Technoleg Iechyd Cymru
Health Technology Wales

Evidence Appraisal Report [1]

Artificial Intelligence (AI) assisted review of prostate biopsies in the detection and diagnosis of prostate cancer

Appraisal summary

Why did Health Technology Wales (HTW) appraise this topic?

Prostate Cancer is the most common cancer in men in the UK and is diagnosed in over 2000 men every year in Wales. People with suspected prostate cancer usually undergo a blood test to assess levels of Prostate Specific Antigen (PSA). If this is raised, many are offered a multiparametric MRI (mpMRI) scan. If deemed appropriate following the mpMRI, patients undergo a prostate biopsy. The biopsy is converted into slides and stained with haematoxylin and eosin (H&E), following which the morphology is assessed by pathologists either digitally or under a microscope. The slides are then graded using the Gleason grading system and Grade Group and the results included in a pathology report for review by the patient's clinician and a multi-disciplinary team. Information including number of cores containing cancer and maximum length of the cancer is usually also provided in the pathology report. In some cases, where initial results are unclear, additional tests such as immunohistochemistry (IHC) or additional biopsies may be requested. The disease grade assigned to the biopsies is taken into account alongside other information such as PSA level and any spread of the disease, to assign an overall risk group. This information is used to determine which treatments are offered to the patient.

Biopsy assessment by pathologists requires expertise and can be open to human error and variability depending on level of experience. Artificial Intelligence (AI) technologies have been developed to assist in the review of H&E-stained prostate biopsies, with the aim of improving diagnostic accuracy and case review time. The AI highlights areas of interest to the pathologist, usually by colour coding areas of varying Gleason cell patterns and assessing their proportions, to assign Grades. The technologies can also pre-request additional tests such as IHC, if required.

The topic was submitted to HTW by a company producing AI technology.

What evidence did HTW find?

This report aims to identify and summarise evidence that addresses the following question: What is the clinical and cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer?

[1] Cyfieithu dogfennau HTW wedi'u cyhoeddi o'r Saesneg i'r Gymraeg
Translation of published technical HTW documents from English into Welsh

We identified one Medtech Innovation Briefing by NICE, eight primary studies, three ongoing studies and one unpublished report which compared pathologist review of slides to AI-assisted review in a paired read, crossover study design. AI assistance tended to improve sensitivity of diagnosis, without impact on specificity. Both inter-observer and intra-observer concordance was improved with the AI, suggesting addition of this technology could increase consistency of biopsy review both for cases reviewed by the same individual, and between pathologists across Wales. Case review time decreased overall, as did the number of additional tests being requested. Generally, pathologists seemed to like using the technology and found it useful. There were no reported results describing the impact of AI-assisted review of prostate biopsies (AIPB) on patient outcomes.

Due to a lack of applicable published evidence on the cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer, HTW developed an economic analysis to estimate cost effectiveness compared to pathologist review alone. The model comprises a short-term decision tree and lifetime (40 years) predictions of cost, quality of life and mortality to evaluate the cost-effectiveness of the two strategies.

The results of the economic analysis show that using AI-assistance is expected to increase costs by £207 per patient and provide an additional 0.02 quality adjusted life years compared to pathologist review alone, translating to an incremental cost-effectiveness ratio of £13,278. This is below the cost-effectiveness threshold of £20,000, and so using AI-assistance is deemed a cost-effective strategy. Results of the probabilistic sensitivity analysis suggest that using AI-assistance has a 69% probability of being cost effective.

Both the identified patient and public involvement (PPI) articles and the focus groups hosted by Velindre Cancer Centre on behalf of HTW indicated that patients understand the use of AI in healthcare. There was some level of transparency expected when it comes to being informed of the use of AI in tests and procedures, but this should not be too detailed to avoid additional information burden on patients. Only a small percentage of patients are disinterested in understanding how test results have been generated. Acceptance of AI in prostate biopsy was found to be directly related to an understanding of the practitioner's role, patients overall tended to show preference for AI that is controlled by practitioners, provided the practitioners retain ultimate responsibility for outcomes. With this reassurance, most patients are welcoming of the introduction of AI in prostate biopsy and are hopeful it would result in quicker and more accurate diagnosis.

What was the outcome of HTW's appraisal?

HTW is a national body working to improve quality of care in Wales. We collaborate with partners across health, social care, and industry to issue independent guidance that informs commissioning within Wales health and social care. We are supported by an Assessment Group, who ensure our work adheres to high standards of methodological and scientific rigour, and an Appraisal Panel, who consider evidence within the Welsh context and produce HTW guidance. More details on our appraisal process, the assessment group, and the appraisal panel can be found on the HTW website.

In this case, the HTW Assessment Group considered the evidence presented in this Evidence Appraisal Report (EAR057) and concluded there was sufficient evidence for the development of guidance. Please refer to the HTW website for full guidance details. Evidence Appraisal Report 057 follows below and provides full details for this topic. More comprehensive details of the HTW Guidance and HTW Appraisal Panel considerations can be found on the HTW website.

1. Purpose of the Evidence Appraisal Report

This report aims to identify and summarise evidence that addresses the following question: What is the clinical and cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer?

Evidence Appraisal Reports are based on rapid systematic literature searches, with the aim of identifying the best published evidence on the effectiveness and cost-effectiveness of health and social care technologies and models of care and support. Researchers critically evaluate this evidence. The draft Evidence Appraisal Report is reviewed by experts and by Health Technology Wales multidisciplinary advisory groups before publication.

2. Context

Prostate cancer develops when abnormal cells of the prostate gland divide and grow in an uncontrolled manner. It is the most common cancer in men in the UK and has been increasing in recent years. In Wales, 2261 men were diagnosed with the disease in 2020, with incidence peaking around 65-74 years of age (Public Health Wales 2023b). Although 591 people died from prostate cancer in Wales in 2021, survival rates are increasing with 97% of men who are diagnosed with prostate cancer in Wales surviving for five years or longer in 2016-20, after adjusting for age (Public Health Wales 2023b).

The National Institute for Health and Care Excellence (NICE) guideline NG131 describes the UK guidance for prostate cancer diagnosis and management (NICE 2021b). Tests for prostate cancer start with a digital rectal examination and a prostate specific antigen (PSA) blood test. If there are any results of concern, men undergo a scan, usually multiparametric Magnetic Resonance Imaging (mpMRI). Following the results of the scan, they may then be referred for biopsy if there are areas of concern. A biopsy is a small core of tissue taken from the suspicious area for further review. Prostate biopsies can be undertaken using several different methods, including ultrasound-guided and transperineal template, and there are usually 10-12 small pieces of tissue taken from various areas of the prostate. The biopsies are chemically processed and thinly sectioned, these thin slices are applied to slides and then stained by haematoxylin and eosin (H&E) before being assessed by a pathologist. Pathologists either use a conventional microscope or review a digital image of the slide on a computer, some use the methods in parallel. Tissue on the slides is identified as benign or cancerous. If tissue is identified as cancerous, it is graded using the Grade Group or, previously, the Gleason score (Section 4.2). Where diagnosis is uncertain, a second opinion is sought, and sometimes additional testing via immunohistochemistry (IHC) or further biopsies are requested. This information is then combined with other data such as Tumour/ Node/ Metastasis (TNM) score and PSA blood test level to allocate a Cambridge Prognostic Group (CPG) for the cancer and helps determine the treatment options given to the patient. The NHS England rapid diagnostic and research pathways handbook states that a biopsy should be performed within 9 days of GP referral, with a target of 5 days for pathology being reported (NHS England 2022). However according to Prostate Cancer UK, biopsy results can take up to two weeks, and sometimes longer, to be reported back to the patient (Prostate Cancer UK 2022).

Treatment for prostate cancer can include surgery, radiotherapy, hormone therapy and chemotherapy depending on the severity of the disease at diagnosis and the fitness of the patient.

3. Health technology

Grading of prostate biopsies by pathologists is open to human error and inter- and intra-reading variability due to the subjectivity of grading the cells visible on slides. Therefore, there have been several Artificial Intelligence (AI) tools developed to assist pathologists in the review of prostate biopsy slides via digitised whole slide images (WSIs). Use of WSIs is already commonplace in Wales, and most pathologists are validated to assess samples via this method, although it is not consistently utilised. The AI can differentiate between cancerous and benign tissue on H&E WSIs. They flag WSIs which require further review, indicate high versus low grade areas of tissue, and assist in the grading of WSIs and pre-ordering of IHC testing. It is hoped that the use of AI in assessment of prostate biopsies in addition to pathologist review will improve consistency, efficiency and accuracy of biopsy review, grading and diagnosis when compared to pathologist review alone. It is also hoped the AI will reduce how many requests are made for additional IHC tests and increase the number of cancers which are identified, which can reduce the need for additional biopsies and the risk of missing cancers.

AI is a term used to refer to computer systems capable of performing tasks that typically require human intelligence, such as learning, problem-solving and decision-making. AI systems are based on mathematical formulae and rules called algorithms. The algorithms allow computers to learn patterns from data and make decisions. This is known as machine learning. There are subsets of machine learning such as decision trees, clustering algorithms, linear regression and deep learning that employ diverse algorithms to make predictions, classify data or discover patterns. Deep learning is one of the more complex forms of machine learning and uses neural networks. Neural networks are algorithms layered together to create complex, computational models based on the structure and functioning of the human brain. Deep learning neural networks are commonly used in tasks like image or speech recognition (Chahal & Byrne 2020).

AI systems are taught to recognise patterns on training datasets and are then tested on a separate dataset. Typically, training datasets for medical AI technologies are labelled or supervised datasets. To avoid bias, training and testing datasets must be large, varied, representative of the population and condition of interest and labelling of the training dataset must be robust (Chahal & Byrne 2020). AI licensing and adoption are subject to similar guidelines and rules as other health technologies and are described in further detail in Section 4.1.

The three AI tools which have evidence identified in the literature search and are of focus within this review are all CE marked, and are Galen Prostate (Ibex Medical Analytics), Paige Prostate (Paige AI, Inc) and DeepDx Prostate (DeepBio). All are convolutional neural network (CNN) AI, trained on WSIs. Galen Prostate was trained on a dataset of over 549 H&E WSIs from Pathology Institute at Maccabi Healthcare Services' centralised laboratory (MegaLab) in Israel (Pantanowitz et al. 2020) and has since been validated on multiple datasets from across the world including 2201 WSIs at Betsi Cadwaladr University Health Board in Wales (Aslam & Heath 2023). It classifies WSIs into benign, cancer or borderline, creating cancer and Gleason pattern heatmaps. It calculates tumour and tissue lengths, as well as indicating non-cancer findings, and can automatically pre-order IHC testing according to results identified (Ibex Medical Analytics 2024). Paige Prostate was trained on multiple datasets increasing in size and validated on a dataset of 12,000 WSIs from the Memorial Sloan Kettering Cancer Centre and is described in detail by Campanella et al. (2019). It marks areas of suspicion and automatically grades WSIs according to Gleason patterns and grade (Paige AI Inc 2024). DeepDx was trained on 1133 WSIs and validated on 700 cases from two hospitals in Seoul, South Korea (Ryu et al. 2019). It works by detecting and localising areas of interest, showing coloured overlays based on Gleason patterns, quantifying the proportions of these to assist in scoring, and automatically measures tissue and tumour lengths (Deep Bio 2022).

These AI tools can be integrated into the clinical workflow in a few ways. They can be used as first-read, second-read or concurrent read tools. If used as first-read, the AI reviews and assesses the scan first, and the pathologist then checks the outputs to identify any errors, or sign it off as correct. If used as second-read, the pathologist reviews the biopsies digitally as usual, and the AI re-reviews the reports and flags any contradictions or other discrepancies. Concurrent use is a mix of these two, whereby the AI reviews the scans first, but the output is not seen by the pathologist who reviews as usual, with any discrepancies flagged by the AI.

The use of these AI tools is becoming more accepted and widespread in Wales and across the UK. Galen Prostate was funded for use in six health boards across Wales as part of the Welsh Governments Innovation Fund, supported via the Small Business Research Initiative (SBRI) Centre of Excellence (SBRI Centre of Excellence 2023). The initial phase of this has now completed and has been reported (Aslam & Heath 2023, Nicholson & Theunissen 2023). Paige Prostate has been used across three NHS Trusts in England in collaboration with the University of Oxford, after being awarded funding from the NHSx AI Health and Care Awards (University of Oxford 2021).

4. AI guidelines and prostate cancer grading

4.1 Guidelines

AI-assisted review of prostate biopsy is considered a digital health technology and was determined to be a Tier C technology according to the Evidence Standards Framework for Digital Health Technologies. Technologies within this classification provide information that will be used to aid treatment or diagnosis, to triage or identify early signs of a disease or condition or will be used to guide next diagnostics or next treatment interventions. For technologies of this classification, it is recommended that satisfactory evidence is produced to demonstrate effectiveness of the technology. This includes studies conducted in a setting like the UK health and care system, peer-reviewed studies, and prospective studies. Therefore, evidence to support the claimed benefits of AI-assisted review of prostate biopsies should include real-world evaluations of its clinical utility and include one or more high-quality studies that support the claimed benefits in a relevant setting, showing improvements in relevant outcomes. Similarly, appropriate assessment of the economics of AI-assisted review of prostate biopsy should be undertaken.

All AI technologies are required by law to have a UK Conformity Assessed (UKCA) or relevant CE mark prior to being put on the market. They also need to comply with the UK Data Protection Act during development (if personal data will be used at this time), and following implementation (NHS AI and Digital Regulations Service 2023).

There is also a recommendation from the NHS AI and Digital Regulations Service that AI technologies undergo in-house validation prior to use (NHS AI and Digital Regulations Service 2023). This ensures they are properly integrated and demonstrated via a small pilot test before being used in the real-world in the clinical workflow, potentially having impact on diagnosis decisions.

4.2 Prostate cancer grading

There are several different methods of grading prostate biopsies and prostate cancer. Those described below are the most commonly used in the UK and/ or are referred to within the papers included in this review.

4.2.1 Gleason Score (GS)

The GS was commonly used to grade prostate biopsies, by grading patterns of cells and how they look compared to normal cells. There are five patterns, where 1 is like normal tissue and 5 has the greatest difference to normal tissues and/or appears to be more aggressive. Pathologists allocate and add two scores; the most common grade seen, and the highest grade seen. Examples of GS are 3+4=7 where grade 3 is most common but grade 4 is present, or 4+3=7 where grade 4 is commonly seen and is the highest grade present but there is some grade 3 identified. An overall score of between 6 and 10 means cancer is present; 6 is a slow growing cancer, 7 is intermediate grade, and scores between 8 and 10 indicate high grade cancer which is likely to grow quickly (Prostate Cancer UK 2023).

4.2.2 Grade Group (GG)

A GG is assigned based on the GS and is between 1 and 5 depending on the likelihood of the cancer growing and spreading. GG1 is a GS of 6, GG2 is a GS of 7 (3+4), GG3 is a GS of 7 (4+3), GG4 is a GS of 8, and GG5 is a GS of 9 or 10 (Prostate Cancer UK 2023). This may also be referred to as 'Gleason Grade Group'.

5. Clinical effectiveness

We searched for evidence that could be used to answer the review question: What is the clinical and cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer?

For details on the methodology used to identify evidence for this report, refer to Section 12.

5.1 Overview

The evidence identification criteria and research question is available in Appendix 1, and the PRISMA flowchart summarising study selection is available in Appendix 3. There was one NICE Medtech Innovation Briefing (MIB280) identified, along with eight primary studies, three ongoing studies, and one unpublished report. The studies are described in Appendix 4 and below. There was no evidence identified on the outcomes: 1) resource use, 2) change in patient management, 3) overall survival, 4) progression-free survival or 5) health related quality of life.

MIB280 reviewed evidence related to Paige Prostate, and was published in 2021 (NICE, 2021a). They included five observational studies, one of which is also included in this review (da Silva et al. 2021), across 3,444 WSIs. In general, they note that Paige Prostate may be an effective addition to standard care, that it can increase sensitivity and provide more efficiency to support high caseload demand in this area. Key uncertainties noted that studies were mainly retrospective, were not based in the UK, and only two reported statistical significance of using Paige Prostate when compared to standard care. The expert review included in the MIB was positive, noting both detection and efficiency improvements but highlighting the need for training for pathologists, and a need for prospective data. NICE also note that further evidence is needed in order to quantify potential real-world cost savings and the impact of Paige Prostate on the system.

Three included primary studies assessed Paige Prostate. da Silva et al. (2021) obtained 682 slides from 100 patients, half of whom had been diagnosed with prostate cancer. The slides were reviewed by pathologists and then digitised and reviewed by the AI separately, the results of which constitute the majority of the paper. However, following this and of relevance to this review,

an exploratory analysis was carried out whereby pathologists re-analysed images in conjunction with Paige Prostate. Outcomes reported from this exploratory analysis included sensitivity, specificity and case review time. Ground truth was defined as agreed diagnosis between pathologists and AI, or where there was a disagreement, the pathologists agreed diagnosis following additional review of IHC results. Eloy et al. (2023) obtained 105 slides from 41 patients. All slides were digitised and read twice by pathologists, once unaided and once aided by Paige Prostate after a washout period of at least two weeks. Outcomes reported included case review time (from opening the WSI to recording results), final diagnosis, grade group, requests for IHC and second opinion, and agreement with the AI. Ground truth was defined as agreement between the pathologists or if there was a disagreement, the consensus between two additional pathologists with a common WSI session. Pathologist experience ranged from 2 years to 12 years. Raciti et al. (2023) obtained 610 slides and digitised them. Sixteen pathologists were involved in the study, and all received training prior to assessing WSIs using Paige Prostate. Each pathologist reviewed every case twice sequentially, with the first read being unassisted and the second read, assisted by AI, taking place immediately afterwards to assess the AI as a second-read device. Pathologists would classify the WSI as suspicious for cancer, benign, or requiring further information. Ground truth was taken to be the original diagnosis.

