technically fails ingestion, analysis and/or visualisation 2. Paige Prostate Suite processed slides returned late 3. DL model is not triggered on a slide	1. Retry sending notifications for processing slide 2. Warnings and Error Messages 3. Inherent Safety by Design: CPU fallback, software configuration, error handling, integrity verification etc… 4. Slide pre-processing component 5. Code review and test coverage analysis of production code 6.User Training: extensive training is provided to pathologists with examples of potential failure mode, with pathologists reverting to reporting without AI 7. Instructions for Use: the tool only process H&E stained prostate needle biopsy that meet specified quality criteria	Low	Negligible	Minor	Unchanged
Early	N/A						

The Gleason score is also assessed and in this hospital, the maximum Gleason score from any one specimen in the case is reported in the bottom line of the case. Possible variations in other medical centres include embedding individual cores each in a block, putting all levels on 1 or 3 slides, and the use of the overall Gleason score, see RCPath standards for more information [10].

It is important to understand how the hazards identified here might impact the clinical management of prostate cancer in determining the severity level of the hazards and overall risk, as outlined in Section 4.3, since this informs the prioritisation of risk mitigations and the acceptance of the product.

4.2.3. Causes of the hazards

After the identification of the hazards for the clinical workflow, we applied SHARD to the technical workflow in Fig. 5 to identify technical failures that can contribute to the clinical hazards identified above. A technical failure arises where a system or item of software does not carry out its intended function, e.g. does not transmit a WSI. But a failure mode, in itself, is not a hazard. It could act as a trigger event that could lead to harm by activating exposure to one or more hazards. The analysis was carried out with the help from the IT manager of Cellular Pathology in one of the hospitals. It is important for the analysis to consider the interdependencies between different sources of hazards. Once this is established, separating technical failures and clinical hazards might allow more efficient engagement of the relevant expertise and thus optimising the use of resources. Further, it also helps to improve communication about safety issues within the healthcare organisation, e.g., between technical and clinical teams, enabling a collaborative approach to addressing safety concerns which ultimately contributes to a safer and more robust healthcare environment.

The resulting potential technical failures were identified by applying the SHARD method to the technical workflow are as follows:

- Omission — Paige Prostate Suite technically fails ingestion, analysis and/or visualisation.
- Commission — N/A
- Incorrect — Information mismatch, i.e. slide images do not match patients.
- Early — N/A
- Late — Paige Prostate Suite processed slides returned late.

Commission and *early* are considered not applicable as Paige Prostate Suite returning an output when not requested or early is not plausible. After the identification of the potential technical failures, we have further analysed their impact, i.e. what clinical hazards they

can contribute to. This shows that the technical failures are possible causes of the clinical hazards. It is likely that all of these potential technical failures would have been considered when assessing the tools for regulatory approval. However, analysis of the technical workflow would potentially identify more specific causes of these failures. For example, the cause for "Paige Prostate Suite technically fails ingestion, analysis and/or visualisation" can be "WSI and metadata are unable to be processed by Paige", which could be due to image quality issues, file corruption, data issues, or specifically the de-identification step introduced here does not translate the request properly. In this work, we did not further analyse the causes for technical failures as our focus is on DL technologies, but this example highlights the merit in working through the technical workflow to identify the complete set of potential failure causes. Further, there is also value in exploring how varying technical workflows could influence the risk analysis of DL tools, however it is outside scope of this study.

Table 1 presents a summary of the identified clinical hazards, possible causes for the hazards, and the existing risk controls for mitigating the hazards, integrating the results of the analysis of the clinical and technical workflows. The possible causes of the hazards, as shown in Table 1, are mainly from DL models themselves and the high-level technical failures without going back to potential root causes, e.g. file format incompatibilities. For risk controls, note that we have not listed all of the traditional software risk controls in the table, e.g. fixed dependencies in a given software version, merely enumerating some for illustrative purposes. However, we would like to highlight one of the risk controls, i.e. contextual launch where a unique URL will be generated for each patient in the LIMS for pathologists to access the results, which is specifically introduced to prevent information mismatch after patient de-identification. Further, DL model testing/-tuning, analytical performance and clinical validation performed by the vendor are the main risk controls for the DL model itself. There is no institution-specific tuning involved. Once a version of Paige Prostate Suite is released, the algorithm is fixed and remains immutable with no adjustable operating points specific to individual sites. This shows that it is difficult to have inherently safe design for DL due to its "black box" nature. However, the presentation of foci and the tissue map can be thought of as a form of explanation which also provides risk controls for the DL models. Additionally, discussion of the probability, severity and risk rating columns are provided in Section 4.3, while the comparison to development column is elucidated in Section 4.4.

4.2.4. Hazards in real-world scenarios

In this section, we present examples of the hazards encountered during the hospital deployment to illustrate their real-world manifestation. These examples demonstrate that the identified hazards are not merely theoretical but occur in practice, showing the effectiveness of our safety analysis methods. Specially, we show three real-world scenarios associated with hazards, H2, H3, H5, H7 (see Fig. 7, Fig. 8, Fig. 9).

Patient scenario 1: Slide-level false positive classification (H2)

Patient A underwent multiple prostate core needle biopsy samples/specimens and was diagnosed with adenocarcinoma as there was Gleason Score 7 (3+4) adenocarcinoma in specimen 3. However one of the WSIs from specimen 5 of this case was benign, but the Paige system incorrectly predicted it as cancerous, as shown in Fig. 7. Therefore, the area highlighted by the single FOI was benign as well. This shows a slide-level false positive classification (H2), although it did not result in a case-level false positive. Further, the indicated Gleason Score 6 (3+3) in this case should be discarded as well.

Patient scenario 2: True positive slide-level classification but incorrect localisation as displayed by the single FOI (H3) and over-annotated tissue map (H5)

Patient B underwent multiple prostate core needle biopsy samples/specimens and was diagnosed with adenocarcinoma. The Paige system correctly identified a slide from specimen 6 as cancerous with Gleason Score 6 (3+3) adenocarcinoma. However, the area highlighted by the single FOI was actually benign, while the true adenocarcinoma area is indicated by the red arrow in Fig. 8. This is associated with hazard H3. Notably, the invasive carcinoma tissue map successfully detected this adenocarcinoma area, but also incorrectly identified some benign areas as containing adenocarcinoma. This shows the hazards H5 – True positive classification but over-annotated tissue map where all adenocarcinoma are highlighted as well as benign areas. Other specimens in this patient case also contained adenocarcinoma. This scenarios combines the hazards H3 and H5, illustrating that pathologists rarely rely solely on the FOI when making diagnosis, thereby reducing the clinical significance of H3 in diagnostic decision-making.

Patient scenario 3: True positive slide-level classification but Gleason score is discrepant with reporting pathologists (H7)

Patient C underwent multiple prostate core needle biopsy samples/specimens and was diagnosed with adenocarcinoma. One of the WSIs was correctly flagged as cancerous by the Paige system but assigned a Gleason Score of 8 (4+4). However, the reporting pathologist did not agree with this and assigned a Gleason Score of 7 (3+4) in the final report. The pathologist's decision was based on histological features evident at 5x magnification, which revealed predominantly well-formed discrete glands characteristic of a primary Gleason pattern 3, not pattern 4 as the system suggested. The presence of a few poorly formed glands supported the secondary Gleason pattern 4 (see Fig. 9).