One study reviewed DeepDx (Jung et al. 2022). Initially, 593 WSIs were reviewed separately by the pathologists and AI, the results for which are not noted in this report. Following this, and of relevance to this review, one pathologist blindly reviewed the WSIs without any AI assistance, then then assessed again with AI assistance four weeks later following randomisation of cases to ensure they were not in the same order. Reported outcomes included sensitivity, specificity and case review time.

Five of the included primary studies and the unpublished report reviewed Galen Prostate, the technology originally suggested by the topic proposer. Three of the primary studies were conference abstracts (Comperat et al. 2021a, Raoux et al. 2021a, Borkowski et al. 2024), and one was a brief report (Aslam & Heath 2023). No full-text study publications meeting our inclusion criteria were identified for this technology. The unpublished piece of work (Nicholson & Theunissen 2023) reported on the implementation of Galen Prostate across Wales as part of the SBRI, as described previously. The presentation given by Dr Comperat contains more data than the abstract, and is used as the key reference throughout this document (Comperat et al. 2021b). The presentation given by Dr Raoux is also used in this document where data is not provided in the abstract (Raoux et al. 2021b). Given that Galen Prostate is the system that has been used in NHS Wales we considered that this technology would be of interest to stakeholders. Both the studies reported by Comperat et al. (2021a) and Raoux et al. (2021a) had two arms, in which the control arm was slide review under a microscope, and the intervention arm was AI-assisted review of WSIs. Within the study undertaken by Comperat et al. (2021a), 785 WSIs were reviewed and the method used was randomised between pathologists. Outcomes reported include sensitivity and specificity. Raoux et al. (2021a) reviewed 1,231 WSIs and outcomes reported include case review time. Borkowski et al. (2024) undertook a two-arm double read study, where 4366 slides from 180 patients were reviewed once by pathologists alone, and once by pathologists assisted by AI. Outcomes included case review time and number of IHC requests. Aslam & Heath (2023) reported on the implementation of Galen Prostate at Betsi Cadwaladr University Healthboard as part of the SBRI scheme. They performed a qualitative review on the impact of the AI on pathologist confidence using the AI and their overall opinions of it. Nicholson & Theunissen (2023) included some data which was reported by Aslam & Heath (2023), Comperat et al. (2021a) and Raoux et al. (2021a) but additionally reports the results of a user feedback survey across 14 pathologists working across three different healthboards in Wales and of varying pathology experience, which have been included here.

The prevalence of prostate cancer in WSIs was 30 to 37% in the included studies except for Jung et al. (2022) which had a much higher prevalence of 78%.

5.2 Sensitivity

Five of the identified studies reported sensitivity outcomes, three reported data for Paige Prostate (da Silva et al. 2021, Eloy et al. 2023, Raciti et al. 2023), one for DeepDx (Jung et al. 2022) and two for Galen Prostate (Comperat et al. 2021b). Overall, diagnostic sensitivity was reported to increase with the use of AI technologies when compared to pathologist review alone (Table 1).

Two of the papers reporting results for Paige Prostate showed an increase in sensitivity, only one of which was statistically significant, and one showed a slight decrease. Eloy et al. (2023) reported a decrease in sensitivity from 96.8% in the control arm to 95.5% in the intervention arm, with no p value reported. da Silva et al. (2021) reported an increase in sensitivity once AI was introduced, from 93.7% in the control arm to 96.6% in the intervention arm at the WSI level, $p = 0.307$. A similar increase was seen at the patient level, from 94% to 96%, $p = 1.000$. Raciti et al. (2023) also reported an increase in sensitivity following introduction of AI of 8%, from 88.7% to 96.6% with a multi-reader, multi-case (MRMC) 95% confidence interval (CI) of an increase of 4.5% to 11.5%, $p < 0.001$. Raciti et al. (2023) found increases in sensitivity for all GG, with all being over 98% as shown in Table 1, ($p < 0.001$), but the impact on sensitivity did not vary depending on GG ($p = 0.13$). Raciti et al. (2023) also reported that 8.2% of paired reads differed between the control and intervention arms. All of those which were initially incorrect in the control arm and changed to be correct following intervention (341 paired reads) were Paige Prostate driven, 255 of which were malignant and 86 were benign. They noted that this resulted in accuracy gains of 3.5%, all of which could be attributed to Paige Prostate results (i.e. pathologist alone was incorrect, and became correct or deferred to assessment of other test results with AI assistance).

Use of Galen Prostate improved sensitivity in the intervention arm for cancer detection, which increased from 91.6% in control arm to 94.9% following intervention, and cancer grading, which improved from 72.9% to 74.7% (Comperat et al. 2021b). No p-values were reported for this.

In contrast to the results for other AI technologies, use of DeepDx reduced sensitivity from 99.6% in the control arm to 98.5% in the intervention arm, but again, no p-value was reported (Jung et al. 2022). The authors note that this was due to an increased false-negative count from two for the pathologists alone, to six with the AI involvement, all of which related to three malignancies (out of the 52 misinterpreted cases). Two of these malignancies were Gleason 6 (3+3) and one was Gleason 8 (4+4), but all had unusual appearance of bland or poorly formed tumour glands.

We carried out our own pooled analysis of sensitivity which indicated that AI-assistance improved sensitivity for cancer detection by around 4.4% from 91.6% to 96%, p=0.001 (Table 2).

5.3 Specificity

Five of the included studies reported specificity; three reported results for Paige Prostate (da Silva et al. 2021, Eloy et al. 2023, Raciti et al. 2023), and one for Galen Prostate (Comperat et al. 2021b). Specificity data for DeepDx was reported within a figure of the main text, and the supplementary materials of Jung et al. (2022). Overall, specificity was unchanged following the introduction of AI assistance in prostate biopsy review, with some studies showing a reduction in specificity (Table 1).

The addition of Paige Prostate had variable impact on the specificity, according to identified literature. da Silva et al. (2021) reported a slight reduction at the WSI level, from 99.8% to 97.8%

(p = 0.173) and a larger reduction at the patient level, from 98% to 92% (p = 0.739). The authors noted that none of the patients would have had an incorrect malignant diagnosis reported, and the decreased specificity was due to more cases being referred to IHC and other supplementary histopathological evaluation rather than missed diagnoses. Similarly, Eloy et al. (2023) also reported a reduction from 93.9% in the control arm to 92.8% in intervention, with no p-value reported. The only Paige Prostate study reporting an increase in specificity was Raciti et al. (2023), who reported an improvement from 97.3% to 98% in the intervention arm, an increase of 0.7 percentage points (p = 0.02). Improvements were seen for both specialist genitourinary pathologists (0.3% gain) and non-specialist pathologists (0.7% gain), but only reached statistical significance in non-specialists (p = 0.04). They also reported that Paige Prostate drove changes in 85.2% of discordant paired reads which were initially correct but changed to incorrect following intervention (46 out of 54), 38 of these 46 changed results had a ground truth of benign. Raciti et al. (2023) noted that there were accuracy losses of -0.5%, with loss of 0.08% attributed to Paige Prostate, adding that in benign cases that were then incorrectly changed to cancerous or defer, the AI had identified false positive foci sometimes showing benign mimics of cancer, or had falsely identified cancerous lesions as benign.

The use of DeepDx increased specificity from 83.9% in the control arm, to 97.7% in the intervention arm (Jung et al. 2022), and no p-value was reported.

The use of Galen Prostate did not make any difference to cancer detection specificity- with it remaining at 97.3% in both the control and intervention arms (Comperat et al. 2021b). Cancer grading specificity did improve slightly with the addition of Galen Prostate to 98.9% in the intervention arm, compared to 98.2% in the control arm (Comperat et al. 2021b).

We carried out our own pooled analysis of specificity which indicated that AI-assistance did not appear to affect specificity. Pooled analysis showed a change of -0.4% following the introduction of AI, from 97.8% in the control arm to 97.4% in the intervention arm, p=0.447 (Table 2).

5.4 Concordance

Five studies reported concordance; three for Paige Prostate (da Silva et al. 2021, Eloy et al. 2023, Raciti et al. 2023), one for DeepDx (Jung et al. 2022), and one for Galen Prostate (Comperat et al. 2021b). Overall, AI was reported to improve both interpersonal and intrapersonal concordance (Table 3).

Eloy et al. (2023) reported concordance with ground truth, as well as inter- and intra-observer concordance for Paige Prostate. As per the review protocol, only inter- and intra-observer concordance is reported here. For overall diagnostic concordance, linear kappa statistics were calculated. In the control arm, there was a mean interobserver concordance rate of 94.13%. Following the intervention, this was similar, at 93.02%. The intra-observer concordance pre- and post-AI was 98.81%. They also reported concordance for GG assignment, using quadratic weighted kappa statistics. The control arm had a mean interobserver concordance for GG of 73.39%. This remained similar in the intervention arm, at 72.03%, indicating concordance for GG was more difficult to obtain, even with AI assistance. The intra-observer concordance pre- and post-AI was 73.94%. Eloy et al. (2023) also reported that average total agreement with the AI. There was some variability between pathologists, with one pathologist only agreeing with the AI 56.19% of the time overall, and 7.69% of the time for cancer cases. Another pathologist agreed with the AI 79.05% of the time, and 56.41% of the time for cancer cases. Globally, there was 68.10% agreement with the AI, and this was significantly higher in benign cases than malignant cases (89.39% vs 32.05%, p < 0.001). They noted that most diagnostic discordant cases were distinguishing negative cases with atypical small acinar proliferation (ASAP) from small, but well differentiated, GG1 cancers or GG2 cancers with unusual presentation via review of additional tests such as IHC (10 out of 13

cases) or small GG2 cases with unusual presentation. They reported no difference in the detection of cribriform patterns, intraductal carcinoma or perineural invasion but did report a reduction in ASAP in the intervention arm, when compared to the control arm ($p < 0.001$).

Jung et al. (2022) reported both the linear and quadratic weighted kappa results for concordance of GG in the control and intervention arms. The quadratic weighted results only are reported here due to the differences in severity of GG impacting to varying extents on patient care and outcomes, and thus it being considered a more appropriate outcome measure. The control arm reported quadratic weighted kappa value increases, from 0.741 in the control arm to 0.925 in the intervention arm (Table 3), showing an improvement from substantial agreement to almost perfect agreement (Landis & Koch 1977).

The conference presentation by Comperat et al. (2021b) reported an agreement rate between pathologists in the control arm of 92.87%, and major discrepancy rate (defined as one which can impact clinical management, such as cancer v benign or GS 6 v GS 7) of 7.13%. In the intervention arm, the agreement rate was 95.16%, with a major discrepancy rate of 4.84%. Overall, the difference in major discrepancy rate between the two arms was -2.29% (95% CI -4.19% to -0.40%), in favour of the intervention. No p value was reported, and the ground truth for the discrepant samples was unclear. Borkowski et al. (2024) reported a linear kappa improvement in GG agreement from 0.896 for pathologists alone to 0.904 following the introduction of AI. Agreement with the adjudicator increased from 91.3% with pathologists alone, to 95.6% with the AI ($p > 0.05$). The authors also reported that differences between pathologists and adjudicators which were an over-diagnosis by one GG reduced from 7.47% to 2.89% and by two GG from 0.64% to 0.18% ($p > 0.05$) following introduction of AI. They also reported that under-diagnosis with a difference two GG reduced from 0.23% to 0.05%, and a difference of one GG increased from 0.37% to 1.24% ($p < 0.001$).

5.5 Case review time

Two studies reported per-slide or per-WSI case-review time comparisons both before and after introduction of AI assistance in prostate review, one for Paige Prostate (Eloy et al. 2023), and one for DeepDx (Jung et al. 2022). In all cases, review time was reduced by a statistically significant level ($p<0.001$) (Table 4). DeepDx reduced case time from 55.7 to 36.8 seconds and Paige Prostate reduced case time from 139 seconds to 108.5 seconds.

The turnaround time from first review by a pathologist to final review and sign-off for Galen Prostate was reported within three conference submissions (Raoux et al. 2021a, Comperat et al. 2021b). As the AI was able to pre-order IHC tests before pathologist review, the full turnaround time was from 1.8 days down to 9.4 minutes in both Raoux et al. (2021b) and Comperat et al. (2021b) (Table 4). Borkowski et al. (2024) reported a reduction in case review time of 58%, going from 50.6 minutes on average for pathologists alone, to 21 minutes with the introduction of AI ($p<0.01$). This finding was consistent across all diagnoses, with benign diagnoses reducing by 57% from an average of 42.2 minutes to 18.8 minutes, and malignant case review time reducing by 60% from 55.3 minutes to 22.7 minutes following the introduction of AI. The authors reported an increase in productivity from 1.2 cases to 2.9 cases reviewed on average per pathologist, per hour.

One study calculated the potential time-saving possible if Paige Prostate was used prior to pathologist review, with pathologists only reviewing WSIs deemed suspicious for cancer by the AI system (da Silva et al. 2021). Pathologist review of WSIs were found to take longer than glass slides (122 seconds v 98 seconds), resulting in an overall review time for all WSIs being higher than glass slides (19.6 hours vs 15.76 hours). By utilising the AI to classify WSIs prior to review, and requiring the pathologists to only review suspicious cases, only 200/579 WSIs required

additional review by the pathologist, resulting in a review time of 6.77 hours; a reduction of 65.5% from review of the full set of WSIs (Table 4). It should be noted, however, that any use of AI in healthcare should be clinically led (The Royal College of Pathologists (RCP), 2023). As such pathologists should be reviewing every case, whether deemed malignant or benign by the AI, and therefore this result should be interpreted with caution.

5.6 Additional tests

Two studies clearly reported the impact of AI on additional test ordering. Eloy et al. (2023) reported that the proportion of requests for both IHC testing and second opinions was statistically significantly reduced overall across all cases as well as in subgroup analysis splitting cancerous and benign tissue (Table 1) following use of Paige Prostate. In the control arm there were 193 IHC requests across all WSIs and pathologists (45.95% cases), this reduced to 153 requests (36.43% cases) in the intervention arm, a reduction of 9.52% ($p < 0.001$). Second opinions were requested in 51 (12.14% cases) in the control arm, which reduced to 31 (7.38% cases) in the intervention arm, an overall reduction of 4.76% ($p < 0.001$). Borkowski et al. (2024) reported a reduction in IHC overall of 37% following the introduction of AI, from 960 down to 604. There was the greatest proportional reduction seen for benign cases, where the number requested reduced by 59% from 312 to 127. Reductions were also seen for undetermined cases, which reduced by 29% from 487 to 344 following introduction of AI, and by 17% from 161 to 133 for malignant cases. However, it was unclear how many cases were included for each diagnosis.

Some expert reviewers stated that second opinion from another pathologist in sites where this is not routinely done for all cases is usually requested for complex cases. Others also noted that second opinions can be utilised routinely in some sites to cross check coding errors, rarely resulting in differing diagnoses. They noted that even with the introduction of AI, second reporting was likely to continue for complex cases, but where hospitals always require double-reporting, the AI may be able to replace second human review in straightforward cases.

Raoux et al. (2021a) reported IHC requests for Galen Prostate, noting 80% of cases in the control arm had IHC requested, and 0.6% of cases in the intervention arm had IHC manually requested in addition to those automatically ordered by the AI. However, they did not report the proportion of cases for which the AI had requested IHC testing in the intervention arm, so the overall change in proportion of cases for which IHC was ordered was not possible to determine (Raoux et al. 2021b).

No studies reported on the need for additional biopsies.

Expert reviewers of the EAR also noted that the heat maps generated by the AI which indicate areas with higher grade cancer could be utilised by the All Wales Genomic Service to better target areas for analysis.

5.7 User acceptability and usability

Expert review comments reported within NICE MIB280 (NICE, 2021a) indicated general support for the Paige Prostate system. They noted potential for reduced missed cancers, and other areas of suspicion which may benefit from further investigation. They also highlighted the potential to increase efficiencies and improve turnaround times, as well as standardise grading of cancer to improve consistency across the field. They noted the importance of training on the system and noted that there may be a learning curve following introduction of the system before the benefits and potential cost savings could be seen.

Aslam & Heath (2023) undertook some qualitative research following the use of Galen Prostate at Betsi Cadwaladr Health board. All pathologists involved in the study said they felt more confident when making a diagnosis using AI, compared to without the AI. They highlighted the benefits of the AI highlighting areas of interest in otherwise normal looking areas, which may otherwise have been missed. The AI was used as an 'aide memoir', assisting with standardisation of reporting of GS and other features. Additional data on pathologist views of Galen Prostate was also collected and reported by Nicholson & Theunissen (2023). They found that 85% pathologists were highly motivated to continue using Galen Prostate, with an average rating of 4.7 out of 5, and only one pathologist was not very motivated. The majority of pathologists also said Galen Prostate operating system was easy or very easy to use, with an average score of 4.5 out of 5 with one pathologist rating it neutrally and none providing a negative rating. Most pathologists also found it somewhat easy or very easy to triage cases using Galen Prostate, with three pathologists rating it neutrally and none negatively, resulting in an average score of 4.4 out of 5. The pathologists also found other features useful, such as the indication of likelihood of significant clinical findings (average score 4.8 out of 5), heatmaps (average score 4.6 out of 5) and the automated Gleason scoring (average score 4.1 out of 5). They also reported that 50% of surveyed pathologists reported a change in IHC ordering, the majority of which were a reduction. The surveyed pathologists also believed there was potential to increase diagnostic efficiency both for cancerous and benign cases (average score 4.5 out of 5 and 4.6 out of 5 respectively), and felt there was more consistency in case reporting (average score 3.7 out of 5).

No user acceptability or usability results were identified for DeepDx.

Table 1 – AI-assisted biopsy review (intervention) compared to pathologist alone review (control): outcomes

Outcome	Evidence source(s)	Technology	Number of slides/ WSIs	Absolute effect (95% CI)	Difference (95% CI) interpretation	Comments on reliability
Diagnostic accuracy						
Sensitivity	Comperat et al. (2021b)	Galen Prostate	785	*Detection* Control: 91.6% Intervention: 94.9%	Not reported ***Appears to favour AI***	• Not peer reviewed • No confidence intervals or p values reported
				Cancer grading Control: 72.9% Intervention: 74.7%	Not reported ***Appears to favour AI***	• Not peer reviewed
	da Silva et al. (2021)	Paige Prostate	661 part-specimens	Control: 93.7% (89 to 96.8) Intervention: 96.6% (92.7 to 98.7)	2.9% (0.0 to 8.5%) p = 0.307 ***Favours neither***	
			100 patients	Control: 94% (83.5 to 98.7) Intervention: 96% (86.3 to 99.5)	2.0% (0.0 to 13.0%) p = 1.000 ***Favours neither***	
	Eloy et al. (2023)	Paige Prostate	105	Control: 96.8% (range 92.3 to 100) Intervention: 95.5% (range 89.7 to 100)	Not reported ***Appears to favour control***	
	Jung et al. (2022)	DeepDx	593	Control: 99.6% (98.5 to 100) Intervention: 98.5% (96.9 to 99.4)	Not reported ***Appears to favour control***	
	Raciti et al. (2023)	Paige Prostate	All WSIs (610)	Control: 88.7% Intervention: 96.6%	8% (4.5% to 11.5%) p = 0.001 ***Favours AI***	• Intervention read immediately follows control read, with no washout time
			ASAP (15)	Control: 54.2% Intervention: 74.6%	Mode p < 0.001 Factor p < 0.001 Model/ Factor interaction p = 0.13	
			GG 1 (110)	Control: 89.8% Intervention: 98.1%		
			GG 2 (39)	Control: 95.4% Intervention: 99.2%		
			GG3 (10)	Control: 96.9% Intervention: 99.4%		
			GG4 (12)	Control: 90.6%		

Outcome	Evidence source(s)	Technology	Number of slides/ WSIs	Absolute effect (95% CI)	Difference (95% CI) interpretation	Comments on reliability
				Intervention: 99%		
			GG5 (4)	Control: 95.3% Intervention: 98.4%		
Specificity	Comperat et al. (2021b), Comperat et al. (2021a)	Galen Prostate	785	*Detection* Control: 97.3% Intervention: 97.3%	Not reported ***Appears to favour neither***	• Not peer reviewed
				Cancer grading Control: 98.2% Intervention: 98.9%	Not reported ***Appears to favour AI***	• Not peer reviewed
	da Silva et al. (2021)	Paige Prostate Paige Prostate	661 part-specimens	Control: 99.8% (98.6 to 100) Intervention: 97.8% (95.8 to 99)	-2.0% (-4.9 to 0.0%) p = 0.173 ***Favours neither***	
			100 patients	Control: 98% (89.4 to 99.9) Intervention: 92% (80.8 to 97.8)	-6.0% (-23% to 7.7%) P = 0.739 ***Favours neither***	
	Eloy et al. (2023)	Paige Prostate	105	Control: 93.9% (range 89.4 to 100) Intervention: 92.8% (range 87.9to 98.5)	Not reported ***Appears to favour control***	
	Jung et al. (2022)	DeepDx	593	Control: 83.9% (76.4 to 89.7) Intervention: 97.7% (93.4 to 99.5)	Not reported ***Appears to favour AI***	• Control arm specificity much lower than that in other studies, likely due to only one pathologist participating which may result in bias
	Raciti et al. (2023)	Paige Prostate	610	Control: 97.3% Intervention: 98%	0.7% (0.1% to 1.2%) p = 0.020 ***Favours AI***	
Additional tests						
IHC	Eloy et al. (2023)	Paige Prostate	105	Control: 193 / 45.95% Intervention: 153 / 36.43%	p < 0.001 ***Favours AI***	
	Raoux et al. (2021a)	Galen Prostate	1224	Control: 80% Intervention: 0.6% manual ordering on top of auto-ordered by AI	Not reported ***Unclear on whether AI or control is favoured***	• Not peer reviewed • Unclear what proportion of WSIs in intervention arm

Outcome	Evidence source(s)	Technology	Number of slides/ WSIs	Absolute effect (95% CI)	Difference (95% CI) interpretation	Comments on reliability
						required IHC ordering so unable to directly compare to control arm
	Borkowski et al. (2024)	Galen Prostate	4366	Control: 960 Intervention: 604	-37% ***Appears to favour AI***	• Unclear on number of WSIs/ parts, so proportional reduction cannot be calculated
Second reviewer	Eloy et al. (2023)	Paige Prostate	105	Control: 51 / 12.14% Intervention: 31 / 7.38%	p < 0.001 in table, p < 0.006 in text ***Favours AI***	• Discrepancies in p values reported, however both significant at 5% level

Abbreviations: AI = Artificial Intelligence, ASAP = Atypical small acinar proliferation; CI = Confidence interval; GG = Gleason Grade Group; WSI = whole slide images

Table 2 – AI-assisted biopsy review compared to pathologist alone review: summary diagnostic accuracy values for each technology

Technology	Evidence source(s)	Sensitivity (95% CI)			Specificity (95% CI)			Comments on reliability
		AI-assisted	Pathologist alone	Difference with AI-assisted	Pathologist + AI	Pathologist alone	Difference with AI-assisted	
Any AI*	Comperat et al. (2021b), da Silva et al. (2021), Eloy et al. (2023), Raciti et al. (2023)	96.0% (94.1 to 97.3)	91.6% (89.0 to 93.6)	+4.4%, p=0.001 ***Favours AI***	97.4% (95.0 to 98.6)	97.8% (95.8 to 98.9)	-0.4%, p=0.447 ***Favours neither***	• Two AI technologies combined, higher heterogeneity
Paige Prostate	da Silva et al. (2021), Eloy et al. (2023), Raciti et al. (2023)	96.7% (94.3 to 98.1)	91.6% (88.0 to 94.2)	+5.1%, p=0.002 ***Favours AI***	97.2% (93.3 to 98.9)	98.0% (95.0 to 99.2)	-0.8%, p=0.309 ***Favours neither***	

Abbreviations: AI = Artificial Intelligence, CI = Confidence interval
Footnote: *DeepDX was not included in the pooled analysis of any AI due to heterogeneity – Jung et al. (2022) reported much lower specificity for the pathologist alone than the other studies.

Table 3 – AI-assisted biopsy review (intervention) compared to pathologist alone review (control): concordance

Outcome	Evidence source(s)	Technology	Number of slides/ WSIs	Absolute effect (95% CI)	Relative effect [95% CI] (interpretation)	Comments on reliability
Diagnostic						
Inter-observer concordance	Eloy et al. (2023)	Paige Prostate	105	Control: 94.13% (range 90.48% to 98.10%; kappa range 0.802 to 0.961) Intervention: 93.02% (range 90.48% to 97.14; kappa range 0.802 to 0.942)	Not reported ***Appears to favour neither***	
Intra-observer concordance	Eloy et al. (2023)	Paige Prostate	105	98.81% (range 98.1% to 100%; kappa range 0.958 to 100)	Not reported ***Appears to favour neither***	• Concordance comparing control to intervention arm
Concordance with AI	Eloy et al. (2023)	Paige Prostate	105	Overall: 68.10% Benign cases: 89.39% Malignant cases: 32.05%	$p < 0.001$ ***Favours AI***	
	Comperat et al. (2021b)	Galen Prostate	785	**Agreement** Control: 92.87% Intervention: 95.16% **Major discrepancy** Control: 7.13% Intervention: 4.84%	**Major discrepancy** -2.29% (-4.19% to -0.40%) p value not reported ***Appears to favour AI***	• Not peer reviewed • Ground truth for discrepant findings unclear • Unclear as to whether 'major discrepancy' is between pathologists, pathologist and AI, or pathologist/AI and ground truth
Grade group						
Inter-observer concordance	Eloy et al. (2023)	Paige Prostate	105	Control: 73.39% (range 57.5% to 86.11%; kappa range 0.823 to 0.942) Intervention: 72.03% (range 64.71% to 80%; kappa range 0.760 to 0.938)	Not reported ***Appears to favour neither***	

Outcome	Evidence source(s)	Technology	Number of slides/ WSIs	Absolute effect (95% CI)	Relative effect [95% CI] (interpretation)	Comments on reliability
Intra-observer concordance	Eloy et al. (2023)	Paige Prostate	105	73.94% (range 63.42% to 84.09%; kappa range 0.830 to 0.954)	Not reported ***Appears to favour neither***	• Concordance from control to intervention arm
Concordance with AI	Jung et al. (2022)	DeepDx	539	**Linear kappa:** Control: 0.621 Intervention: 0.741 **Quadratic kappa:** Control: 0.876 Intervention: 0.925	Not reported ***Appears to favour AI***	• This study had a reporting error in the linear/ quadratic kappa results whereby the text mismatched the graph. Data from graph used here.
Agreement with adjudicator	Borkowski et al. (2024)	Galen Prostate	2183 parts	**Linear kappa:** Control: 0.896 Intervention: 0.904	Not reported ***Appears to favour AI***	• Not peer reviewed
Abbreviations: AI = Artificial Intelligence; WSI = whole slide images						

Table 4 – AI-assisted biopsy review compared to pathologist alone review: case review time

Outcome	Evidence source(s)	Technology	Number of slides/ WSIs	Absolute effect (95% CI)	Relative effect [95% CI] (interpretation)	Comments on reliability
Case review time	Eloy et al. (2023)	Paige Prostate	105	Median Control: 139s Intervention: 108.5s	Not reported ***Appears to favour AI***	• Per slide review time
	Jung et al. (2022)	DeepDx	539	Control: 55.7s Intervention: 36.8s	Not reported ***Appears to favour AI***	• Per slide review time • Not specified whether case report time included was mean or median
	Borkowski et al. (2024)	Galen Prostate	4366	Control: 50.6 mins Intervention: 21 mins	-58%, $p < 0.01$ ***Favours AI***	• Per case review time to diagnosis – unclear how many slides included per case
Case turnaround time	Comperat et al. (2021a), Raoux et al. (2021a)	Galen Prostate	785	Control: 1.8 days Intervention: 9.4 min	Not reported ***Appears to favour AI***	• Unclear on time between AI ordering IHC and review of slides, and whether control turnaround time also includes wait-time for IHC results • Unclear how many slides included per case • Not peer reviewed
Total case review time	da Silva et al. (2021)	Paige Prostate	579 overall, 200 marked 'suspicious' by AI	Control: 19.6hr (all WSI) Intervention: 6.77hr (only 'suspicious' slides)	Not reported ***Appears to favour AI***	• Only suspicious slides reviewed by pathologist. In practice, pathologist should still review all slides (RCP 2023) so total case review time in intervention arm likely underestimated • Unclear how many slides included per case

Abbreviations: AI = Artificial Intelligence; hr = hours; IHC = immunohistochemistry; min = minutes; s = seconds; WSI = whole slide images

5.8 Ongoing studies

There were no ongoing systematic reviews assessing AI-assisted review of prostate biopsies specifically identified. There were systematic reviews reviewing use of AI in WSI assessment identified, however these were generalised and not specific to prostate cancer or expected to report the outcomes of interest in this review, and as such have not been included.

Three ongoing primary studies which may be expected to complete in the next 6 to 12 months were identified and are described in Table 5.

Table 5 – Summary of ongoing primary studies

Study information	Status	Research question and outcome measures
Registration: ISRCTN14323711 CONFIDENT-P https://www.isrctn.com/ISRCTN14323711 Protocol published as Flach et al. (2023) **Country:** Netherlands **Target recruitment**: 80 participants	Complete Protocol notes study carried out 2022-2023, but intended publication date unclear	A study to assess the clinical implementation of AI assistance in pathology. Patients who undergo prostate biopsy for suspected cancer will be randomised in a 1:1 ratio between the control and intervention arm. Prospective RCT. **Population**: Patients with PCa who underwent prostate needle biopsies **Intervention:** AI-assisted workflow (Paige Prostate). IHC ordered if no cancer detected on slide or on request from pathologist in cases of doubt. **Comparator:** Pathologist assessment alone, IHC routinely performed on all cases. **Primary Outcome Measures:** Number of IHC stains performed. **Secondary Outcome Measures:** Sensitivity and specificity, case review time, number of AI stains that may have been omitted, pathologist evaluation on AI-assisted work process, cost savings.
Registration: ISRCTN91685765 ARTICULATE PRO https://www.isrctn.com/ISRCTN91685765 **Country:** England **Target recruitment:** 1500 participants	Ongoing, recruiting **Last updated:** 31/08/2023 **Intend to publish:** August 2024	Examining the impacts of pathologists using the assistance of computer technology (artificial intelligence software) on the diagnosis of prostate cancer biopsies. Prospective observational case series. **Population:** All patients undergoing prostate biopsy for suspected prostate cancer. **Intervention:** Paige Prostate. **Comparator:** Pathologist review alone. **Primary Outcome Measures:** Changes to diagnostic reports of biopsies, and the corresponding immediate changes in patients' clinical management, as measured by cancer/ASAP identification, Gleason grade 4 identification, tumour burden including length, number of involved cores, treatment pathway recommendation as determined by a MDT **Secondary Outcome Measures:** Case review time, additional tests requested (i.e. resource utilisation),

Study information	Status	Research question and outcome measures
		health economics, diagnostic discrepancies, experience of pathologists, urologists and patients measured by survey and qualitative methods
Registration: NCT05228197 Imperial Prostate 6 https://clinicaltrials.gov/study/NCT05228197 **Country:** England **Target recruitment:** 780 participants	Ongoing, recruiting **Last updated:** 29/03/2022 Intended to complete in 2023, but unclear if has yet commenced.	A Study to Assess the Clinical and Cost-effectiveness of the Galen Prostate Artificial Intelligence Histology System in Diagnosing Clinically Important Prostate Cancer on Prostate Biopsy Tissue. Cohort cross-sectional. **Population:** Patients with suspected prostate cancer who are undergoing or have undergone a prostate biopsy. **Intervention:** Galen Prostate. **Comparator:** Pathologist alone. **Primary Outcome Measure:** Sensitivity, Specificity, PPV, NPV (all on per patient basis), cost-consequence analysis, cost-utility analysis. **Secondary Outcome Measure:** Sensitivity, Specificity, PPV, NPV, AUC, agreement with pathology report for cancer length, Gleason Grade pattern 4, reported Gleason Grade/ Grade Group, cancer area (all on per slide/ patient basis), consent to data linkage, cost effectiveness.
Abbreviations: AI = Artificial intelligence; AUC = area under the curve; IHC = immunohistochemistry; MDT = multidisciplinary team meeting; NPV = negative predictive value; PCa = prostate cancer; PPV = positive predictive value; RCT = randomised controlled trial		

5.9 Certainty of the evidence

- AI development, validation and testing is usually performed on good quality slides. Some studies included within this review did appear to have some WSIs containing unusual morphology but further testing on poor quality slides/ artifacts is needed to determine whether the technologies work well in real-world settings. The expert reviewers noted that there are some histopathological features which are not recognised by AI and others which can be incorrectly identified as cancer. These are more 'unusual' features, which are noted as a potential cause of false positives or false negatives from the AI, within sections 5.2 and 5.3. Features which they say are usually not recognised by AI but can be important in diagnosis include:

    - Cribriform pattern 4
    - Necrosis
    - Extraprostatic extension

    And features which they say can be incorrectly diagnosed as cancer which may result in overdiagnosis include:

    - Benign Cowper's glands
    - ASAP

- No evidence on impact of technology on patient outcomes including overall survival, progression-free survival, or health related quality of life. Minimal impact on change in

patient management was reported, with only one study noting that the differences in the control and intervention arms would not have resulted in a patient receiving an alternative diagnosis (da Silva et al. 2021).