This scenario highlights the hazard H7 – True positive classification but Gleason score is discrepant with reporting pathologists. In routine clinical practice, each patient case is typically reviewed by a single pathologist, unless the pathologist requires a second opinion from senior colleagues. Therefore, in this study as in real life, not all patient cases have had a second opinion. Thus, the hazard H7 refers to discrepancies between AI system and the reporting pathologists' opinion or the final authorised report, rather than a reference standard such as a panel of experts as the former more closely replicates real world practice. Establishing a definitive reference standard or ground truth for Gleason score is difficult in an area with known inter and intra-observer differences. Even among expert panels, consensus is not always reached as more than one Gleason score may be reasonable in reality. However, hospitals regularly hold quality assurance meetings to discuss the difficult cases. For example, in this study, there was a consensus review where participating pathologists discussed how they would grade selected cases – though not all cases. Further, there are efforts to standardise Gleason score through external quality assurance (EQA) schemes as well.

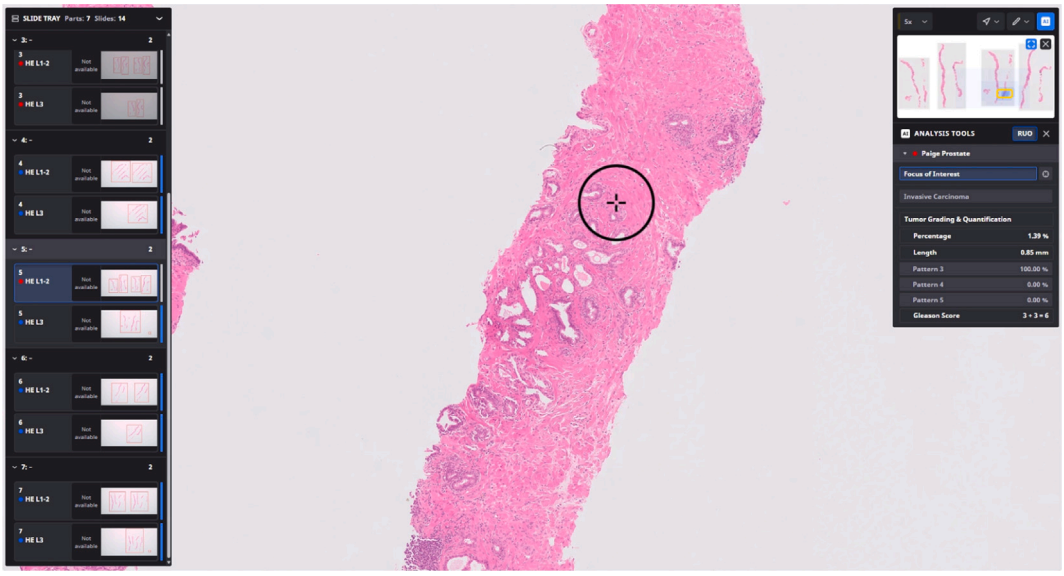
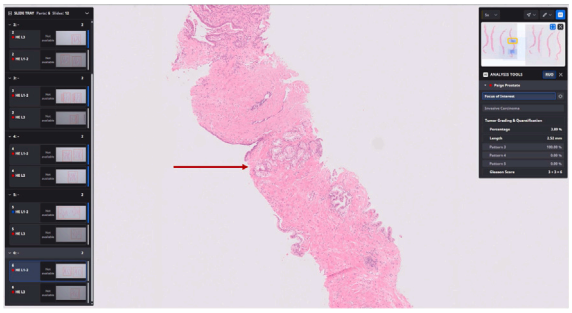


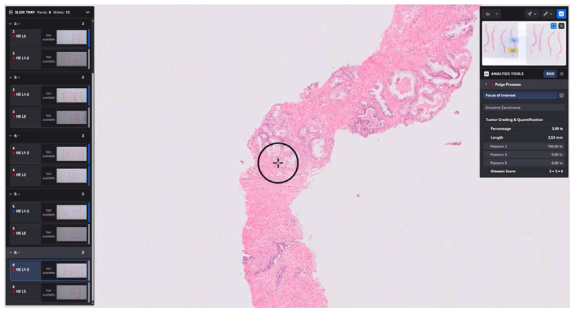
Fig. 7. Patient scenario 1 – slide-level false positive classification (Hazard H2). This area highlighted by the single FOI and this WSI from specimen 5 of this patient case is benign, showing some inflammation which can often produce some mild atypia and there are also prominent basal cells. There was Gleason Score 7 (3+4) adenocarcinoma in another part of the case (specimen 3). **5x zoom**.

(a) One WSI along with the Paige output. The area highlighted by the FOI is benign whilst the area of adenocarcinoma is pointed to by the red arrow. **1x zoom**.

(b) Invasive carcinoma tissue map. The adenocarcinoma area is adjacent to the red arrow and is highlighted by the tissue map. **1x zoom**.



(c) Higher power view of adenocarcinoma area as pointed to by the red arrow. **5x zoom**.



(d) Higher power view of benign area highlighted by the FOI. **5x zoom**.

Fig. 8. Patient scenario 2: True positive slide-level classification but incorrect localisation as displayed by the single FOI (H3) and over-annotated tissue map (H5). This figure shows the different views of the same slide from specimen 6 of the patient case. The panel (a) shows the slide from specimen 6 was correctly flagged as cancerous and showing Gleason Score 6 (3+3) adenocarcinoma. However, the area highlighted as most suspicious for adenocarcinoma in the circle by the Paige system is benign and the 0.6 mm area of adenocarcinoma is adjacent to the red arrow. The area of adenocarcinoma was highlighted in the invasive carcinoma tissue map, as we can see in panel (b). The panels (c) and (d) show the higher power view of adenocarcinoma area near the red arrow and the benign area highlighted by the FOI.

4.3. Risk analysis

As noted in Section 3, after the hazard identification, the level of risk associated with each hazard is estimated. In order to analysis risks, two components, i.e. probability and severity of the harm, should be analysed.

In this study, a qualitative risk analysis is performed by a multi-disciplinary team comprising experts with backgrounds in clinical, AI, and safety. Specifically, the team first collaboratively developed a three-level scale for probability and a four-level scale for severity, presented in Tables 2 and 3. Then the team developed the risk matrix, as shown in Table 4. Combining the probability and severity gives a

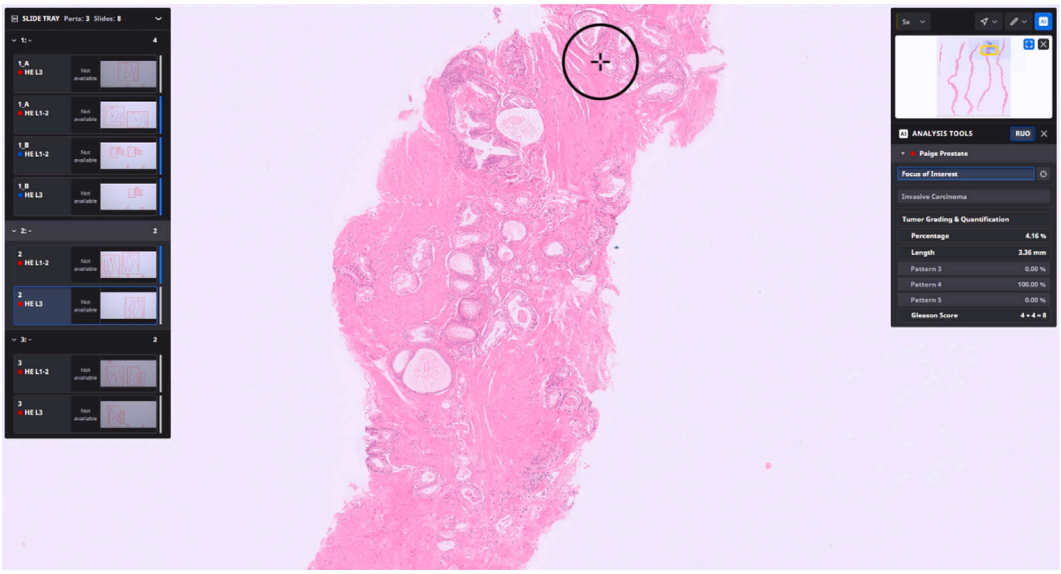


Fig. 9. Patient scenario 3: True positive slide-level classification but Gleason score is discrepant with reporting pathologists (H7). This WSI from specimen 2 was assigned a Gleason Score of 8 (4+4) by the Paige system, but the reporting pathologist did not agree with this and assigned a Gleason Score of 7 (3+4) in the final report. **5x zoom**.