- Many outcomes were reported on a per-slide basis, however only some studies reported the final patient outcome and whether the AI would result in an improvement of overall diagnosis. This means it is unclear whether the addition of AI would result in more accurate cancer diagnoses or just additional identification of malignancy from slides within the same case which would have already had cancer diagnosed.
- There was some variability in the included studies on whether pathologists were trained in the use of AI prior to use, or not. It was unclear on whether this may influence results, however in the real-world setting, it would be expected that all pathologists are trained in the use of AI prior to using it.
- The location of identified studies were mostly international or the USA, so had varying levels of generalisability or applicability to Welsh context. There were some results from Wales for Galen Prostate (from SBRI), but these were not from a diagnostic accuracy study.
- NICE MIB states identified evidence for Paige Prostate was low to moderate quality, with most of it being retrospective. They note there is a requirement for sufficiently powered sample sizes across multiple pathology labs in the UK to show statistically significant clinical benefit compared to the current standard of care.
- There was very limited prospective comparative evidence for DeepDx and Galen Prostate, and most evidence identified for Galen Prostate was not peer reviewed, and was in the form of conference abstracts and presentations. The topic proposer notes work is ongoing for Galen Prostate, but it was not available in time for inclusion in this report.
- It is unclear for all AI technologies included in this report as to the possibility/ expected frequency of updates, the support which will be available to users, how the output of the AI will be monitored, and what patient data is collected and how this is used, if any.
- There is some uncertainty around the cost of training pathologists on the AI systems (although expert reviewers indicated this would be minimal) and maintaining competency, as well as any ongoing costs related to system support and maintenance, these should be considered.

6. Cost effectiveness

6.1 Economic literature review

We conducted a rapid systematic literature review to answer the following research question: What is the cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer compared to pathologist review alone? Appendix 3 summarises the selection of articles for inclusion in the evidence review. The titles and abstracts of 3,457 records identified in the search for this research question were screened but no studies were deemed relevant to the research question as there were no comparative cost studies.

6.2 HTW cost utility analysis

Due to the lack of applicable published evidence on the cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer, HTW developed an economic analysis to estimate cost effectiveness compared to pathologist review alone. The model replicated the cost-effectiveness model built by NICE for Guidance NG131 for the diagnosis and management of prostate cancer (NICE 2021b), however the model was adapted to include an initial decision tree to include the relevant review methods of prostate biopsy. Costs used in the analysis were inflated to 2022 GBP (£) where necessary, and future costs and benefits were discounted at an annual rate of 3.5%. The model comprises a short-term decision tree and lifetime (40 years) predictions of cost, quality of life and mortality to evaluate the cost-effectiveness of the following two strategies:

1. AI-assisted review of prostate biopsy in addition to a pathologist
2. Pathologist review of prostate biopsy alone

A cohort of 1,000 men with a baseline age of 68 who have undergone a prostate biopsy for suspicion of prostate cancer will have their biopsy slides reviewed by either a pathologist alone, or a pathologist in addition to AI assistance.

Costs of biopsy and pathology review have not been included in the analysis as costs will be equivalent across modelled arms. However, the cost of AI-assistance has been provided by two manufacturers and an average cost has been calculated as £41.32 per patient.

The sensitivity and specificity of both diagnostic strategies were sourced from the meta-analysis described in Section 12.1 and the values used in the model are provided in Table 6. As there were no statistically significant differences in specificity between the diagnostic strategies, the specificity for pathologist alone has been assumed for both arms in the model.

Table 6 – Sensitivity and Specificity

	AI-assisted review	Pathologist alone	P-value
Sensitivity	96.0%	91.6%	0.001
Specificity	97.4%*	97.8%	0.447
* Modelled as 97.8% in line with pathologist alone due to non-statistical significance.			

Following a diagnosis of prostate cancer, the disease is categorised as: low-, intermediate- or high-risk cancer, or metastatic disease. All people are assumed to initiate treatment once diagnosed as having cancer. Treatment costs applied in the model are provided in Table 7. Additional monitoring costs are applied per cycle. The probability of disease progression is captured every cycle.

Table 7 – Treatment costs

Disease state	Treatment costs
Low risk	£2,325.59
Intermediate risk	£2,650.14
High risk	£3,871.13
Metastatic disease	£14,752.58

People who are misdiagnosed as not having cancer continue to progress through the disease states, however, each cycle they are subject to a probability of experiencing disease symptoms. It is assumed that symptoms translate to an immediate diagnosis at the current disease state.

People who do not have cancer are captured for the remainder of their lifetime with general population utility and mortality. As there was no statistically significant differences in specificity between diagnostic strategies, no costs or health related quality of life (HRQoL) decrements were applied to people who had an incorrect positive diagnosis, as there would be no differences between modelled arms in this population.

All people within the model are assumed to have quality of life values in line with the general population, as sourced from the NICE Decision Support Unit (Hernández Alava et al. 2022). However, people who have metastatic disease are assumed to have an annual decrement applied to their quality of life, following assumptions used in NG131 (NICE 2021b). In addition, the model captures a decrement in quality of life associated with transitioning between the different disease states. This is to reflect the short-term complications that can occur due to treatments and adverse events.

Table 8 – Quality of life decrements

	Frequency	Decrement
Metastatic disease	Annual	0.137
Transition to low-risk disease	One-off	0.027
Transition to intermediate risk disease		0.029
Transition to high-risk disease		0.027

Mortality was assumed to be captured in line with general population mortality, however metastatic disease was associated with a higher incidence of death.

Full details of the methods and results are available in Appendix 6.

The results of the base case analysis are presented in Table 9. The results show that using AI-assistance is expected to increase costs by £207 per patient and provide an additional 0.02 quality adjusted life years compared to pathologist review alone. This translates to an incremental cost-effectiveness ratio of £13,278, which is below the cost-effectiveness threshold of £20,000, and so using AI-assistance is deemed a cost-effective strategy. Results of the probabilistic sensitivity analysis suggest that using AI-assistance has a 69% probability of being cost effective.

Table 9 – Base case cost-utility analysis results

	Pathologist with AI-assistance	Pathologist alone	Incremental
Total QALYs	8.13	8.11	0.02
Total costs	£8,067	£7,859	£207
ICER			**£13,278**
Abbreviations: ICER: incremental cost-effectiveness ratio; QALY: quality-adjusted life-year			

Results of the deterministic sensitivity analysis demonstrate that cost-effectiveness conclusions are robust to changes in model inputs, with only an increase in age resulting in changes to cost-effectiveness conclusions, due to patients not living long enough to accumulate the benefits of additional prostate biopsy diagnoses.

7. Organisational considerations

Although digitising of slides is commonplace in Wales, not all pathologists utilise them as they do not have the relevant validation and certification. Pathologists also have varying levels of technological capability, and therefore may not use the AI as intended and could therefore see more limited benefits. There is also evidence suggesting there is varying levels of pathologist trust in the AI technologies. Therefore, there is the potential for some pathologists to use the AI and some not, even if all are validated and trained in the use of WSIs and the AI technology. This could cause inconsistencies in biopsy review within hospitals and across Wales. There may also be a risk that a high level of trust in the system could result in reduced attention in the double-checking of results (Meyer et al. 2022). There were transcription issues identified in the included studies, however this would also be the case for manual checking and reporting as is used currently. Additional to this, there is the possibility for AI to be used as either a first or second read, or 'on demand' (or 'concurrent'); which could result in variation of use between hospitals. Finally, one of the potential benefits of AI which has been noted in section 5.6 is the pre-ordering of IHC, however one of the expert reviewers noted that this may not be permissible at NHS Trusts without human sign-off, and not all laboratories would be comfortable with this practice which could cause variation across Wales.

Consideration needs to be given to the technology and staff support available at NHS Trusts across Wales to allow the scanning of slides, integration of AI with pre-existing local laboratory information management systems, and troubleshooting of issues. There may be some variation in slide quality due to variable sampling, H&E testing and scanners used between NHS trusts. Expert reviewers noted that there had been some availability issues with scanners and scanning capacity at sites, which could limit the number of biopsies which could be reviewed by the AI system. They also noted the importance of good IT support for the system, to assist with integrating the use of digital pathology and AI into the routine clinical pathway. The NHS AI and Digital Regulations Service for health and social care recommend integrating and validating digital healthcare technologies on local systems prior to implementing them, and state piloting the technology would also be considered 'best practice' (NHS AI and Digital Regulations Service 2023). The time and costs required for this therefore also need consideration.

Expert reviewers also noted current difficulties around recruitment and retention of pathologists, and ongoing backlogs of case review. They noted that the addition of digital pathology and AI would allow for easier remote review as necessary, and prioritisation of cases

and potential decisions around outsourcing of analysis when demand is high and backlogs build up.

The costs of the AI system, training and maintaining competency and any ongoing costs and system support and maintenance also need to be considered. Expert reviewers noted that in the current financial climate, these costs need to be reasonable and maintainable. They noted that training for use of the AI system was quick and straightforward, and did not have a significant impact on costs or staff time.

8. Patient, carer and family considerations

8.1 Introduction

Health Technology Wales partnered with Velindre NHS Trust to hold focus groups with the aim of determining an understanding of a) patient understanding, experiences of and expectations around the use of Artificial Intelligence (AI) in healthcare and b) patient's views, opinions and acceptance of AI-assisted review of prostate biopsies (AIPB) in the detection and diagnosis of prostate cancer.

Additionally, two reviews of patient experiences of AI in diagnosis for prostate cancer were identified in the clinical evidence search and are reported here.

Please note that the focus groups also discussed Artificial Intelligence (AI) assisted endoscopy (AIAE) in the detection of lower gastrointestinal cancer and pre-cancerous lesions. As such, the results reported here are also reported in EAR055, with references to AIPB removed.

8.2 Velindre Focus Groups

The focus groups were open to current and former cancer patients of Velindre, their family and/or carers. One virtual and one in-person group was held. A total of 22 people attended.

8.2.1 Perceptions and experiences of AI in healthcare

Patient understanding of AI and their experiences of encountering AI in healthcare are vastly varied. Some attendees advised that they have "*no experience or understanding whatsoever*". Some attendees had a basic understanding of AI, whereas others had a comprehensive and well-informed knowledge of what AI can comprise of. This was either obtained through self-research or due to the patient's profession.

> *"AI Intelligence (AI) is a multidisciplinary field of computer science that aims to create intelligent machines capable of performing tasks that typically require human intelligence."*

From discussions, it would appear that science-fiction does not have undue influence on how patients think. Patients were able to recognise the difference between fictionalised ideas of where AI may lead and are more realistic in the perception of how AI may currently be used in healthcare. Patient understanding of AI may be somewhat influenced by how it is reported and communicated through media, from news to socials. Patients were aware of 'big' news stories involving the misuse of AI in other fields, such as the arts and music, and somewhat aware of news coverage of AI concerns, mostly those in relation to the rising use of AI in society at large.

Some attendees were able to identify areas where they have previously encountered AI in their healthcare, from administrative functions such as 'chatbox' services on websites, appointment booking systems, paperwork etc, to other uses such as domestic and surgical. Others were

unaware of where or how AI is currently being used in healthcare in the UK. It was suggested by several attendees that patients are probably encountering AI in healthcare without realising it. With such varied ideas of what AI can look like, do and where it is or is not currently being utilised, patients may not recognise that they are using, or being subjected to, AI.

8.2.2 Expectations and concerns

When asked broadly about their expectations when it comes to the use of AI in healthcare, patients were unsure. Most advised that they would not expect their practitioner to inform them if AI was involved in their care, be that diagnostic, surgical or otherwise. Upon further enquiry, it became apparent that for most of the attendees, this was because they did not think that their practitioner would inform them, and therefore would not expect it, and not because they would not like to be informed. Most attendees agreed that they would like to be informed of the use of AI and several advised that they believed this should form part of the explanation of a procedure, test or result that practitioners would usually give to patients. Attendees advised this should be provided for carers also, where necessary. Most attendees agreed that they would expect to be asked to consent to the use of the AI, but as part of the usual consent they would be giving to the procedure, and not specific consent regarding the AI. A small number of attendees advised that they were not at all concerned with being informed of, or consenting to, the use of AI.

> *"I would expect this as part of any consent forms that I signed"*

> *"I would prefer to be told if being used - I had an operation in October and the surgeon explained what he was going to do so would have expected to have been told then if appropriate"*

> *"It wouldn't bother me at all"*

Patients had many concerns when it comes to how AI can and is being used in healthcare. These could be grouped into the following categories: data and privacy concerns, AI programming, legal responsibility concerns, loss of human skill, and ethical and philosophical concerns.

Patient concerns around the performance of the AI were directly related to their personal understanding. Concerns such as the AI collecting or relaying patient sensitive information, the potential for AI to weaponised, and AI's capability to understand the complexities and nuances of differences between patients when making decisions may not be applicable to current AI uses in healthcare, but are legitimate concerns around AI's use in future. There was concern expressed about how information potentially stored in AI could be used in the future if procured by private companies, such as insurance companies. Concerns around AI bias and programming are already prevalent in digital healthcare communities and it is interesting to know that patients also share the same concerns.

> *"Because we don't know where we're going in the future, with AI in health care. And I think if that was my business, would I insure everybody on the flat rate, or would I wanna look at where the spikes and troughs of different diseases are, and would I want them on the same premiums as the next person."*

Patients also expressed concerns regarding the legal implications around the use of AI. Particularly, the potential for practitioners to off-load responsibility for mistakes, misdiagnoses or serious incidents to the AI to avoid blame or to prevent patients from seeking recompense. Patients would value a clear protocol in the case of AI malfunction and practitioner's responsibility.

"who's accountable if the AI makes a mistake? The AI company? the health board? a named clinician?"

"Who decides? Who is the decision maker?"

8.2.3 Artificial Intelligence (AI) assisted review of prostate biopsies (AIPB) in the detection and diagnosis of prostate cancer.

Patients discussed at length the importance of human connections during very difficult times. The process of cancer diagnosis through to treatment and outcome was described as "one of the most difficult periods of my life". Patients were passionate about the importance of having emotional support from family and friends but were equally passionate about how much of a difference the support of kind and empathetic practitioners can make. There was concern that, with the potential for more automated or simple procedures and tests to be passed over from practitioners to AI, more and more human interactions will be lost and the support of reassurance, kind words, a hand to hold etc that patients rely on to get them through the challenges they are facing will be diminished.

Attendees were given a brief overview of the function and purpose of AI in prostate biopsies. The role of the AI in the technology was explained in simple language and took no more than three minutes. Three key messages were emphasised: that the AI was designed to assist the practitioner by highlighting areas to examine more closely, that the AI was not learning or changing in real-time, and therefore can only preform as programmed, and that the practitioner makes the decisions.

All attendees agreed that it was clear to them what the function of the AI was. Attendees advised that they were reassured about the purposes of the AI. They were particularly reassured to hear that the AI would not be making any decisions and that the practitioners were ultimately responsible for the decisions made.

"So I would feel more reassured with the extra AI"

Patients were then asked if, understanding how the AI works, they would be happy to have their results assessed by both the AI and a pathologist. All attendees unanimously agreed that they would be happy to consent to the use of the AI technology.

"I think AI will be an exciting additional tool to help with both the programmes mentioned as well as many others but it is my hope that the human will always ultimately make the final decision. It should complement what they are already using whether that's for diagnosis or treatment"

Attendees discussed the potential benefits of using AI in prostate biopsy. The principal benefit attendees could see was the potential to identify and diagnose cancer early.

"I think anything that's got the potential to alleviate patient worries when someone is already anxious is a great thing"

One concern that remained was the potential for the presence of the AI to cause the practitioner to rely too much on it to catch areas of concern, and that this could lead not only to loss of skill but also to complacency, consciously or unconsciously.

"the disadvantages there was they might not sort of learn how to be better because they're relying on the AI"

8.3 Literature review

Two articles were found during the clinical literature review, which collected patient views by questionnaire or survey. Rodler et al. (2023) found that overall, patients trust their practitioners for diagnosis more than they do AI, but trust in the AI component is higher where the AI is controlled by, or use in conjunction with, a practitioner. Patients also showed a preference for AI-assisted practitioners against practitioners alone or AI alone. Trust in the AI was higher when the treatment situation was less complex or severe. For complex decision situations, lower trust by patients in AI systems was apparent. Rodler et al. (2023) advised that individual patient factors contributing to trust have to be considered to ensure good patient communication and implementation of AI-based diagnostic and therapeutic tools. They note a correlation between cumulative use of technology and trust in AI. However, other potential factors such as level of education, type of intervention, and subjective perception of the illness were not correlated with trust in AI. Patients still show a preference for discussing their situations with a practitioner.

Rakovic et al. (2022) found that there was a strong perception from patients that adoption of a digital workflow, including AI, would reduce turnaround time, thus reducing patient anxiety. Some felt that reviewing digital images would lead to greater diagnostic accuracy, if images were to be of sufficient resolution and viewed on large digital displays. Many respondents acknowledged the possibility of human error and fatigue in manual reporting of histopathological slides, thus raising the potential advantage of AI as an adjunct to reduce such errors. Participants discussed the need for digital images and associated data to be held on a secure server, well protected from illicit access and accidental loss. Others felt a greater sense of trust in reports generated purely by a human without any computational intervention, or distrust in computer systems. Most participants viewed the AI component favourably, although a few were not favourable due to concerns around the technical performance and a preference for human review. Overall, the authors found that the majority of participants undergoing a prostate biopsy are supportive of the use of technology in the form of digital pathology or AI for diagnostic assistance in assessing their prostate biopsy.

8.3.1 Summary

The results of both papers support the views, opinions and general trends reported from the HTW focus groups. From all three reviews, it would appear that patient understanding of AI is relatively true to its current uses in healthcare. Most patients expect some level of transparency when it comes to being informed of the use of AI in tests and procedures, although this does not have to go into great detail, and should avoid doing so, in order to lessen the burden of information on patient's experience. Only a small percentage of patients are disinterested in understanding how test results have been generated.