Table 2
Three qualitative probability levels.

Levels	Description
Low	Unlikely to happen, rare, remote
Medium	Can happen but not frequently
High	Likely to happen, often, frequently

Table 3
Four qualitative severity levels.

Levels	Description
Negligible	No injury or slight injury
Minor	Minor injury from which recovery is expected
Moderate	Severe injury or severe incapacity from whichrecovery is expected
Significant	Death or permanent harm

Table 4
A qualitative 3 × 4 Risk Matrix.

Risk assessment		Qualitative severity levels			
		Negligible	Minor	Moderate	Significant
Qualitative probability levels	High	Minor	Moderate	Major	Major
	Medium	Minor	Moderate	Moderate	Major
	Low	Minor	Minor	Minor	Moderate

risk matrix with twelve combinations. Each cell in Table 4 is given an overall risk rating (minor, moderate or major). For example, low probability of the occurrence of the harm with a significant severity is classified as Moderate.

The overall risk level also identifies the approach to mitigation, viz:

- Overall Risk = **Minor**: no injury or damage to health possible; Mitigation: No mitigation is required
- Overall Risk = **Moderate**: non-serious injury possible; Mitigation: If no mitigation is applied, there must be justification that the benefits outweigh the risk
- Overall Risk = **Major**: unacceptable/death or serious injury possible; Mitigation: All major risks must be mitigated

Finally, utilising the aforementioned matrices, the team assessed the probability and severity separately for each hazard entry, as outlined in Table 1, based on the authors' real-world experience of using the Paige system. However, we acknowledge that hazards may be of differing significance depending on the specifics of different hospitals and their local clinical practices.

When assessing probability, it is important to take a Paige-oriented focus. This means that we estimate the likelihood of the hazard itself occurring, rather than the likelihood of its consequences, i.e. clinical harm. In healthcare, clinicians are highly effective at preventing harm from reaching patients. Therefore, understanding system behaviour is more important than focusing solely on the ultimate harm. When assessing severity, it is important to understand the clinical impact of the hazards, i.e. how they might change the clinical management of prostate cancer, as we mentioned in Section 4.2.2. Case-level diagnosis will determine whether patients enter a prostate cancer pathway or not. A false negative case-level diagnosis could result in a missed opportunity for timely cancer treatment while a false positive diagnosis could lead to unnecessary repeated biopsy, exposing patients to associated comorbidities. Therefore, we classified the severity of H1 and H2 as significant. Although, incorrect Gleason scoring could also potentially lead to severe consequences, e.g. over-treatment, under-treatment or potential long term side effects from inappropriate interventions we classified the severity of H7 as moderate rather than significant. There are two reasons for this. The first reason is related to the specific architecture of the Paige Prostate Suite. As illustrate in Fig. 4, the WSIs first are processed by PPD, which identifies suspicious tissue in WSIs. Only WSIs with tissue defined as "suspicious" by PPD are passed to PPGQ. This workflow means that if cancer detection is not accurate, such as in cases of false negative classifications, those cases will not proceed to the Gleason scoring stage. Therefore this severity rating reflects the system's operation logic. Secondly, once the presence or absence of cancer is established, Gleason scoring, while crucial, only exists within a broader diagnostic process. The overall clinical decision-making for treatment involves multiple factors, such as PSA levels, cancer staging and other clinical data. This integrative approach means no single diagnostic metric determines treatment in isolation, therefore

further supports the rating of Gleason score misclassification as moderate. Finally, we classified the severity of H4 to H6 as minor and the remainder as negligible. The severity classification directly reflects the degree of importance of the key factors impacting clinical management, i.e. case-level diagnosis, Gleason score, and tumour assessment. In general, it is crucial to take a clinical focus when conducting risk estimation rather than just focusing on the technical level of the medical devices. For example, for hazard H8: delayed output, the severity was deemed as negligible as the pathologists can revert to reporting without AI. However, this would be a significant hazard if we took a technical perspective that a piece of software is not producing the intended output. Finally, the resulting risk rating for the Paige system indicates the risk associated with the hazards post-implementation of the existing risk controls within the Paige system and before the intervention of clinicians, which are all documented in Table 1. Based on the risk matrix in Table 4 and risk assessment approach mentioned above, the overall risk level for all of the hazards falls within acceptable level, thus does not require further mitigations or risk controls.

4.4. Comparison to development hazard log

To continually assure the safety of DL-based medical devices in deployment, it is important to understand safety-relevant changes that have occurred since their development. In this section, we compare the hazards identified in Table 1 to the development hazard log to identify changes. As a result, the risk associated with one of the hazards identified during development has been reduced due to the implementation of additional risk controls. More refined hazard definitions have also been established from deployment based on one of the general hazards that was identified in development, i.e. false quantification and/or grading for cancer. Note that the development hazard log submitted to the regulator is far more comprehensive than Table 1, as our intention here is to highlight the changes in hazards specially related to DL technologies after the initial development.

Specifically, we have derived four more refined hazard definitions, see Table 1: H7 arises out of the clinical workflow due to the use of two Paige AI tools (PPD and PPGQ) together in the deployment context (see the green area in Fig. 4). Another three, i.e. H4, H5, H6, are related to the tissue map overlay which was introduced for the UKCA and CE-IVD marked version of PPD compared to the FDA-approved version. These detailed hazard definitions enable clinical risks to be further contextualised and elucidated by performing this analysis in a real-life deployment setting. Further, the risk associated with H3 "True positive classification but incorrect localisation as displayed by the single FOI" has been reduced due to the introduction of the tissue map as it highlights all of the suspicious areas of cancerous tissue discouraging the pathologist from focusing on just one location. Extra risk controls are also identified, i.e. contextual launch to prevent information mismatch, which could further contribute to hazards H1 "False negative classification" and H2 "False positive classification". However, the risk associated with these two hazards, H1 & H2, are not deemed reduced compared to development due to the extra risk introduced by the de-identification step. Therefore, we consider the overall risk associated with these two hazards, H1 & H2, to be unchanged. The risk associated with H8 "Delayed output" is deemed unchanged as well.

This underscores the significance of ongoing and systematic safety assurance for such systems in deployment and demonstrates that we are not merely duplicating the regulatory efforts or clinical governance & validation requirements or scrutinising the manufacturer's work.