Acceptance of AI in prostate biopsy is directly related to understanding the practitioner's role. Patients show an overall preference for AI that is controlled by practitioners and where practitioners retain ultimate responsibility for outcomes. With this reassurance, most patients are welcoming of the introduction of AI in prostate biopsy and it is seen as beneficial to patients in the hope that it could lead to faster, correct diagnosis and reduce the need for further testing.

9. Conclusions

This evidence review summarised published evidence on the effectiveness and cost effectiveness of AI-assisted review of prostate biopsies.

The evidence included in this review suggests there are statistically significant improvements in diagnostic sensitivity, which are consistent across GGs. There is also a reduced use of additional tests such as IHC and referring for second opinions. User acceptance was high, with most pathologists finding the technology easy to use and useful in identifying low grade cancer compared to ASAP. These improvements in sensitivity did not come at the expense of specificity, which remained similar when comparing pathologists alone to AI plus pathologists, which was again seen consistently across GGs. There also appeared to be a reduction in case review time and turnaround time following introduction of AI alongside pathologist review, however the statistical significance of this was unclear. There was also indication of an improvement in inter- and intra-observer concordance, suggesting an improvement in consistency of biopsy review. The impact of the use of AI on patient outcomes was not reported. Overall, the addition of AI to assist pathologist review of prostate biopsies did appear to be beneficial.

The identified PPI evidence and focus groups all indicated that patients understood the ways in which AI could be used in healthcare. On the whole, patients wished to be kept informed when AI was going to be used in the assessment of their biopsies, but this should be kept simple to avoid additional information burden. Patients expected practitioners to retain ultimate responsibility for outcomes, and also expected the AI to be an addition to human review of tests, rather than the AI alone being relied on.

Statistical significance was not commonly reported in the identified clinical effectiveness evidence, meaning it was challenging to determine whether differences in outcomes between control and intervention arms were significant or not. Outcomes reported within the evidence was fairly consistent across all AI technologies identified, however the evidence for Galen Prostate which was relevant to, and eligible for, inclusion in this review was reliant on conference abstracts and unpublished evidence and as such were not peer reviewed.

Experts noted that there was variation in utilisation and confidence of pathologists with digital technology, which could result in differences in use throughout Wales. There were several other local considerations which were important to take into account, including linking the AI results with laboratory information management systems and patient notes, and ensuring there is sufficient support from other departments such as IT. However, they did note that addition of AI to workflow can improve efficiencies and allow for remote assessments, easier prioritisation of cases, and increased ease of external outsourcing where required.

Results of HTW's economic analysis show that using AI-assistance is expected to increase costs by £207 per patient and provide an additional 0.02 quality adjusted life years compared to pathologist review alone, translating to an incremental cost-effectiveness ratio of £13,278. This is below the cost-effectiveness threshold of £20,000, and so using AI-assistance is deemed a cost-effective strategy. Results of the probabilistic sensitivity analysis suggest that using AI-assistance has a 69% probability of being cost effective.

10. Contributors

This topic was proposed by Richard Nicholson, UK Sales Director, Ibex Medical Analytics.

The HTW staff involved in producing this report were:

- A Evans, Patient and Public Involvement (PPI) Manager – PPI author
- E Hasler, Information Specialist – Literature search & information management
- L Batten, Health Services Researcher - Effectiveness author
- M Prettyjohns, Principal Researcher – Cost effectiveness quality assurance
- N Bromham, Senior Health Services Researcher – Effectiveness quality assurance
- R Boyce, Health Economist – Cost effectiveness author
- R Shepherd, Project Support Officer – Project Management

The HTW Assessment Group advised on methodology throughout the scoping and development of the report.

We are grateful to the following subject experts, who also contributed to this appraisal:

- Gokul Vignesh Kanda Swamy; Consultant Urological Surgeon, Swansea Bay University Health Board
- Ioulia Evangelou; Consultant Histopathologist and Lead Uropathologist, Swansea Bay University Health Board
- Jon Oxley; Consultant in Cellular Pathology, North Bristol NHS Trust
- Muhammad Basar Aslam; Consultant Pathologist and Clinical Director, Betsi Cadwaladr University Health Board
- Margaret Horton; Vice President of Evidence Generation and Clinical Partnerships, Paige Prostate
- Pearl Huey; Digital Pathology Project Support, Betsi Cadwaladr University Health Board
- Richard Nicholson; Commercial Director, Ibex Medical Analytics

Subject experts contributed to this appraisal by commenting on a draft of this report, and in some cases providing other advice to HTW's staff and decision making groups. All contributions from reviewers were considered by HTW's Assessment Group and actioned accordingly. However, subject experts had no role in authorship or editorial control, and the views expressed are those of Health Technology Wales.

11. References

American Cancer Society. (2023). Prostate cancer stages. Available at: https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/staging.html [Accessed 27 Feb 2024].

Aslam M, Heath A. (2023). Successful deployment of an artificial intelligence solution for primary diagnosis of prostate biopsies in clinical practice. Trillium Pathology. doi: https://doi.org/10.47184/tp.2023.01.03

Borkowski P, Dettloff J, Zhou J, et al. (2024). AI-empowered digital workflow for prostate pathology in clinical routine: a reader study for prostate biopsies [Pathology Visions 2023, the 14th Annual Meeting of the Digital Pathology Association (DPA)]. Journal of Pathology Informatics. (100362): Conference abstract 100304. doi: https://doi.org/10.1016/j.jpi.2024.100362

Campanella G, Hanna MG, Geneslaw L, et al. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine. 25(8): 1301-9. doi: https://doi.org/10.1038/s41591-019-0508-1

Cancer Research UK. (2021). Prostate cancer incidence statistics. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/incidence#heading-One [Accessed 26 Feb 2024].

Chahal D, Byrne MF. (2020). A primer on artificial intelligence and its application to endoscopy. Gastrointestinal Endoscopy. 92(4): 813-20 e4. doi: https://doi.org/10.1016/j.gie.2020.04.074

Comperat E, Rioux-Leclercq N, Levrel O, et al. (2021a). Clinical level AI-based solution for primary diagnosis and reporting of prostate biopsies in routine use: a prospective reader study [33rd European Congress of Pathology]. Virchows Archiv. 479(Supplement 1): Conference abstract CP-03-001. doi: https://doi.org/10.1007/s00428-021-03157-8

Comperat E, Rioux-Leclercq N, Levrel O, et al. (2021b). Presentation by Dr. Eva Comperat: Clinical level AI-based solution for primary diagnosis and reporting of prostate biopsies in routine use: a prospective reader study. Available at: https://youtu.be/pQPFpfT-Lho [Accessed 08 Feb 2024].

da Silva LM, Pereira EM, Salles PG, et al. (2021). Independent real-world application of a clinical-grade automated prostate cancer detection system. Journal of Pathology. 254(2): 147-58. doi: https://doi.org/10.1002/path.5662

Deep Bio. (2022). DeepDx Prostate - CNB. Available at: https://deepbio.co.kr/products/prostate-cnb/ [Accessed 25 Jan 2024].

Eloy C, Marques A, Pinto J, et al. (2023). Artificial intelligence-assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies. Virchows Archiv. 482(3): 595-604. doi: https://doi.org/10.1007/s00428-023-03518-5

Flach RN, Stathonikos N, Nguyen TQ, et al. (2023). CONFIDENT-trial protocol: a pragmatic template for clinical implementation of artificial intelligence assistance in pathology. BMJ Open. 13(6): e067437. doi: https://doi.org/10.1136/bmjopen-2022-067437

Hernández Alava M, Pudney S, Wailoo A. (2022). Estimating EQ-5D by age and sex for the UK. Available at: https://www.sheffield.ac.uk/nice-dsu/methods-development/estimating-eq-5d [Accessed 25 May 2024].

Ibex Medical Analytics. (2024). Galen Prostate; AI powered pathology. Available at: https://ibex-ai.com/galen-prostate/ [Accessed 31 Jan 2024].

Jones KC, Weatherley H, Birch S. (2022). Unit costs of health and social care 2022. Personal Social Services Research Unit, University of Kent, Canterbury. Available at: https://www.pssru.ac.uk/unitcostsreport/ [Accessed 26 Feb 2024].

Jung M, Jin MS, Kim C, et al. (2022). Artificial intelligence system shows performance at the level of uropathologists for the detection and grading of prostate cancer in core needle biopsy: an independent external validation study. Modern Pathology. 35(10): 1449-57. doi: https://doi.org/10.1038/s41379-022-01077-9

Landis JR, Koch GG. (1977). The measurement of observer agreement for categorical data. Biometrics. 33(1): 159-74. doi: https://doi.org/10.2307/2529310

Meyer J, Khademi A, Têtu B, et al. (2022). Impact of artificial intelligence on pathologists' decisions: an experiment. Journal of the American Medical Informatics Association. 29(10): 1688-95. doi: https://doi.org/10.1093/jamia/ocac103

NHS AI and Digital Regulations Service. (2023). All adopters' guidance. Available at: https://www.digitalregulations.innovation.nhs.uk/regulations-and-guidance-for-adopters/all-adopters-guidance/ [Accessed 9 Apr 2024].

NHS England. (2022). Implementing a timed prostate cancer diagnostic pathway: guidance for local health and care systems. Available at: https://www.england.nhs.uk/publication/rapid-cancer-diagnostic-and-assessment-pathways/ [Accessed 08 Feb 2024].

NICE. (2021a). Paige Prostate for prostate cancer. Medtech innovation briefing [MIB280]. National Institute for Health and Care Excellence. Available at: https://www.nice.org.uk/advice/mib280 [Accessed 24 Jan 2024].

NICE. (2021b). Prostate cancer: diagnosis and management [Published: 09 May 2019; Last updated: 15 December 2021]. NICE guideline [NG131]. National Institute for Health and Care Excellence. Available at: https://www.nice.org.uk/guidance/ng131/chapter/Recommendations#assessment-and-diagnosis [Accessed 24 Jan 2024].

Nicholson R, Theunissen J. (2023). SBRI Outpatients Transformation Challenge. Phase 2 - Project Closure Report; Ibex Medical Analytics [unpublished report]. SBRI Centre of Excellence & Welsh Government.

Office for National Statistics. (2024). National life tables: Wales. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetableswalesreferencetables [Accessed 26 Feb 2024].

Paige AI Inc. (2024). AI to support prostate cancer diagnosis. Available at: https://paige.ai/diagnostic-ai/prostate-suite/ [Accessed 31 Jan 2024].

Pantanowitz L, Quiroga-Garza GM, Bien L, et al. (2020). An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. The Lancet Digital Health. 2(8): e407-e16. doi: https://doi.org/10.1016/S2589-7500(20)30159-X

Prostate Cancer UK. (2022). Prostate biopsy. Available at: https://prostatecanceruk.org/prostate-information-and-support/prostate-tests/prostate-biopsy [Accessed 24 Jan 2024].

Prostate Cancer UK. (2023). What do my test results mean? Available at: https://prostatecanceruk.org/prostate-information-and-support/just-diagnosed/what-do-my-test-results-mean [Accessed 24 Jan 2024].

Public Health Wales. (2023a). Cancer incidence in Wales [2002-2020]. Available at: https://phw.nhs.wales/services-and-teams/welsh-cancer-intelligence-and-surveillance-unit-wcisu/cancer-reporting-tool-official-statistics/cancer-incidence/ [Accessed 26 Feb 2024].

Public Health Wales. (2023b). Prostate cancer: overview. Available at: https://publichealthwales.shinyapps.io/Cancer_Reporting_Tool_PHW/ [Accessed 24 Jan 2024].

Raciti P, Sue J, Retamero JA, et al. (2023). Clinical validation of artificial intelligence-augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. Archives of Pathology & Laboratory Medicine. 147(10): 1178-85. doi: https://doi.org/10.5858/arpa.2022-0066-OA

Rakovic K, Colling R, Browning L, et al. (2022). The use of digital pathology and artificial intelligence in histopathological diagnostic assessment of prostate cancer: a survey of Prostate Cancer UK supporters. Diagnostics (Basel). 12(5): 1225. doi: https://doi.org/10.3390/diagnostics12051225

Raoux D, Yazbin I, Arbov S, et al. (2021a). Novel AI-based solution for supporting prostate cancer diagnosis increases the efficiency and accuracy of reporting in clinical routine. Laboratory Investigation. 101(Supplement 1): Conference abstract 497. doi: https://doi.org/10.1038/s41374-021-00558-w

Raoux D, Yazbin I, Arbov S, et al. (2021b). Presentation by Dr. Delphine Raoux: Novel AI-based solution for supporting prostate cancer diagnosis increases the efficiency and accuracy of reporting in clinical routine. Available at: https://youtu.be/9XuWK8PmUlw [Accessed 08 Feb 2024].

Rodler S, Kopliku R, Ulrich D, et al. (2023). Patients' trust in artificial intelligence-based decision-making for localized prostate cancer: results from a prospective trial. European Urology Focus. [article in press]. doi: https://doi.org/10.1016/j.euf.2023.10.020

Ryu HS, Jin MS, Park JH, et al. (2019). Automated gleason scoring and tumor quantification in prostate core needle biopsy images using deep neural networks and its comparison with pathologist-based assessment. Cancers (Basel). 11(12): 1860. doi: https://doi.org/10.3390/cancers11121860

SBRI Centre of Excellence. (2023). Artificial intelligence technology to improve prostate cancer diagnosis. Available at: https://sbriwales.co.uk/case-study/artificial-intelligence-technology-to-improve-prostate-cancer-diagnosis/ [Accessed 24 Jan 2024].

Takwoingi Y, Dendukuri N, Schiller I, et al. (2023). Undertaking meta-analysis. In: Deeks J, Bossuyt P, Leeflang M & Takwoingi Y (eds.) Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. London: The Cochrane Collaboration. Chapter 10: p.249-325. doi: https://onlinelibrary.wiley.com/doi/10.1002/9781119756194.ch10

The Royal College of Pathologists. (2023). Position statement from the Royal College of Pathologists (RCPath) on Digital Pathology and Artificial Intelligence (AI). Available at: https://www.rcpath.org/static/90e5e248-4ad3-4d61-8247223f9faffc80/RCPath-AI-position-statement-2022.pdf [Accessed 23 Feb 2024].

University of Oxford. (2021). Oxford University and partners win government funding to evaluate Paige Prostate Cancer Detection System. Available at: https://www.ox.ac.uk/news/2021-06-16-oxford-university-and-partners-win-government-funding-evaluate-paige-prostate-0 [Accessed 24 Jan 2024].

12. Evidence review methods

We searched for evidence that could be used to answer the review question: What is the clinical and cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer?

The criteria used to select evidence for the appraisal are outlined in Appendix 1. These criteria were developed following comments from the Health Technology Wales (HTW) Assessment Group and UK experts.

The systematic search followed HTW's standard rapid review methodology. A search was undertaken of Medline, Embase, Cumulated Index to Nursing and Allied Health Literature (CINAHL), KSR Evidence, Cochrane Library, and the International Network of Agencies for Health Technology Assessment (INAHTA) HTA database. Ongoing study registers (WHO ICTRP, ISRCTN Registry, Clinicaltrials.gov, EU Clinical Trails Register, EU Clinical Trials Information System, and PROSPERO) and key websites were also searched. The searches were conducted in November 2023, with an update search of the key databases and forward citation tracking of included studies undertaken in April 2024. Appendix 2 gives details of the search strategy used for Medline. Search strategies for other databases are available on request.

Appendix 3 summarises the selection of articles for inclusion in the review.

12.1 Meta-analysis

We conducted meta-analysis of diagnostic test accuracy by fitting the bivariate model to compare summary sensitivity/specificity points using generalized linear mixed models with the lme4 package in R version 4.3.2. The methods and code were adapted from Chapter 10 of the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Takwoingi et al. 2023). Differences between sensitivity and specificity with and without AI-assistance were assessed by comparing models without covariates to models with a covariate representing AI assistance. Likelihood ratio tests (via the lrtest function from the lmtest package in R) were used to compare the models and assess the significance of AI assistance. To determine whether a model assuming equal or unequal variances for each test provided a better fit, both were compared using likelihood ratio tests. There was no statistically significant difference in the fit between the model assuming equal variances and the model allowing for different variances, so the simpler equal variances model was used.

The unit of analysis in the studies was whole slide images and there were typically around 6 to 8 of these per patient. The data were not reported in enough detail to allow us to account for the correlation between slide images from the same patient in our model. This is a shortcoming of our approach because failing to account for intra-patient correlation in our model could affect the confidence intervals and p-values, making them more narrow and potentially more significant than they actually are.

Heterogeneity was assessed visually by looking at the receiver operating curve (ROC) plot of sensitivity and specificity. This led to the exclusion of Jung et al. (2022) from the pooled analysis as the reported specificity of the pathologists alone was much lower than the other studies. This study was also an outlier in terms of having much higher cancer prevalence in the included WSIs than the other studies.

Appendix 1 – Inclusion and exclusion criteria for evidence included in the review

	Inclusion criteria	Exclusion criteria
Population	People with suspected prostate cancer who have undergone a biopsy	
Intervention	AI-assisted review of prostate biopsies where the AI assesses digitised biopsy slides prior to pathologist review, as an adjunct to standard care for diagnosis	AI-assisted review of MRI scans
Comparison/ Comparators	Pathologist alone prostate biopsies review, without AI	
Outcome measures	• Sensitivity • Specificity • Case review time • Concordance between AI and clinician reviewers • Need for/ use of additional tests (including repeat biopsies and immunohistochemistry) • Resource use • Change in patient management • Overall survival • Progression-free survival • Health related QoL (EQ-5D) • User acceptability and usability	
Study design	We will prioritise the following study types, in the order listed: • Systematic reviews of diagnostic test accuracy studies. • Diagnostic test accuracy studies • Systematic review of RCTs. • Randomised controlled trials. • Non-randomised comparative trials. • Single-arm (no control group) trials that report any relevant outcome. We will only include evidence from "lower priority" sources where this is not reported by a "higher priority" source. This could be because higher priority evidence: • Does not cover all relevant populations • Does not compare the technology of interest to all relevant comparators • Does not cover all outcomes of interest • Reports over short-term follow up periods, and longer follow up data is required to facilitate decision making.	

	Inclusion criteria	Exclusion criteria
	Where relevant and well-conducted systematic reviews exist we will use these by: • Reporting or adapting their reported outcome measures where these are fully relevant to the scope of our review, and appropriate synthesis methods have been used • Using these reviews as a source of potentially relevant studies where the review cannot be used as a source of outcome data We will prioritise systematic reviews in terms of the sources of evidence they include, using the order described above. We will exclude studies describing development of systems including validation of AI-algorithms, both internal and external, in training and testing datasets and non-human models.	
Search date limits	Literature published in 2010 and after	
Language limits	English language only	
Publication status	We will include evidence from studies that are published in full. We will only include evidence from conference abstracts if there are critical gaps in the fully published evidence. We will include details of any ongoing trials that have a planned completion or reporting date within 24 months of the date searches are carried out. We will only include trials of a design that is likely to add to the existing evidence in terms of certainty; for example, if we report evidence from randomised controlled trials in the EAR, we will only report details of ongoing trials if they also use a randomised design.	
Subgroup analysis	Where the evidence allows, we will report outcomes separately according to list any factors identified as potentially influential on outcomes such as: • Likert score as measured by MRI • Reclassification of risk group • Ethnicity	

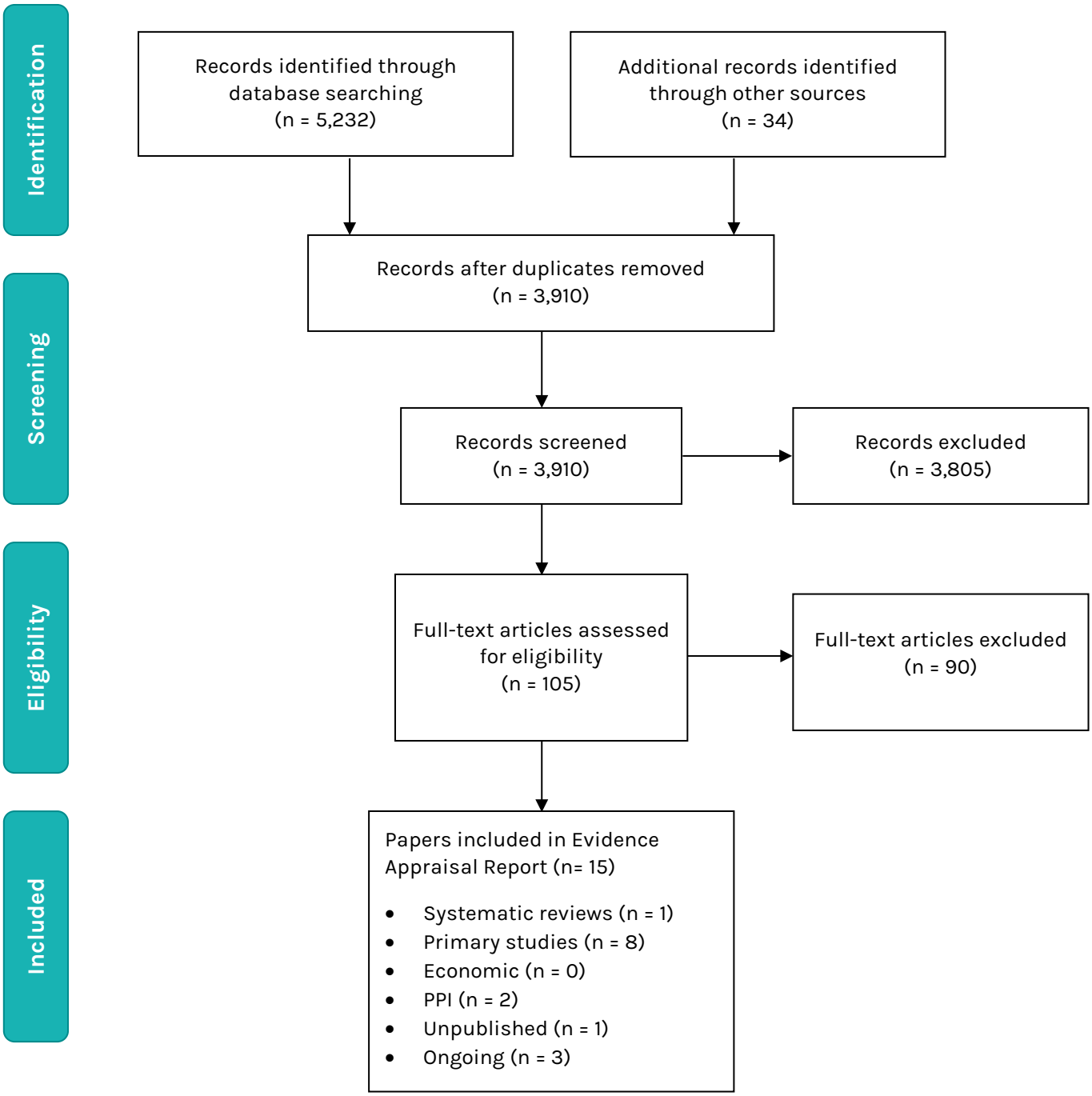
Appendix 2 – Medline strategy

Ovid MEDLINE(R) ALL 1946 to April 01, 2024		
Prostate biopsy		
1	Prostate/ and exp Biopsy/	7172
2	(prostat* adj3 biops*).tw,kf.	14387
3	1 or 2	16752
Prostate cancer		
4	Prostate/	43035
5	exp Prostatic Neoplasms/	152504
6	(prostat* adj4 (neoplas* or cancer* or carcinoma* or adenocarcinom* or tumo?r* or malignan* or metasta*)).tw,kf.	184649
7	or/4-6	226390
Biopsy / Pathology		
8	Biopsy/	189794
9	Biopsy, Needle/	49854
10	Biopsy, Fine-Needle/	15741
11	Biopsy, Large-Core Needle/	2550
12	Image-Guided Biopsy/	5863
13	biops*.tw,kf.	481538
14	Pathologists/	1642
15	Pathology/	33187
16	Pathology, Clinical/	5892
17	Histology/	4347
18	Histological Techniques/	25873
19	Cytodiagnosis/	17425
20	Tissue Array Analysis/	8964
21	(tissue adj1 (array* or microarray*)).tw,kf.	17243
22	(whole* adj1 slide*).tw,kf.	3183
23	((histolog* or histopatholog* or patholog*) adj3 imag*).tw,kf.	26856
24	((digital or digiti?e* or virtual) adj3 (patholog* or histolog* or histopatholog* or slide*)).tw,kf.	6242
25	or/8-24	722337
Artificial intelligence		
26	Artificial Intelligence/	44919
27	((artificial or machine or computer or augment*) adj2 intelligen*).tw,kf.	50001
28	AI.tw,kf.	54476
29	exp Machine Learning/	66348
30	((machine or transfer or deep or hierarch* or computer) adj2 (learn* or reasoning)).tw,kf.	167038
31	(machinelearn* or deeplearn*).tw,kf.	53
32	Neural Networks, Computer/	51637
33	((neural or convolut*) adj2 network*).tw,kf.	113286
34	(CNN or CNNs).tw,kf.	20177
35	(vector adj2 machine*).tw,kf.	28570
36	Fuzzy Logic/	5381
37	(fuzzy adj2 logic*).tw,kf.	2861
38	((computer or machine) adj1 (aid* or assist* or support*)).tw,kf.	58755
39	CADe.tw,kf.	420
40	Diagnosis, Computer-Assisted/	24308
41	Image Interpretation, Computer-Assisted/	47920
42	Automation, Laboratory/	2811
43	(automat* adj1 (approach* or analys*)).tw,kf.	10266

44	((digital or digiti?e* or virtual) adj3 (patholog* or histolog* or histopatholog* or slide*)).tw,kf.	6242
45	or/26-44	461573
Set combination part 1		
46	(3 and 45) or (7 and 25 and 45)	1506
Draft HTW systematic review filter		
47	systematic review.pt.	256889
48	systematic reviews as topic/	12995
49	((systematic$ or evidence$) adj (review$1 or overview$1)).tw,kf,kw.	332989
50	meta-analysis.pt.	198106
51	exp meta-analysis as topic/	29424
52	(meta-analy$ or metaanaly$ or metanaly$).tw,kf,kw.	303154
53	exp review literature as topic/	24923
54	or/47-53	522906
55	(medline or pubmed or medlars).ab.	353253
56	embase.ab.	170491
57	cochrane.ab,jw.	149491
58	(cinahl or cinhal).ab.	50755
59	(psychlit or psyclit or psychinfo or psycinfo).ab.	64946
60	science citation index.ab.	3938
61	cancerlit.ab.	640
62	british nursing index.ab.	429
63	hmic.ab.	397
64	current contents.ab.	1273
65	or/55-64	398551
66	reference list$.ab.	23035
67	bibliograph$.ab.	23671
68	(handsearch$ or hand-search$).ab.	11530
69	relevant journals.ab.	1392
70	manual search$.ab.	6455
71	(search adj (strategy or criteria)).ab.	26563
72	(search$ adj4 literature).ab.	106448
73	or/66-72	173529
74	review.pt.	3301299
75	((selection or inclusion or exclusion) adj criteria).ab.	206906
76	data extraction.ab.	35690
77	74 and (75 or 76)	82739
78	54 or 65 or 73 or 77	709356
79	comment.pt.	1033735
80	letter.pt.	1247839
81	editorial.pt.	686255
82	or/79-81	2235908
83	78 not 82	687899
Draft HTW guidelines filter		
84	exp Evidence-Based Medicine/	76783
85	practice guideline/	31227
86	guideline/	16381
87	exp guidelines as topic/	173138
88	guideline$.ti,kf.	112376
89	exp technology assessment, biomedical/	12297
90	((technology adj (apprais$ or assess$)) or HTA or HTAs).tw,kf,kw.	11998
91	rapid review*.ti,kf,kw.	1431

92	(evidence* adj2 (base* or synthes*)).ti,kf,kw.	49209
93	or/84-92	369175
94	83 or 93	1015482
Set combination part 2 (using systematic review/guidelines filters)		
95	(3 or 7) and 45 and 94	168
Final set combinations		
96	((Ibex* adj3 Medical* adj3 Analytic*) and prostat*).mp.	1
97	(Galen* adj3 Prostat* adj3 Solution*).mp.	0
98	(Deep Bio and prostat*).mp.	0
99	((DeepDx* or Deep-Dx* or Deep Dx*) and prostat*).mp.	1
100	(Paige* and prostat*).mp.	6
101	("Paige.AI" and prostat*).mp.	0
102	("Paige.AI" and "digital pathology").mp.	1
103	or/96-102	9
104	46 or 95 or 103	1631
105	limit 104 to (english language and yr="2010 -Current")	1301

Appendix 3 – Flow diagram outlining selection of relevant evidence sources



Appendix 4 – Full sources of evidence and outcome data

Table A1 – Primary studies: design and characteristics

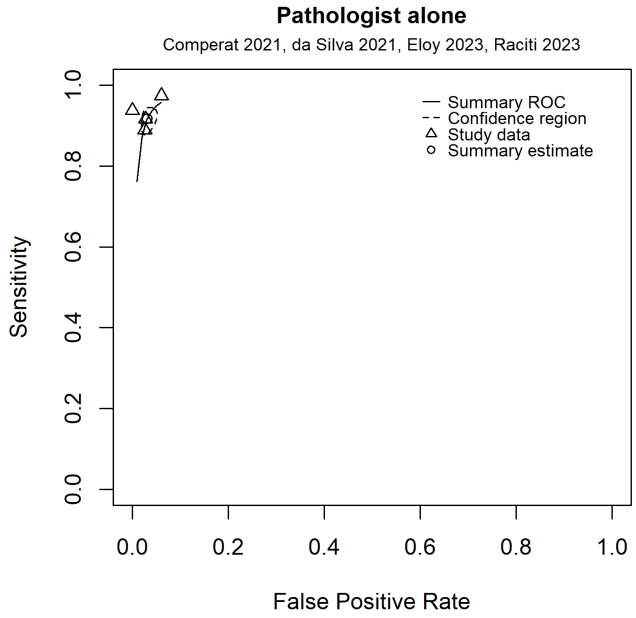
Study reference	Design, Setting	Participants	Intervention	Comparator	Procedures	Outcomes	Comments
Aslam & Heath (2023)	Double reporting and comparison of workflows 2019-2022 Unclear number of pathologists Betsi Cadwaladr UHB, Wales	Number of pathologists unclear 2201 WSIs/ 860 patients 917 PCa, 1,260 benign, 15 ASAP, 9 ungradable Prevalence of malignancy in WSIs 42%	Galen Prostate CNN assisted review	N/A	N/A	Pathologist confidence Benefits of AI	Only qualitative results regarding pathologists opinions included in report, as other results not reported in a way which is suitable for inclusion Not peer reviewed
Borkowski et al. (2024)	Two-arm prospective parallel double-read Three pathologists	180 patients (4366 WSIs in 2183 parts) Prevalence of malignancy in WSIs 20% Unclear number of PCa v benign	Galen Prostate CNN assisted review	Pathologist alone, using WSIs and reviewing digitally	Twice reported slides, once with AI and once without 'AI assistance'; AI output accessed on demand Time between reads unclear	Case review time Pathologist reporting efficiency AI accuracy Usability and productivity	Number of included slides varies throughout abstract and slides, insufficient data to calculate sensitivity and specificity Not peer reviewed
Comperat et al. (2021a)	Two-arm prospective reader parallel crossover Three pathologists France (Medipath)	100 patients (785 slides) Prevalence of malignancy in WSIs 35% Unclear number of PCa v benign	Galen Prostate CNN assisted review	Pathologist alone, using microscope Ground truth is consensus report by two subspecialists, with third reviewing discrepancies, as per H&E/ IHC results	Twice reported slides, once with AI and once without AI used as first read Time between reads unclear	Difference in major discrepancy rates Cancer detection and Gleason scoring 6 v 7+ (Sensitivity/ Specificity) Diagnostic efficiency	All pathologists trained prior to using in reviews Not peer reviewed Consecutive cases used

Study reference	Design, Setting	Participants	Intervention	Comparator	Procedures	Outcomes	Comments
da Silva et al. (2021)	Paired read Two pathologists Brazil	100 patients (682 slides, 661 WSIs assessable by AI) Prevalence of malignancy in WSIs 30% 50 PCa; median age 69 50 benign; median age 65 7 GG1, 18 GG2, 12 GG3. 6 GG4, 7 GG5	Paige Prostate CNN assisted review	Pathologist alone, using WSIs and reviewing digitally Ground truth consensus of pathologist and AI. IHC assessment by pathologists used for discrepant cases	Initial review by pathologist, then AI alone. Re-review by pathologist with AI assistance AI used as second read Time between reads unclear	Sensitivity Specificity PPV/NPV Concordance Case review time	No indication pathologists were trained in use of AI
Eloy et al. (2023)	Multi-reader, multi-case crossover Four pathologists Portugal	105 core needle biopsies 66 benign 39 PCa Prevalence of malignancy in WSIs 37% Median age 69 41 patients 19 GG1, 8 GG2, 5 GG3. 2 GG4, 4 GG5, 1 GG N/A	Paige Prostate CNN assisted review	Pathologist alone, using WSIs and reviewing digitally Ground truth was agreement between all four pathologists, or additional two pathologists with sight of IHC/ AI results in case of discrepancy	Pathologists reviewed slides unaided, then assisted by Paige Prostate 'AI assistance'; AI output accessed on demand At least two weeks between reads	Diagnostic accuracy Concordance IHC ordering Second opinions requested Case read and report time	
Jung et al. (2022)	Double read/ crossover One pathologist South Korea	593 WSIs 130 benign 463 PCa Prevalence of malignancy in WSIs 78% 85 GG1, 92 GG2, 110 GG3, 85 GG4, 91 GG5	DeepDx CNN assisted review	Pathologist alone, using WSIs and reviewing digitally Reference standard is GS and GG determined by 3 uropathology experts	Reviewed dataset without AI, then WSIs order randomised and re-reviewed with AI assistance 'AI assistance'; AI output accessed on demand	Diagnostic accuracy Concordance Slide examination time	No indication pathologists were trained in use of AI Reporting error identified and noted in Table 3