4.5. Risk control visualisation

Bow tie diagrams are a useful means for visualising how hazards are controlled. Here we present a bow tie diagram for Hazard H3: "True positive classification but incorrect localisation as displayed by the single FOI" in Fig. 10 for illustration. There are five important elements of a bow tie diagram:

- Context (square with the black and yellow border) — an activity or condition that is part of normal operation, but which can be a source of harm. In this case, the context is "Output from the Paige Prostate Suite" in Fig. 10;
- Top event (amber circle) — the occurrence of the hazard. In this case, it is H3: "True positive classification but incorrect localisation as displayed by the single FOI" occurred;
- Threats (round-cornered blue box) — a cause that contributes to the hazard. In this case, there is one threat identified for H3, arising from the DL model itself, as shown in Table 1;
- Consequences (Orange box) — the potential result from the occurrence of the hazard. In this case, it might cause "fail to detect potential cancerous tumours for the slide", but if the focus was not activated by the pathologist, then the right result might still arise;
- Risk controls (Grey boxes) — measures taken to reduce the risk. If they appear on the left side, they help to reduce the likelihood that a threat can cause the top event. if they appear on the right side, they help to mitigate the impact of the top event. In this case, five risk control measures, as shown in Table 1, have been presented in addition to pathologists. Notably, the "tissue map overlay" emerges as an additional control to help to mitigate the impact of hazard H3 by offering a holistic view of areas exhibiting suspicious cancerous tumours on a slide image. Therefore, the overall risk associated with hazard H3 is deemed reduced as we indicated in Section 4.4.

Hazard and risk analysis is exploratory, so it is important to assess the credibility of the analysis. One of the difficult aspects for producing the bow tie diagram in Fig. 10 is assessing the consequences of hazard H3 as even for the same hazard, different credible consequences might occur. In one of the retrospective studies conducted previously without the tissue map using Paige Prostate Suite [33], one pathologist changed the initial correct classification (Cancerous) to (Benign) after reading the Paige system output, which presented a correct cancerous prediction but with an incorrect location as indicated by the focus, i.e. it was benign where it was highlighted in the WSI.

This shows that one of the possible consequences of hazard H3 is "Fail to detect potential cancerous tumours for the slide", but it occurred prior to the inclusion of the tissue map, so it also shows the value of the tissue map as an additional risk control. However, the worst credible consequence from this hazard H3 is successful detection based on the current deployment experience. Therefore, the severity for this hazard is negligible, as indicated in Table 1. However, this is limited by the observation period and if over time, this led the pathologist not to vigilantly search the rest of the tissue for the presence of cancer or more aggressive cancer, then the severity of this hazard H3 would need to be updated.

The bow tie diagram does not just repeat Table 1. Instead, it offers a more refined view of the relationship of the causes of the hazards and the controls for the hazards. It helps to show the whole process of the propagation of a cause to the consequence if all of the controls fail. Further, it helps us to see more clearly what types of risk controls we have introduced. If all of the controls appear on the right hand side, i.e. for mitigation after the hazards occur, then it is likely to be useful to think about what controls can be introduced to prevent the hazards occuring or to reduce the probability of occurrence.

Finally, it is important to continue to monitor systems in operation to confirm the credibility of the analysis, the effectiveness of the controls, and to prompt action if new hazards arise which were not predicted.

4.6. Safety case

All the phases of the methodology in Fig. 2 feed into the safety case. The safety case draws together and integrates the work in the different

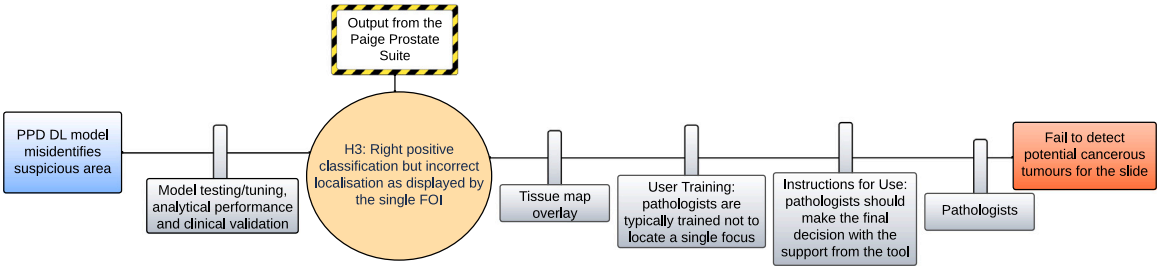


Fig. 10. Bow tie diagram for H3: "True positive classification but incorrect localisation as displayed by the single FOI".

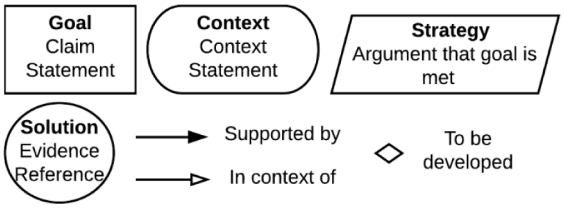


Fig. 11. Goal Structuring Notation.

phases of the methodology, showing and critically evaluating how the information produced might assure the safety of the "system" which, in this work, is taken to mean the Paige Prostate Suite. Before we describe the safety case we have developed, we briefly introduce the notation.

In this work we present the safety argument using the Goal Structuring Notation (GSN) [34]; a legend showing the key elements of the notation is presented in Fig. 11. The goals – claims that we wish to make and support – are shown as rectangles and they can be decomposed into sub-goals thus forming a tree. Goals are understood in a context – for example, the clinical setting of the system. Where the decomposition of goals is not obvious this is explained through a strategy, represented as a rhombus. In a complete safety case all leaf-level goals are supported by solutions, represented as circles; the solutions provide evidence references to support the argument. Incomplete parts of the argument are shown with a diamond, meaning that part of the argument is to be developed. The detailed description of the notation can be found in https://scsc.uk/r141B:1?t=1

Here we present a snapshot of the safety argument in Fig. 12 with the top goal G0 "Sufficient controls are in place for risks associated with deploying Paige Prostate Suite into the hospital". As we mentioned above, goals should be understood within their context, thus G0 should be interpreted in the context of the deployment hospital as the risks associated with deploying Paige Prostate Suite might vary in different hospitals along with the risk matrix. G0 is decomposed using the strategy "argument over risks associated with hazards when deploying Paige Prostate Suite".

In Section 4.2, we have identified eight hazards in total after applying SHARD method to the clinical workflow in Fig. 4. Hazards H3 to H7 are the detailed hazards arising from "part of the output is incorrect", so we combine them into one goal G3. Therefore, G0 is decomposed into four subgoals G1 to G4. For each subgoal G1 to G4, we use the strategy "argument over risk controls for hazards" to decompose it further. For G1, two hazard causes are identified, therefore it is further decomposed to G5 – "PPD DL model produced false negative prediction is controlled by model testing/tuning, analytical performance and clinical validation, user instructions, user training and IHC testing" and G6 – "Information mismatch is controlled by contextual launch" with each representing one of the hazard causes. For G2, one of the hazard causes is the same, i.e. information mismatch, therefore G6 also supports G2.

There are five hazard causes for G3, i.e. PPD DL model misidentifies suspicious areas, PPD DL model produced under-annotated tissue map, PPD DL model produced over-annotated tissue map, PPD DL model produced incorrectly annotated tissue map, and PPGQ DL model produced discrepant Gleason score. However, the risk control for all of them is model testing/tuning, analytical performance and clinical validation, user instructions and training except H3 has an additional control of tissue map overlay, therefore we combined them into one subgoal G8. For G4, there are three hazard causes associated with it. However, the risk controls for these three hazard causes are intertwined in that retry sending notifications, warnings and error messages, inherently safe design etc could act as a barrier for all of the three hazard causes, therefore we have combined these into one subgoal G9 "possible causes for Paige Prostate Suite to process slides unsuccessfully are controlled by retry sending notifications, warnings and error messages, inherent safe design etc".

Finally, the evidence to support the leaf goals G5, G6, G7, G8, G9 all should come from the data, documents and evaluation conducted by the wider evaluation and the local clinical safety teams (hence the use of the 'to be developed' diamond symbols for these goals).