Study reference	Design, Setting	Participants	Intervention	Comparator	Procedures	Outcomes	Comments
					At least four weeks between reads		
Nicholson & Theunissen (2023)	Survey following implementation of AI 14 pathologists Wales	N/A	Galen Prostate CNN assisted review	N/A	N/A	Usability and acceptability	Only survey data included in report Not peer reviewed (unpublished)
Raciti et al. (2023)	Sequential paired-read 16 pathologists (2 GU, others not specialist) 50% New York, 50% other (location unclear)	610 WSIs 420 benign 190 PCa Prevalence of malignancy in WSIs 31% GG 1 110, GG2 39, GG3 10, GG4 12, GG5 4, not graded 1, ASAP 10, treated 4	Paige Prostate CNN assisted review	Pathologist alone, using WSIs and reviewing digitally Ground truth based on initial diagnosis by GU subspecialised pathologists at time of initial reporting, based on all test results	Read all WSIs in randomised order sequentially- first read unassisted, second read assisted AI used as second-read Assisted read immediately follows unassisted read	Sensitivity Specificity Paired reads with AI-driven changes/ accuracy Efficiency gains/ losses	All pathologists trained and had to demonstrate competency at the end of training prior to participation
Raoux et al. (2021a)	Parallel crossover 8 pathologists France (Medipath)	1224 slides overall from 160 cases Unclear number of PCa v benign Prevalence of malignancy in WSIs unclear	Galen Prostate CNN assisted review	Pathologist alone, using microscope Ground truth confirmed by two subspecialists reviewing major discrepancies H&E/ IHC	Cases reported twice, pathologist alone and AI randomised between pathologists AI used as first-read Time between reads unclear	Accuracy of AI Efficiency of reporting Turnaround times Pathologist satisfaction	All pathologists trained prior to using in reviews Not peer reviewed

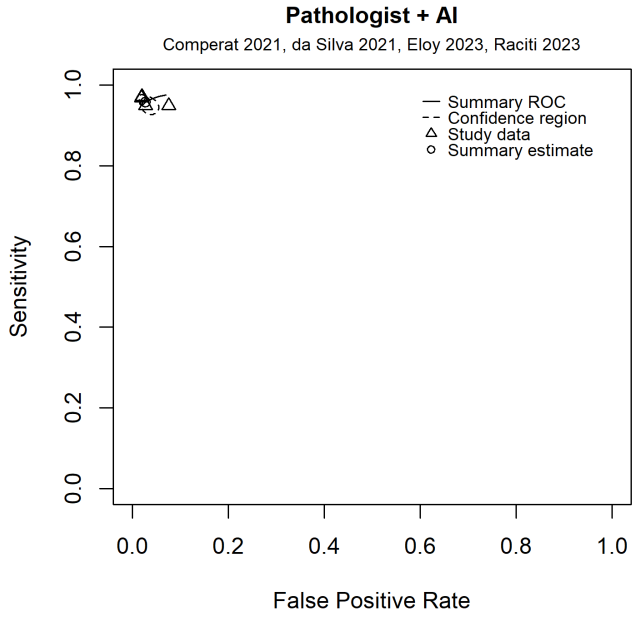
Abbreviations: AI = Artificial Intelligence; ASAP = atypical small acinar proliferation; CNN = convolutional neural network; GG = Grade Group; GU = genitourinary; H&E = haematoxylin and eosin; IHC = immunohistochemistry; NPV = negative predictive value; PCa = Prostate cancer; PV = positive predictive value; WSI = whole slide images

Appendix 5 – Receiver Operating Curve (ROC) plots



Jung et al. (2022) was excluded from the pooled analysis due to heterogeneity.

Figure 1 – Summary ROC and estimate of the sensitivity and specificity of pathologists alone



Jung et al. (2022) was excluded from the pooled analysis due to heterogeneity.

Figure 2 – Summary ROC and estimate of the sensitivity and specificity of AI-assisted pathologists

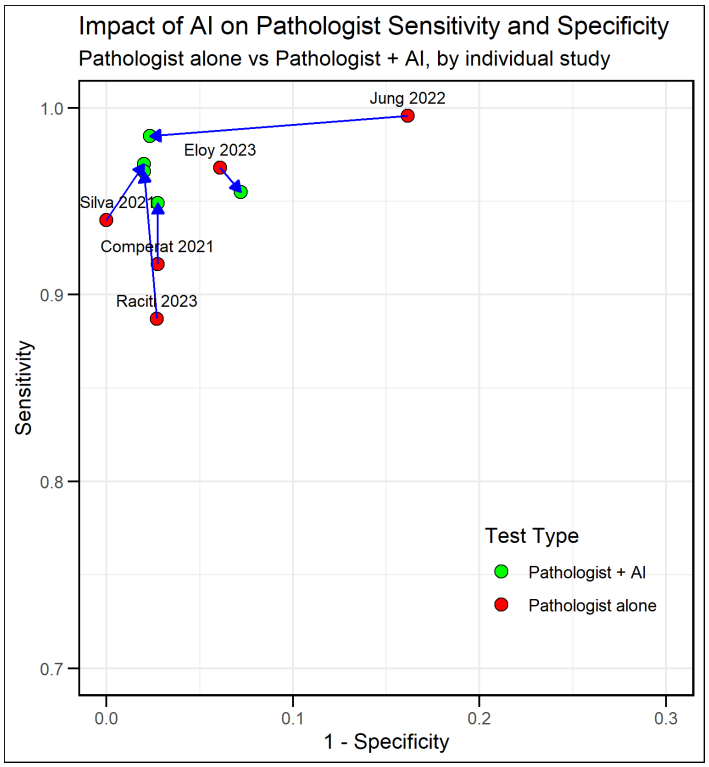


Figure 3 – Impact of AI assistance on the diagnostic accuracy of pathologists, by individual study

Appendix 6 HTW cost utility analysis

1. Background and objective

An economic analysis was developed to estimate the cost effectiveness of AI-assisted review of prostate biopsy in identifying prostate cancer compared to pathologist review alone.

The basic structure of the economic analysis followed that of a cost-utility analysis developed by NICE in development of Guidance NG131 for the diagnosis and management of prostate cancer (NICE 2021b). The model was adapted to include only the diagnostic strategy of pathologist review of a prostate biopsy compared to AI-assisted review of the biopsy, and it captured both costs and benefits of both interventions up to a time horizon of a lifetime (40 years).

2. Methods

2.1 Model approach

A hybrid decision tree and Markov model was developed using Microsoft Excel to compare the cost effectiveness of diagnostic strategies in identifying prostate cancer. The analysis took the perspective of the Welsh NHS and personal social services (PSS). The model comprises a short-term decision tree and lifetime (40 years) predictions of cost, quality of life and mortality to evaluate the cost effectiveness of the following two strategies:

1. AI-assisted review of prostate biopsy in addition to a pathologist
2. Pathologist review of prostate biopsy alone

Future costs and benefits were discounted at a rate of 3.5% per annum.

Figure 4 depicts the basic decision tree model structure. People who have undergone a prostate biopsy for suspicion of prostate cancer will have their biopsy slides reviewed by either a pathologist alone, or a pathologist in addition to AI-assistance. Following concurrent agreement of pathologist and AI, or pathologist opinion alone, a diagnosis is provided. Following a diagnosis of prostate cancer, disease is categorised as: low-, intermediate- or high-risk cancer, or metastatic disease. All people are assumed to initiate treatment once diagnosed. The probability of disease progression is captured every cycle. The model has a cycle length of 3 months.

People who are misdiagnosed as not having cancer continue to progress through the disease states, however, each cycle they are subject to a probability of experiencing disease symptoms. It is assumed that symptoms translate to an immediate diagnosis at the current disease state.

People who do not have cancer are captured for the remainder of their lifetime with general population utility and mortality. As there was no statistically significant differences in specificity between diagnostic strategies, no costs or health related quality of life (HRQoL) decrements were applied to people who had an incorrect positive diagnosis, as there would be no differences between modelled arms in this population.

Mortality was assumed to be captured in line with general population mortality, however metastatic disease was associated with a higher incidence of death.

The model structure is an adaption of the NICE cost-utility model developed for NG131 (NICE 2021b) on the diagnosis and management of prostate cancer. As such, this model includes many of the same assumptions and inputs.

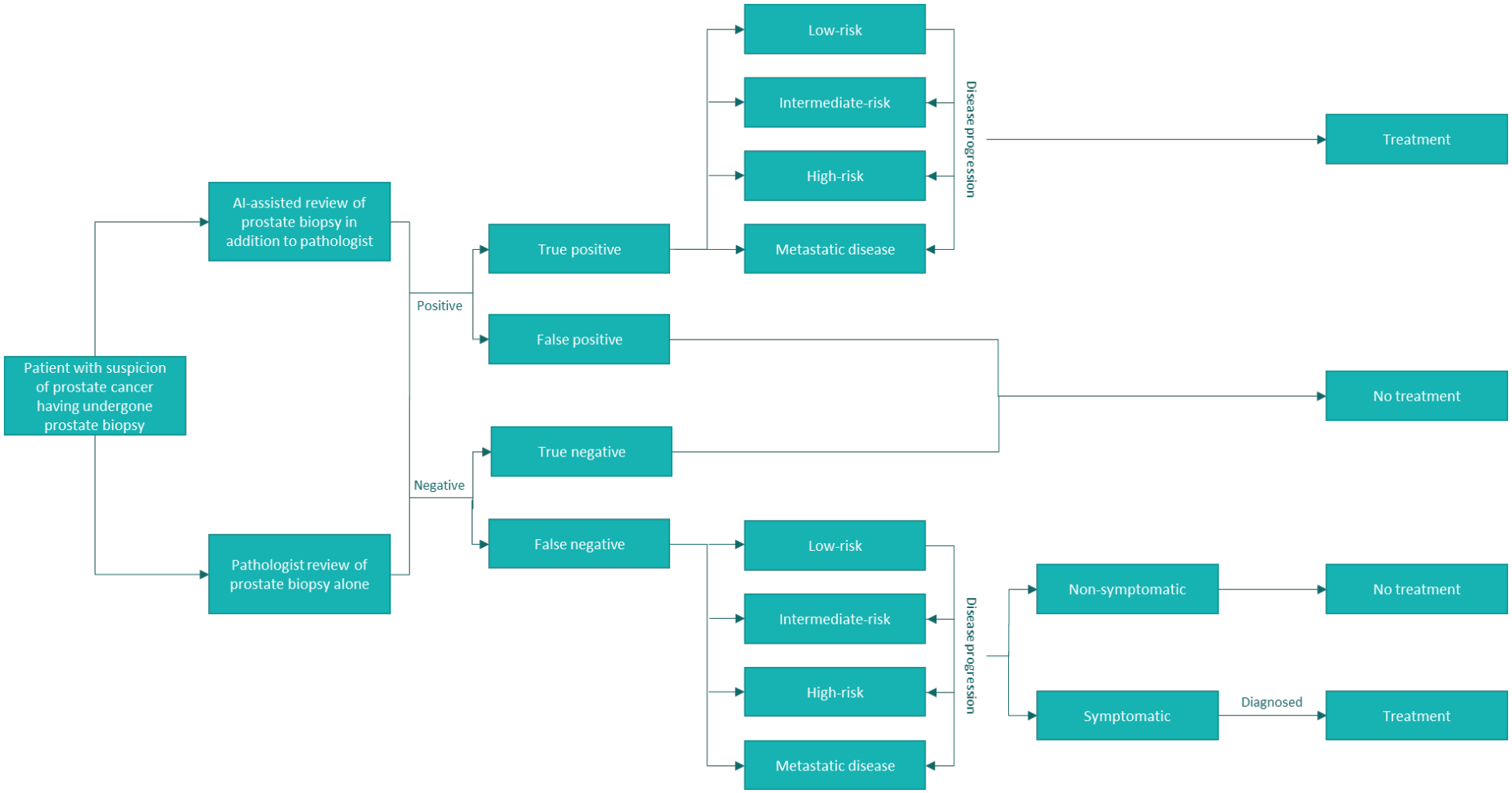


Figure 4 – Schematic of Model

2.2 Clinical data

2.2.1 Diagnostic accuracy

A cohort of 1,000 men are initiated in the model with a baseline age of 68. Although the average age of prostate cancer diagnosis in the UK is 71 (Cancer Research UK 2021), across the studies used to inform the analysis, 2 reported baseline age of their populations. One study reported a mean age of 66.8 (da Silva et al. 2021) and the other a median age of 69 (Eloy et al. 2023).

Following a prostate biopsy, slides are analysed either by a pathologist with AI-assistance, or by a pathologist alone. The sensitivity and specificity of both diagnostic strategies have been sourced from the meta-analysis described in Section 12.1 and the values used in the model are provided in Table A2. For simplicity of sensitivity analyses, a hazard ratio has been calculated for the sensitivity of AI-assisted review based on the results of the meta-analysis. As there were no statistically significant differences in specificity between the diagnostic strategies, the specificity for pathologist alone has been assumed for both arms in the model.

Data on the prevalence of prostate cancer following a prostate biopsy was provided by experts, sourced from laboratory information management system software (LIMS) used in Wales. A prevalence rate of 70.6% was calculated from the data and applied within the analysis.

It should be noted that amongst the studies included in the meta-analysis (Comperat et al. (2021b), da Silva et al. (2021), Eloy et al. (2023), Raciti et al. (2023)), there was an average prevalence rate of 31%; however, this prevalence figure refers to prevalence amongst WSIs rather than people. For the purposes of this analysis, it is assumed appropriate that the sensitivity and specificity derived from WSIs can be applied at a patient level.

Table A2 – Sensitivity and Specificity

	AI-assisted review	Pathologist alone	Hazard Ratio	SE
Sensitivity	96.0%	91.6%	1.05	0.018**
Specificity	97.4%*	97.8%	-	-

Abbreviations: AI – artificial intelligence, SE – standard error
* Modelled as 97.8% in line with pathologist alone due to non-statistical significance.
** Modelled with a lognormal distribution.

2.2.2 Disease states and progression

People with prostate cancer were defined as having low-, intermediate- or high-risk cancer, or metastatic disease. Disease states were determined using a combination of Gleason grades and PSA scores, as outlined in Table A3.

Table A3 – Risk categories of cancer

Risk category	Gleason score	PSA Score
Low risk	6 or lower	10 or lower
Intermediate risk	More than 6	10 to 20
High risk	8 or higher	20 or higher

Abbreviations: PSA, prostate specific antigen

The baseline distribution of patients across disease states was derived from data by Public Health Wales (2023a) which reports disease stage at diagnosis. Data is presented for disease stages I, II, III and IV, and it has been assumed that these stages correspond to the disease states used within the model. This has been deemed an appropriate assumption based on disease categorisation provided by the American Cancer Society (2023). Although data is available up to 2020, 2019 data has been used as this is the last year available which represents expected trends due to the COVID-19 pandemic. The model does not account for unknown stage at diagnosis. Reported data and the distribution applied in the model is provided in Table A4. The distribution at model initiation is equivalent between people with diagnosed and undiagnosed disease at baseline, as it is assumed that all people have an equal chance of having a false negative result. Those who are undiagnosed, however, are at an increased risk of disease progression.

Table A4 – Baseline distribution across disease states

	Disease stage	Distribution	SE*
Reported by Public Health Wales	I	31.27%	0.008
	II	23.88%	0.008
	III	21.24%	0.007
	IV	17.02%	0.007
	Unknown	6.60%	0.005
Modelled	Low risk	33.48%	-
	Intermediate risk	25.56%	-
	High risk	22.74%	-
	Metastatic disease	18.22%	-

Abbreviations: SE – standard error
* Modelled with a beta distribution

The model captures a cyclic probability that people from one disease state may transition to the next, obtained from NG131 (NICE 2021b), Table A5. Progression occurs in people with and without a diagnosis, with a faster progression modelled in those who are undiagnosed. It is assumed that people are only able to progress to the next disease state per cycle, i.e. a person may progress from low-risk prostate cancer to intermediate-risk, but may not progress to high-risk in the same cycle.

Table A5 – Cyclic probability of disease progression

Progression	Diagnosed		Undiagnosed	
	Mean	SE*	Mean	SE*
Low risk – intermediate risk	3.5%	0.011	3.8%	0.006
Intermediate risk – high risk	3.1%	0.006	8.5%	0.030
High risk – metastatic disease	0.8%	0.001	1.4%	0.003

Abbreviations: SE – standard error
* Modelled with a beta distribution

2.2.3 Disease detection

There is a cyclic probability that people with undiagnosed disease experience symptoms which result in an appropriate diagnosis. The probability of experiencing these symptoms per disease state have been derived from data provided in NG131 (NICE 2021b), and are provided in Table A6. All patients who experience symptoms are assumed to be diagnosed correctly as having disease, and no additional diagnostic costs have been included in the model.

Table A6 – Probability of developing disease symptoms

Disease category	Reported Probability			Calculated Cyclic Probability
	Probability	Time period	SE*	
Low risk	2.6%	1 year	0.010	0.66%
Intermediate or high risk	28.4%	5 years	0.020	1.66%
Metastatic disease	61.4%	22 months	0.025	12.17%

Abbreviations: SE – standard error
* Modelled with a beta distribution

2.2.4 Mortality

Mortality for the general population is derived using published life tables for Wales (Office for National Statistics 2024).

Mortality rates for people with prostate cancer is assumed to be in line with general population until progression to metastatic disease. Once disease has progressed to be metastatic, an increased risk of mortality is modelled, using hazard ratios sourced from NG131 (NICE 2021b), Table A7. The risk of mortality is higher in those with undiagnosed metastatic disease.