By developing the safety case for deploying Paige Prostate Suite in the hospital, it shows how the different phases in the methodology link together and support each other to evaluate the safety of the DL tools in their clinical context. Further, safety cases can enhance transparency by documenting the risk controls implemented (enabling reviews) and also critically challenging the reasons behind them, which is crucial for organisations to continue monitoring the safety of the system in operation.

5. Related work

The number of regulatory-approved AI/ML-based medical devices has increased significantly over the past decade. However, even AI tools that have received regulatory clearance for clinical use may underperform when deployed in new clinical settings due to poor generalisation or off-label use [35]. Studies have highlighted that some FDA-approved AI/ML medical devices lack sufficient published information about their validation datasets, making it difficult to justify their clinical applications [36]. These cases highlight the difficulties faced for deploying AI in healthcare. Some commentaries have emphasised the necessity of establishing a comprehensive framework of quality assurance and training procedures to ensure consistent safety and effectiveness of AI tools in deployment [37–39]. However, there is little published work investigating continual assurance of the safety of AI/ML-based clinical decision support systems in deployment, based on real-world experience. This is unsurprising as the introduction of such technology in hospitals is still at an early stage. More work exists on exploring the safety of the AI tools in development [40]. Some work focuses on how to assure the technology, and other researchers have explored the robustness of the underlying AI/ML model. For example, Assurance of ML in Autonomous Systems (AMLAS) [41] provides a systematic approach to evaluating the safety of an ML component during development, producing an associated safety case. Here, we mention some relevant work on exploring the robustness of the AI/ML algorithms for prostate cancer detection based on WSIs.

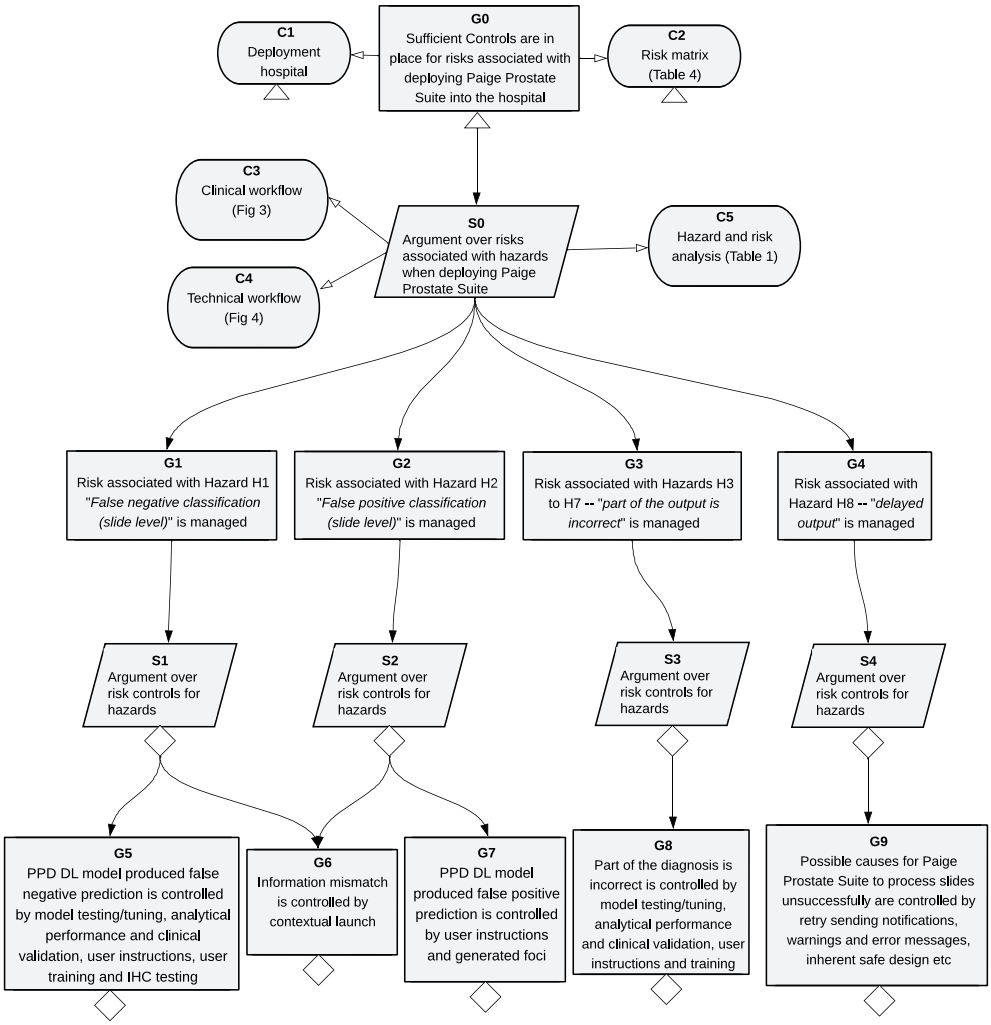


Fig. 12. Safety argument for deploying Paige Prostate Suite in the hospital.

Swiderska-Chadaj et al. [42] investigated the impact of scanner types and staining protocols on the performance of the DL model they produced in WSI classification of prostate cancer. They found that scanner variation only partly affects the performance of the DL model, whilst the staining protocols are more critical to the DL model performance. Other studies have investigated the error patterns that have been made by the DL models for prostate cancer detection. Mun et al. [43] has found that for the false negative cases, the DL model missed small-sized cancers which consisted of only several cancer glands or cancer glands located on the outer sample margin of the WSI, whilst for false positive cases, they often exhibited diffuse infiltration of lymphocytes and atrophic glands. Singhal et al. [44] found that misclassifications were occasionally discovered, particularly in the stromal regions and at the margins of the tissue borders. The majority of tissue border misclassifications are caused by preparation artefacts that the network does not recognise. They proposed to train an extra neural network to detect artefact regions and eliminate them as a pre-processing step to avoid such misclassifications. Arvaniti et al. [45] found that most misclassifications at the tissue borders are due to tissue preparation artefacts, which are not recognised by the DL model. They also propose to train an additional neural network to detect stromal and artefact regions and exclude them as a pre-processing step. This shows that there are many variables outside of this study that could affect the performance and potentially the safety impact of the DL models. In this study, the same Paige DL model is deployed across all three NHS sites without adjusting the operating point, unlike other systems which often calibrate their algorithms for each deployment site so that they are performant.

There is considerable value if a comprehensive set of systematic AI failure modes tailored to pathology applications can be established. Such failure modes could be used as a dictionary of prompts for evaluating DL-based medical devices. This does not mean that every system will have the same problems but it could help to discover potential issues in the design of the DL-based medical devices. Further, regulators could also use such information to ensure the analysis has addressed common failure modes. We view this as complementary to our work as it can support a more granular analysis for false positive and false negative classifications, which could further help to achieve inherently safe design of the DL model or to identify risk controls.

6. Discussion

Obtaining regulatory approval for AI/ML-based medical devices marks a crucial milestone in ensuring their safety. However, it is imperative to recognise that regulatory approval is only the initial step in assuring safety. Even if a device receives clearance from regulators, real-world application may still reveal unforeseen risks or challenges that were not apparent during the approval process. This underscores the fact that safety remains an ongoing, lifelong process that requires continual evaluation and assurance, especially at the point of deployment where actual harm can occur. Through the course of this study, we have acquired several valuable insights which could

potentially contribute to safety assurance of such systems, especially in deployment.

First, the use of DL-based medical devices together with other tools during deployment presents a potential source of hazards that needs careful investigation. Moreover, different technical deployment workflows present different risks. In this study, the three NHS trust hospitals all used a different integration strategy. For example, one site had direct LIMS integration; while another site had specimen tracking system integration; and the third site's integration involved daily image exports from the local image management system that were directly uploaded to the cloud. This can further complicate the process for pathologists who not only need to access LIMS when making the final diagnosis but also navigate additional interfaces if the integration is not directly within LIMS, opening up the possibility of discrepancies in interactions (perhaps mis-associating patients and WSIs). This indicates that some integration strategies can be more hazardous than others and it reinforces the need for careful analysis of the DL-based medical devices in their technical as well as clinical workflow.

Second, whilst it is important to identify hazards in the deployment environment and the associated risks, understanding potential AI failure modes that could contribute to these hazards is equally, if not more, important in the context of a clinical decision support tool. Currently, all of the regulatory cleared AI/ML-based medical devices are clinical decision support tools in the sense they cannot produce a decision automatically without human oversight. Therefore, educating clinicians on AI failure modes will enable the clinicians to develop a deeper understanding of the tools they interact with. Thus, it will empower the clinicians to critically evaluate the AI-generated recommendations, enabling them to discern when to trust AI recommendations and when to exercise caution or seek additional opinions. For example, in this study, when cores were spaced very closely together in one slide, the Paige system might over-annotate the tissue map which can lead to incorrect calculations of lengths of tumour. If the clinicians are equipped with such knowledge, they will know in this kind of scenario they should be more careful with the AI recommendations. Therefore, educating and training clinicians about potential AI failure modes for DL-based medical devices is paramount to foster more effective human-AI teaming to ultimately achieve better diagnostic performance and deliver safer patient care. Moreover, imparting this knowledge encourages a culture of transparency in healthcare, fostering collaboration between clinicians and AI developers to continuously improve AI algorithms, although when the AI algorithm has been updated, it has to undergo a new review based on the current health regulations, and clinicians need to be appraised of new behaviours and changes in failure modes as well.

Third, when assessing the probability for Hazard H1 "False negative classification" and H2 "False positive classification", it is noted that both are determined to be low in Table 1. However, it is observed that the probability for H2 is higher than that for H1. This reflects the design choice for the Paige system to prioritise sensitivity over specificity. As a result, the Paige system sometimes marks areas of tissue with Atypical Small Acinar Proliferation (ASAP) as being suspicious for cancer as well despite that PPD was trained only to distinguish adenocarcinoma from benign tissue, i.e. ASAP was not used in training PPD. ASAP, although not diagnostic of cancer itself, is considered as a marker for heightened risk of developing prostate cancer. Further, some studies have found that pathologists are more likely to make false negative errors than false positive ones [46,47]. This shows the merit to enhance the sensitivity of the DL-based medical device to support pathologists in order to avoid missing cancer. Again, for a clinical decision support tool to realise its full potential, it is important to understand what kind of errors human pathologists are likely to make, so the tool can be tuned to best support the pathologists to achieve better human-AI teaming diagnostic performance. Vice versa, it is also important to understand what kind of errors that AI might make, so that humans learn to be more cautious in such scenarios as we discussed above. However, it inevitably leads to an increase in false positive rates in order to avoid missing cancer. Therefore, clinicians utilising the Paige system may encounter a higher volume of false positive results, necessitating additional scrutiny, resources, and time for follow-up procedures such as repeat biopsies and patient consultations.

Fourth, in biomedical image analysis, the current ML performance metrics are not necessarily reflecting the domain of interest, and thus can hinder translation of ML techniques into practice. Currently for multiple class classification, weighted kappa score [48] is often used when reporting the accuracy of the DL model. However, if Gleason pattern 3 classified as 4 (and vice versa) is potentially more harmful than a Gleason pattern 4 classified as 5, then the kappa score might not be sufficient to report the real performance - indeed it may be misleading in terms of the safety of the model. Thus, standard ML performance metrics might not always reflect the true clinical safety, consequently they should not be relied upon as indicators for how successful a system will be when applied in practice. Rather, refined or adjusted performance metrics that can reflect the real clinical safety situations should be used and the reasons should be documented in the safety case. This also shows that collaboration between ML developers, safety engineers and domain experts (pathologists in this case) is necessary for effective and safe deployment of DL tools in healthcare.

Finally, it became clear that introducing DL-based medical devices into hospitals presents a challenge to, and a significant demand for, specialist IT support. Unlike conventional medical equipment, DL-based systems require sophisticated digital infrastructure, extensive data management, and seamless interoperability with existing hospital systems. For example, in order to use the Paige Prostate Suite, it is necessary to have a digital pathology solution, e.g. a scanner, to process the glass slides to produce WSI. Achieving this integration requires substantial investment in IT infrastructure, including hardware upgrades, network optimisation, and cybersecurity measures to safeguard sensitive patient data, exacerbating the strain on scarce healthcare IT resources. Therefore, there is a need for strategic planning and resource allocation in hospitals to facilitate successful integration and to maximise the potential benefits of DL-based medical devices. Fortunately, in this case, the hospital has their own IT specialists who addressed the above issues by developing extra tools, such as contextual launch, to integrate the Paige system smoothly into their existing workflow.

6.1. Limitation and future work

The work reported here is constrained by lack of operational data. Although the hazard and risk analysis is systematic, it is primarily qualitative as it has not been possible to calculate the exact frequency of occurrence of events in practice to validate the results at this stage. Further, the probability assigned to each hazard reflects the likelihood of the hazard itself occurring, i.e. how the Paige system behaves, rather than the likelihood of its consequences, which depends on how effective clinicians are as a risk control in this context. Consequently, we are unable to fully evaluate the effectiveness of the human-AI teaming. However, we view our work as an important first step in the safety assessment of human-AI interactions by fully establishing and characterising the risk profile of the AI system, which has been the focus of this study.

Thus, future work can build on this study and explore two key areas:

- **Designing a comprehensive monitoring framework for AI-based tools**. Safety assurance of the AI system is a continuous and ongoing effort. Thus, having a monitoring framework is important for maintaining the safety and effectiveness of the AI-based tools. Our view is that addressing this issue will require a multi-faceted approach that considers both technological and clinical validation perspectives. A robust framework should integrate automated technical monitoring – such as detecting data drift and model performance degradation. For example, when the distribution

of new data significantly deviates from the training data and impacts the model performance, it may be necessary to update the AI system. Further, it is also important to systematically collect and monitor data related to each hazard associated with the AI system, e.g. false negative, false positive and incorrect localisation of the focus, and the consequences of the hazards, which means clinical validations will be necessary to monitor all outputs from the AI system and also the final decisions made by pathologists. In cases of human-AI disagreement, documenting the reason behind the pathologists' differing decision would also be valuable to understand the limitations of the AI system.

Furthermore, it is also important to establish a structured process to analyse collected data in order to support continuous improvement. This will include using the understanding of hazards consequences to update the hazard log, to inform updates of the AI system, to introduce new risk controls when necessary and to support clinicians' training. These changes need to be considered holistically. For example, significant guideline changes may result in adjustments to the AI-based decision support tool being required. Further, continuous monitoring needs to be backed up with periodical reviews to implement a feedback integration mechanism to use insights effectively.

- **Understanding the dynamics of human-AI interaction**. In this study, we have not explored these issues as the Paige system is still at an early stage of clinical use. However, it is important to assess the risks associated with human decision-making when supported by AI systems. Some hazards may be mitigated by the human users correcting the system, but other hazards may be introduced due to under- and over-reliance on AI tools. For example, over time pathologists may gradually develop implicit trust in AI systems, which could undermine their role as an effective risk control as repeated confirmations of AI predictions might reduce their critical scrutiny. Conversely, if the pathologists perceive the system as limited, they may frequently disregard its outputs even when it is correct. This shows the complexity of human-AI interaction in healthcare and highlights a crucial potential hazard associated with the introduction of AI tools.