Table A7 – Hazard ratios for increased risk of mortality with metastatic disease

Metastatic disease status	Hazard ratio	SE*
Undiagnosed	13.38	0.053
Diagnosed	9.07	0.085

Abbreviations: SE – standard error
* Modelled with a lognormal distribution

2.3 Costs

The costs considered in the analysis reflect the perspective of the analysis, thus only costs that are relevant to the UK NHS & PSS were included. Where possible, all costs were estimated in 2022 prices. Where costs were reported in a different cost year, they were inflated to 2022 prices using data from the PSSRU (Jones et al. 2022).

Costs within the model have been sourced from NG131 (NICE 2021b) to account for all treatment cost, adverse event costs and monitoring costs. Costs of the AI device itself have been provided by the manufacturers of both Paige Prostate and Galen Prostate.

Costs have not been included for biopsy or pathologist review of the slides as it was assumed that these would be equivalent between modelled arms and so negligible in the analysis.

2.3.1 Treatment costs

Initial upfront treatment costs differ dependent on the disease state at diagnosis. The distribution of treatment options has been sourced from NG131 (NICE 2021b), Table A8. All people with metastatic disease are assumed to receive a combination of hormone therapy and chemotherapy.

Table A8 – Distribution of treatment options per disease state

Treatment	Disease state		
	Low risk	Intermediate risk	High risk
Active Surveillance	46.7%	25.3%	5.4%
Brachytherapy	6.7%	2.9%	0.6%
Hormone therapy	8.6%	21.6%	47.8%
Radical prostatectomy	17.8%	15.6%	11.6%
External radiotherapy	20.3%	34.7%	34.6%

The corresponding costs of these treatment options, including associated adverse event costs, are provided in Table A9, and have been sourced directly from NG131 (NICE 2021b). Treatment costs are applied at the point of diagnosis or once disease has progressed to the next state.

Table A9 – Treatment costs

Disease state	Treatment costs	SE*
Low risk	£2,325.59	£250.72
Intermediate risk	£2,650.14	£201.20
High risk	£3,871.13	£421.06
Metastatic disease	£14,752.58	£1,636.33

Abbreviations: SE – standard error
* Modelled with a gamma distribution

2.3.2 Monitoring costs

Once diagnosed, people are assumed to be subject to monitoring costs for the duration of their lifetime. These costs, sourced from NG131 (NICE 2021b), are applied on a cyclic basis in the model and are provided in Table A10.

Table A10 – Monitoring costs

Disease status	Monitoring costs	SE*
Low risk	£34.69	£3.47
Intermediate risk	£34.69	£3.47
High risk	£84.33	£8.43
Metastatic disease	£133.97	£13.40

Abbreviations: SE – standard error
* Assumed 10% of the mean - modelled with a gamma distribution

2.3.3 AI costs

Costs of the AI intervention have been provided by Paige Prostate and Galen Prostate, and an average cost has been applied in the analysis. An average cost of £41.32 has been calculated to represent the cost of the AI systems, based on an expected 4,202 men undergoing a prostate biopsy per year.

2.4 Health-related quality of life

As recommended in the NICE reference case, the model estimates effectiveness in terms of QALYs. These are estimated by combining life year estimates with quality of life (QoL) values associated with being in a particular health state.

All people within the model are assumed to have quality of life values in line with the general population, as sourced from the NICE Decision Support Unit (Hernández Alava et al. 2022). However, people who have metastatic disease are assumed to have an annual decrement applied to their quality of life, following assumptions used in NG131 (NICE 2021b). In addition, the model captures a decrement in quality of life associated with transitioning between the different disease states. This is to reflect the short-term complications that can occur due to treatments and adverse events.

Table A11 – Quality of life decrements

	Frequency	Decrement	SE*
Metastatic disease	Annual	0.137	0.036
Transition to low-risk disease	One-off	0.027	0.004
Transition to intermediate risk disease		0.029	0.004
Transition to high-risk disease		0.027	0.004

Abbreviations: SE – standard error
* Modelled with a beta distribution

3. Results

3.1 Base case results

The base case health economic results of the analysis are provided in Table A12. The results show that using AI-assistance in the review of prostate biopsies is expected to increase costs by £207 per patient compared with pathologist review alone; however, it is also associated with 0.02 more QALYs per patient, resulting in a corresponding ICER of £13,278 per QALY. This is below the threshold of £20,000 and therefore AI-assistance is considered to be cost effective compared to pathologist review alone.

Table A12 – Base case results

	Pathologist with AI	Pathologist alone	Incremental
Total QALYs	8.13	8.11	0.02
Total Costs	£8,067	£7,859	£207
ICER (cost per QALY)			**£13,278**

Abbreviations: ICER: incremental cost-effectiveness ratio; QALY: quality-adjusted life year

The base-case analysis predicts that using AI-assistance when analysing prostate biopsy slides could reduce the number of false negative diagnoses by 31 people per 1,000 people undergoing biopsy. This means that these patients will have earlier access to treatment and a lower rate of disease progression compared to when the slides were analysed by pathologist alone.

3.2 Deterministic sensitivity analysis results

A series of deterministic sensitivity analyses (DSA) were conducted, whereby an input parameter is changed, the model is re-run, and the new cost-effectiveness result is recorded. This is a useful way of estimating uncertainty and determining the key drivers of the model result. All inputs in the model were varied by 20% in either direction, with the exception of time horizon, which was tested for 2 years and 5 years, and discounting, which was tested at 0% and 6%.

The results of the 20 most influential parameters on the ICER are presented in Figure 5. The only scenario that resulted in AI-assistance no longer being a cost-effective strategy was when age was increased. Under this scenario, whilst incremental costs are comparable to the base case scenario, patients do not live long enough to experience as much of a QALY benefit due to the additional diagnoses of prostate cancer.

Under all other scenarios, using AI-assistance in the review of prostate biopsies is expected to be a cost-effective strategy.

In addition to the DSA, we also conducted a scenario analysis to explore the assumption around prevalence in the model. As the model is a cohort Markov model, the prevalence of prostate cancer following biopsy has been applied at a patient level. However, the diagnostic accuracy within the model is derived at a WSI level, and so there is a discrepancy between the two parameters in the model. It has been assumed that it is appropriate to model the prevalence at a patient level, however, a scenario using the prevalence at a WSI level has been tested to evaluate whether cost-effectiveness conclusions would be changed.

Table A13 provides the model results when a prevalence of 31% is assumed within the model. Absolute costs across treatment arms are much lower in this scenario, due to a lower prevalence resulting in less patients with prostate cancer undergoing treatment. QALYs are higher under this scenario for the same reason. Incremental QALYs and costs are both lower in this scenario as with a lower prevalence, the AI-assistance is picking up fewer additional cases of prostate cancer. However, using AI-assistance to review prostate cancer biopsies is still expected to be a cost-effective strategy, even with a lower prevalence rate, with an ICER of £16,658.

Table A13 Scenario analysis results

	Pathologist with AI	Pathologist alone	Incremental
Total QALYs	8.89	8.88	0.01
Total Costs	£3,565	£3,451	£114
ICER (cost per QALY)			**£16,658**
Abbreviations: ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life year			

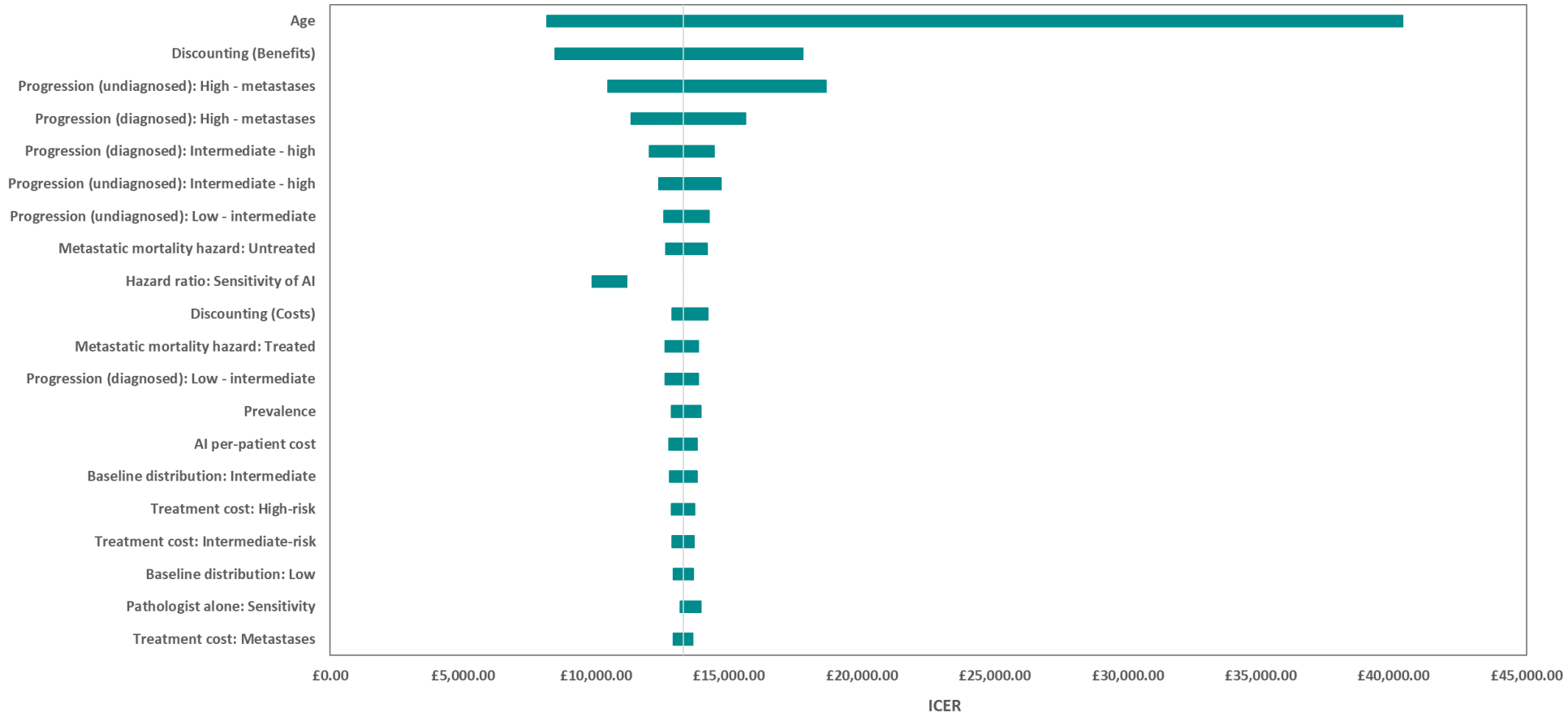


Figure 5 – Deterministic sensitivity analysis results

3.3 Probabilistic sensitivity analysis results

Probabilistic sensitivity analysis (PSA) was conducted to assess the combined parameter uncertainty in the model. In this analysis, the mean values that were utilised in the base case were replaced with values drawn from distributions around the mean values. The results of 10,000 runs of the PSA are presented using ICER scatterplots and cost-effectiveness acceptability curves (CEACs). The ICER scatter plots show the incremental costs and QALYs associated with each of the 10,000 runs of the PSA along with the mean result. The CEAC graphs show the probability of pathologist review with AI-assistance being considered cost effective at the various cost-effectiveness thresholds on the x-axis.

Table A14 presents the health economic results from the PSA. Under this analysis, pathologist review with AI-assistance is a cost-effective strategy with an ICER of £13,807 per QALY, and a 69% chance of being cost effective.

Table A14 – Probabilistic Sensitivity Analysis results

	Pathologist with AI	Pathologist alone	Incremental
Total QALYs	8.16	8.15	0.01
Total Costs	£8,005	£7,801	£204
ICER (cost per QALY)			**£13,807**
ICER: incremental cost-effectiveness ratio; QALY: quality-adjusted life year			

Figure 6 shows the ICER scatterplot for the PSA. Most points reside in the top right of the graph, indicating that AI-assistance is usually more expensive but provides greater benefit to patients. The results of the analysis are more skewed to the cost-effective side of the cost-effectiveness threshold, indicating that AI-assistance is likely to be a cost-effective strategy. It should be noted that when sampling led to the sensitivity of AI-assistance exceeding 100%, 100% was modelled and the sensitivity of pathologist alone was decreased to ensure that the ratio between treatment arms was maintained.

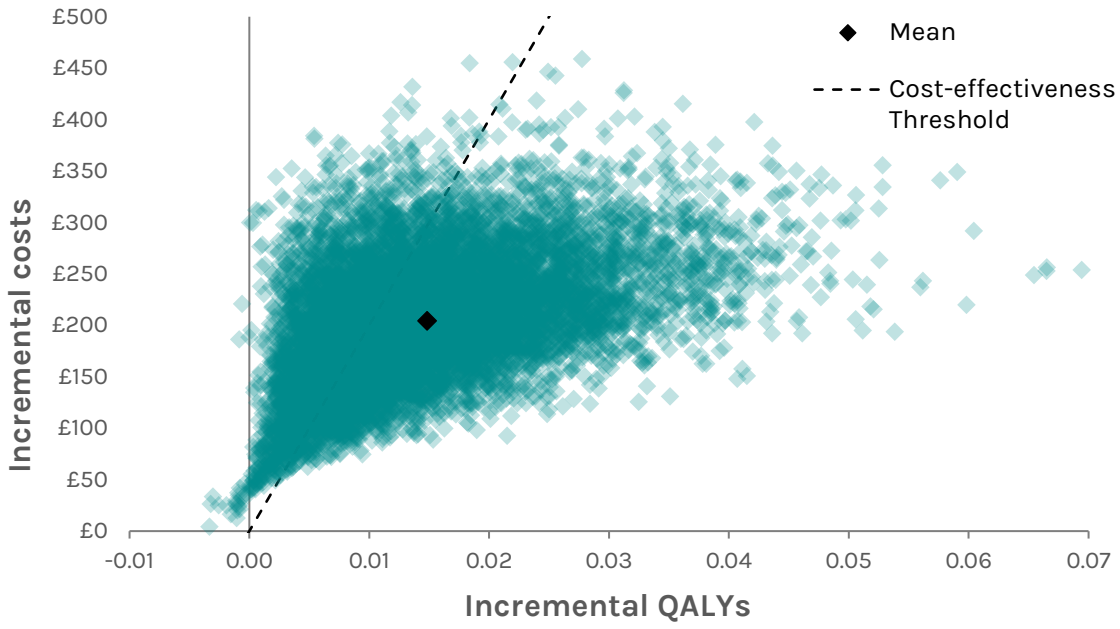


Figure 6 – Cost-effectiveness plane

Figure 7 presents the probability of AI-assistance being considered cost-effective at various cost-effectiveness thresholds.

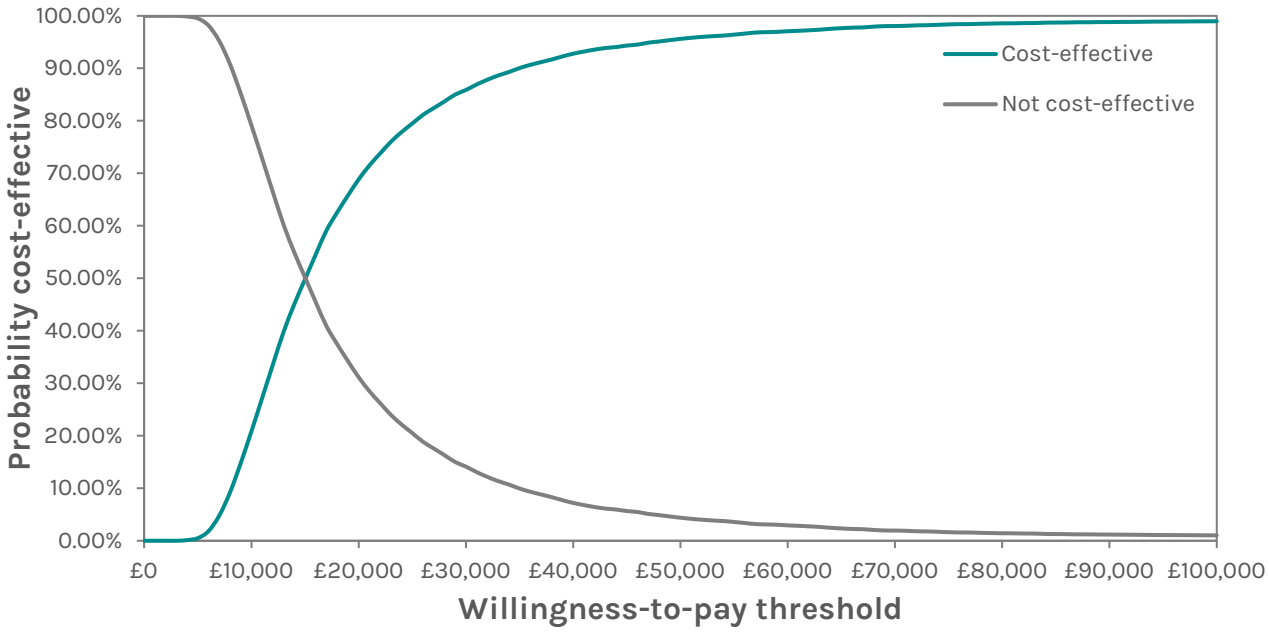


Figure 7 – Cost-effectiveness acceptability curve