  To ensure pathologists continue to perform their independent assessment, it is important to design the system in a way to encourage pathologists to exercise their independent assessment. One practical example in this study is a feature in Paige system that requires pathologists to explicitly click the "AI button" in the interface to reveal the AI prediction otherwise the AI prediction remains hidden. Further, it is also important to maintain a comprehensive log of diagnostic outputs from both human and AI and conduct regular detailed analysis of these data. Thus, establishing a structured process for periodic clinical audits to compare pathologists' diagnostic outputs with AI outputs will be necessary. We understand that best practice recommendations for pathologist validation and ongoing audit with use of AI is likely to be produced by the Royal College of Pathologists in the future. Beyond that, nationwide surveillance schemes may need to be developed, with clinical validations conducted for extending product use cases, evaluating/ comparing new AI offerings when workflow components are altered, etc. Significant planning will be required to develop a suitable nationwide surveillance programme, to ensure consistent and safe reporting for AI products across multiple vendors.

  To contextualise any future work, we hypothesise that (i) the issues relating to over- and under-reliance on AI will be very specific to individual behaviours and preferences, (ii) this needs to be assessed in actual clinical settings to reflect real-world risk, and not in a retrospective simulated setting; and (iii) we would not have a robust baseline on the complexity and risks of human decision-making in the absence of AI for comparison. While a very important area of research, we anticipate that it will be quite complex to fully assess and measure these factors in any future studies and that it will be critically dependent on having effective monitoring frameworks.

In summary, the stand-alone performance of the AI system, the dynamics of human-AI interaction and the clinical consequences of the system's use can evolve over time. Ensuring the safety of AI-based systems is therefore a continuous process, requiring regular evaluations, particularly at critical points like deployment and system updates. Addressing through-life safety will be an important aspect of any future work in this field.

7. Conclusion

In this paper, we investigated how to bridge the gap from regulatory approval of DL models to their safe deployment in the clinical context. We presented a systematic method to show how to assure the safety of the DL-based system in deployment taking into account the clinical workflow. We applied our methodology to the use case of deploying the Paige Prostate Suite in the UK. In doing so, we identified new hazards which arise from the deployment, which cannot be identified in the development, and hence would not be addressed by regulatory approvals. Further, we have also summarised the insights we gained from this study, which should be valuable for developing guidelines for the deployment of such technology in healthcare. This work is timely and should not only inform the safety community but also influence the regulatory and ML communities. New regulatory requirements need to reflect these insights, e.g. there needs to be a two-stage regulatory process with the second stage focusing on the specifics of the deployment context, as explored here.

In conclusion, the safety assurance of real-world DL-based decision-support tools demands a joint effort from ML developers, safety experts, healthcare professionals, and regulatory bodies. By fostering collaboration, we can facilitate the introduction of AI technology that enhances diagnostic capabilities while prioritising patient safety. This work contributes to this aim, encouraging a responsible trajectory for the safe integration of AI in healthcare.

CRediT authorship contribution statement

Yan Jia: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Clare Verrill:** Writing – review & editing, Resources, Formal analysis. **Kieron White:** Writing – review & editing, Visualization. **Monica Dolton:** Writing – review & editing, Project administration. **Margaret Horton:** Writing – review & editing, Resources, Formal analysis. **Mufaddal Jafferji:** Writing – review & editing. **Ibrahim Habli:** Writing – review & editing, Visualization, Methodology, Formal analysis.

Declaration of competing interest

Prof Clare Verrill is the principal investigator of the study (ArticulatePro) evaluating Paige Prostate. University of Oxford and Oxford University Hospitals NHS Foundation Trust were part of the PathLAKE consortium. PathLAKE has received in-kind industry investment from Philips. Margaret Horton reports she was employed by Paige AI at the time. Mufaddal Jafferji reports he is employed by Paige AI. The remaining authors declare no competing interest.

Acknowledgements

ARTICULATE PRO is a two-year project funded by the Department of Health and Social Care (DHSC), in collaboration with the National Institute for Health and Care Research (NIHR) and the Accelerated Access Collaborative (AAC) through the Artificial Intelligence in Health and Care Award. The project was also overseen by DHSC

(AI_AWARD02269). The project's remit is to investigate the deployment of AI in the prostate cancer pathway by using Paige Prostate to assist pathologists when reading prostate biopsies. Dr Yan Jia and Prof Ibrahim Habli are supported by the Center for Assuring Autonomy funded by the Lloyd's Register Foundation and University of York. Prof Clare Verrill is supported by the NIHR Oxford Biomedical Research Center (BRC4 -SITE; sub-theme 5 - Pre-operative optimisation & enhanced recovery). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the DHSC.

References

[1] Prostate Cancer U.K., About prostate cancer, 2024, https://prostatecanceruk.org/prostate-information-and-support/risk-and-symptoms/about-prostate-cancer. (Accessed: 03 June 2024).

[2] American Cancer Society, Key statistics for prostate cancer, 2024, https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html. (Accessed: 03 June 2024).

[3] World Cancer Research Fund International, Prostate cancer statistics, 2024, https://www.wcrf.org/cancer-trends/prostate-cancer-statistics/. (Accessed: 03 June 2024).

[4] J. Gordetsky, J. Epstein, Grading of prostatic adenocarcinoma: current state and prognostic implications, Diagn. Pathol. 11 (2016) 1–8.

[5] J.R. Srigley, B. Delahunt, H. Samaratunga, A. Billis, L. Cheng, D. Clouston, A. Evans, B. Furusato, J. Kench, K. Leite, et al., Controversial issues in Gleason and International Society of Urological Pathology (ISUP) prostate cancer grading: proposed recommendations for international implementation, Pathology 51 (5) (2019) 463–473.

[6] R.N. Flach, P.-P.M. Willemse, B.B. Suelmann, I.A. Deckers, T.N. Jonges, C. van Dooijeweert, P.J. van Diest, R.P. Meijer, Significant inter-and intralaboratory variation in Gleason grading of prostate cancer: a nationwide study of 35,258 patients in the Netherlands, Cancers 13 (21) (2021) 5378.

[7] G.J. Van Leenders, T.H. Van Der Kwast, D.J. Grignon, A.J. Evans, G. Kristiansen, C.F. Kweldam, G. Litjens, J.K. McKenney, J. Melamed, N. Mottet, et al., The 2019 International Society of Urological Pathology (ISUP) consensus conference on grading of prostatic carcinoma, Am. J. Surg. Pathol. 44 (8) (2020) e87–e99.

[8] Paige, https://paige.ai. (Accessed: 03 June 2024).

[9] Nuffield Department of Surgical Sciences, The articulate PRO project, 2024, https://www.nds.ox.ac.uk/research/verrill-pathology-group/articulate-pro. (Accessed: 03 June 2024).

[10] The Royal College of Pathologists, Standards and datasets for reporting cancers, dataset for histopathology reports for prostatic carcinoma, 2024, https://www.rcpath.org/static/8cc88604-2c8d-4df4-a99542df41c102af/G084-dataset-for-histopathology-reports-for-prostatic-carcinoma.pdf. (Accessed: 20 December 2024).

[11] Y. Niu, S. Foerster, M. Muders, The role of perineural invasion in prostate cancer and its prognostic significance, Cancers 14 (17) (2022) 4065.

[12] P. Ström, T. Nordström, B. Delahunt, H. Samaratunga, H. Grönberg, L. Egevad, M. Eklund, Prognostic value of perineural invasion in prostate needle biopsies: a population-based study of patients treated by radical prostatectomy, J. Clin. Pathol. 73 (10) (2020) 630–635.

[13] A. Sciarra, M. Maggi, A. Del Proposto, F.M. Magliocca, A. Ciardi, V. Panebianco, E. De Berardinis, S. Salciccia, G.B. Di Pierro, A. Gentilucci, et al., Impact of uni-or multifocal perineural invasion in prostate cancer at radical prostatectomy, Transl. Androl. Urol. 10 (1) (2021) 66.

[14] C. of American Pathologists, Protocol for the examination of specimens from patients with carcinoma of the prostate gland, 2017, https://documents.cap.org/protocols/cp-prostate-2017-v4020.pdf. (Accessed: 20 December 2024).

[15] International Organization for Standardisation, ISO 15189:2022 Medical laboratories — Requirements for quality and competence, 2022.

[16] The Royal College of Pathologists, Position statement from the Royal College of Pathologists (RCPath) on Digital Pathology and Artificial Intelligence (AI), 2023, https://www.rcpath.org/static/90e5e248-4ad3-4d61-8247223f9faffc80/RCPath-AI-position-statement-2022.pdf. (Accessed: 03 June 2024).

[17] College of American Pathologists, How to Validate AI Algorithms in Anatomic Pathology, 2024, https://www.cap.org/member-resources/clinical-informatics-resources/how-to-validate-ai-algorithms-in-anatomic-pathology. (Accessed: 03 June 2024).

[18] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K.J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature Med. 25 (8) (2019) 1301–1309.

[19] J.I. Epstein, An update of the Gleason grading system, J. Urol. 183 (2) (2010) 433–440.

[20] J.I. Epstein, Prostate cancer grading: a decade after the 2005 modified system, Mod. Pathol. 31 (2018) 47–63.

[21] J.I. Epstein, L. Egevad, M.B. Amin, B. Delahunt, J.R. Srigley, P.A. Humphrey, G. Committee, et al., The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system, Am. J. Surg. Pathol. 40 (2) (2016) 244–252.

[22] International Organization for Standardisation, ISO 14971:2019 Medical devices — Application of risk management to medical devices, 2019.

[23] T.A. Kletz, Hazop & Hazan: Identifying and Assessing Process Industry Hazards, CRC Press, 1999.

[24] I. Habli, Y. Jia, S. White, G. Gabriel, T. Lawton, M. Sujan, C. Tomsett, Development and piloting of a software tool to facilitate proactive hazard and risk analysis of Health Information Technology, Heal. Inform. J. 26 (1) (2020) 683–702.

[25] International Organization for Standardisation, ISO/TR 24971:2020 Medical devices — Guidance on the application of ISO 14971, 2020.

[26] R. Bloomfield, P. Bishop, Safety and assurance cases: Past, present and possible future–an adelard perspective, in: Making Systems Safer: Proceedings of the Eighteenth Safety-Critical Systems Symposium, Bristol, UK, 9-11th February 2010, Springer, 2009, pp. 51–67.

[27] NHS Digital, DCB0160: Clinical risk management: its Application in the Deployment and Use of health IT Systems, 2018.

[28] US Department of Health and Human Services, Health information privacy, 2024, https://www.hhs.gov/hipaa/index.html. (Accessed: 03 June 2024).

[29] International Organization for Standardisation, ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements, 2022.

[30] D.J. Pumfrey, The Principled Design of Computer System Safety Analyses (Ph.D. thesis), University of York, 1999.

[31] S. Perincheri, A.W. Levi, R. Celli, P. Gershkovich, D. Rimm, J.S. Morrow, B. Rothrock, P. Raciti, D. Klimstra, J. Sinard, An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy, Mod. Pathol. 34 (8) (2021) 1588–1595.

[32] National Institute for Health and Care Excellence (NICE), Prostate cancer: diagnosis and management, 2024, https://www.nice.org.uk/guidance/ng131/resources/prostate-cancer-diagnosis-and-management-pdf-66141714312133. (Accessed: 03 June 2024).

[33] P. Raciti, J. Sue, J.A. Retamero, R. Ceballos, R. Godrich, J.D. Kunz, A. Casson, D. Thiagarajan, Z. Ebrahimzadeh, J. Viret, et al., Clinical validation of artificial intelligence–augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection, Arch. Pathol. Lab. Med. 147 (10) (2023) 1178–1185.

[34] T. Kelly, R. Weaver, The goal structuring notation–a safety argument notation, in: Proceedings of the Dependable Systems and Networks 2004 Workshop on Assurance Cases, vol. 6, Citeseer Princeton, NJ, 2004.

[35] A.F. Voter, E. Meram, J.W. Garrett, J.Y. John-Paul, Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage, J. Am. Coll. Radiol. 18 (8) (2021) 1143–1152.

[36] S. Ebrahimian, M.K. Kalra, S. Agarwal, B.C. Bizzo, M. Elkholy, C. Wald, B. Allen, K.J. Dreyer, FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies, Academic Radiol. 29 (4) (2022) 559–566.

[37] U. Mahmood, A. Shukla-Dave, H.-P. Chan, K. Drukker, R.K. Samala, Q. Chen, D. Vergara, H. Greenspan, N. Petrick, B. Sahiner, et al., Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing, BJR| Artif. Intell. 1 (1) (2024) ubae003.

[38] R.K. Samala, K. Drukker, A. Shukla-Dave, H.-P. Chan, B. Sahiner, N. Petrick, H. Greenspan, U. Mahmood, R.M. Summers, G. Tourassi, et al., AI and machine learning in medical imaging: key points from development to translation, BJR| Artif. Intell. 1 (1) (2024) ubae006.

[39] Z. Huo, R.M. Summers, S. Paquerault, J. Lo, J. Hoffmeister, S.G. Armato III, M.T. Freedman, J. Lin, S.-C. Ben Lo, N. Petrick, et al., Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use a, Med. Phys. 40 (7) (2013) 077001.

[40] A.V.S. Neto, J.B. Camargo, J.R. Almeida, P.S. Cugnasca, Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work, IEEE Access 10 (2022) 130733–130770.

[41] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, I. Habli, Guidance on the assurance of machine learning in autonomous systems (AMLAS), 2021, arXiv preprint arXiv:2102.01564.

[42] Z. Swiderska-Chadaj, T. de Bel, L. Blanchet, A. Baidoshvili, D. Vossen, J. van der Laak, G. Litjens, Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer, Sci. Rep. 10 (1) (2020) 14398.

[43] Y. Mun, I. Paik, S.-J. Shin, T.-Y. Kwak, H. Chang, Yet another automated Gleason grading system (YAAGGS) by weakly supervised deep learning, Npj Digit. Med. 4 (1) (2021) 99.

[44] N. Singhal, S. Soni, S. Bonthu, N. Chattopadhyay, P. Samanta, U. Joshi, A. Jojera, T. Chharchhodawala, A. Agarwal, M. Desai, et al., A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies, Sci. Rep. 12 (1) (2022) 3383.

[45] E. Arvaniti, K.S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P.J. Wild, J.H. Rueschoff, M. Claassen, Automated Gleason grading of prostate cancer tissue microarrays via deep learning, Sci. Rep. 8 (1) (2018) 12054.

[46] M.F. Santana, L.C.L. Ferreira, Diagnostic errors in surgical pathology, J. Bras. de Patol. E Med. Lab. 53 (2017) 124–129.

[47] J.D. Oxley, C. Sen, Error rates in reporting prostatic core biopsies, Histopathology 58 (5) (2011) 759–765.

[48] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, 2020, arXiv preprint arXiv:2008.05756.