Development and Feasibility Evaluation of an Edge AI as Medical Device System for Breast Cancer Multidisciplinary Team Meetings

Authors:

Aarzoo Dhiman[1] (A.Dhiman@hull.ac.uk), Farzana Haque[2] (Farzana.Haque4@nhs.net), Kartikae Grover[2] (kartikaegrover@nhs.net), Lydia Brian Smith[3] (L.Bryan-Smith-2014@hull.ac.uk), William Stephen Jones[1] (Will.Jones@hull.ac.uk)

Affiliations:

[1] Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM), Faculty of Science and Engineering, University of Hull, Hull, United Kingdom

[2] Queen's Centre for Oncology and Haematology, Hull University Teaching Hospitals NHS Foundation Trust, Castle Hill Hospital, Cottingham, United Kingdom

[3] Department of Computer Science, Faculty of Science and Engineering, University of Hull, Hull, United Kingdom

Corresponding author: Aarzoo Dhiman, Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM), Faculty of Science and Engineering, University of Hull, Hull, United Kingdom. Email: a.dhiman@hull.ac.uk

Abstract

Cancer Multidisciplinary Team (MDT) meetings manage increasingly complex cases under considerable time pressure, and documentation requirements can reduce clinical efficiency and decision quality. Although Artificial Intelligence (AI) has the potential to support MDT workflows, most existing systems rely on cloud-based processing, limiting their use because patient discussions contain identifiable information. We developed a fully on-device AI pipeline using an open-source Automatic Speech Recognition (ASR) model and Large Language Models (LLM) that transcribes breast cancer MDT discussions, structures them with a locally hosted LLM, and generates treatment recommendations using retrieval-augmented generation (RAG) grounded in National Institute for Health and Care Excellence (NICE) guidance. All components operate on a single NVIDIA Jetson AGX Orin device, ensuring that patient audio, transcripts, and model outputs remain within institutional infrastructure. The system was evaluated using two recorded simulated MDT discussions, ten clinically validated synthetic discussions, and 1,270 acoustically augmented variants. Silence-aware preprocessing and noise-adaptive decoding on Whisper Large v3 reduced word error rate by 20.7% and 24.4%, respectively, on the recorded discussions compared with the default configuration, and achieved performance within 0.58%-word error rate and 1.58% word information lost of a commercial clinical ASR benchmark on augmented audio. The local MedGemma-RAG model identified 2.3 times more MDT-concordant interventions than a proprietary comparator (recall 0.318 vs 0.136; difference 0.18, 95% CI 0.041–0.327; $p = 0.020$), with no significant difference in overall accuracy. Stakeholders identified automated documentation, treatment recommendation support, and case triage as the most credible near-term applications, while highlighting workflow integration, governance, and clinician trust as important barriers to implementation. This study demonstrates that privacy-preserving, fully on-device AI can support MDT documentation and guideline-informed decision support without cloud infrastructure. These findings provide a practical foundation for secure clinical AI deployment, although prospective evaluation in routine MDT practice is required before clinical implementation.

Keywords: clinical decision support; multidisciplinary team meeting; large language models; retrieval-augmented generation; edge computing; AI as a medical device

1. Introduction

Breast cancer is the most frequently diagnosed cancer in the United Kingdom, with over 60,000 new cases annually [1, 2]. Management is inherently multidisciplinary, requiring the integration of radiological, pathological and clinical findings with tumour biology, disease stage, comorbidities, prior treatment, patient preferences, and eligibility for evolving systemic therapies and clinical trials [1, 3]. Multidisciplinary team (MDT) meetings, or tumour boards, are the mechanism through which this integration occurs and remains the standard for complex cancer care delivery [1, 3-5].

Despite their central role, MDT meetings face significant operational and cognitive challenges [1, 3, 6]. Cancer MDTs review large and growing caseloads with only a few minutes per patient [7], and prolonged sequential decision-making is associated with measurable decision fatigue, reduced information quality and diminished team

contribution across a meeting [6]. Missing results and incomplete records compound this, and national guidance has called for MDT processes to be streamlined [8, 9]. Consequently, a widening gap exists between the information relevant to a decision and what can realistically be synthesised during the meeting, particularly in breast oncology, where recommendations depend on detailed clinicopathological features and guidance distributed across multiple NICE guidelines and technology appraisals [10, 11].

Digital MDT platforms have improved case preparation and decision capture [12], and interest in artificial intelligence (AI) for oncology decision support has expanded rapidly [13, 14]. In the UK, AI software with a medical purpose is regulated as AI as a medical device (AIaMD) within the software as a medical device framework [15, 16]. Recent developments in AI have introduced complementary technologies that address different stages of the MDT workflow. Automatic speech recognition (ASR) captures discussion in real time, supporting documentation, traceability and audit [17, 18], although performance deteriorates with background noise, overlapping speech, and specialist terminology [19-21]. Large language models (LLMs) can summarise and structure clinical information, with tumour board studies reporting moderate-to-high but variable concordance with MDT decisions; however, they perform less consistently for treatment sequencing and complex therapeutic recommendations, underscoring the need for expert oversight [22-26]. Retrieval-augmented generation (RAG) grounds outputs in retrieved documents at inference, improving factual accuracy and guideline alignment while reducing unsupported responses [27-29]. A rapid scoping review of LLM, RAG and AI Decision Support in Breast Cancer MDT Settings is provided in Supplementary Table 1.

Nevertheless, important limitations remain in the current literature. First, existing MDT decision-support studies predominantly evaluate structured referral letters or curated case summaries rather than live MDT discussions and therefore do not address the conversational and unstructured nature of tumour board meetings; the closest RAG work for breast tumour boards similarly uses curated case text as input [30, 31]. Consequently, they neither address the challenges of speech-derived clinical information nor evaluate end-to-end workflows from discussion to recommendation. Second, and more consequentially for adoption, most reported systems rely on proprietary cloud-hosted models. Besides governance, confidentiality, cost and reproducibility concerns, dependence on external inference and structured summaries introduces additional opportunities for information loss, materially limiting deployment in healthcare systems such as the NHS [32].

This study addresses these limitations by developing and evaluating an AIaMD pipeline built using open-source ASR and LLM models that captures MDT discussions, transcribes and structures clinical information, and generates NICE guideline-informed recommendations from MDT conversations in addition to curated case summaries. The entire pipeline runs on a single edge device using quantised models, enabling local inference without external data transmission while allowing comparison with a proprietary cloud-based system. We also quantify the performance gap between synthetic and authentic MDT audio, providing evidence on the extent to which synthetic speech reflects real MDT discussions for ASR evaluation. Following the DECIDE-AI framework, this work represents analytical validation and bench feasibility testing, with prospective clinical evaluation reserved for future studies.

2. Methods

2.1. Study Design and Reporting

This development and bench (non-clinical) feasibility study evaluated an edge-deployed AIaMD pipeline integrating ASR, a locally hosted LLM, and RAG. No patients were recruited, and no identifiable patient data were processed. Evaluation comprised component-level assessment of transcription accuracy, clinical terminology preservation, and recommendation quality, together with system-level assessment of clinical relevance and concordance with MDT recommendations. As this work represents analytical validation and bench testing, items relating to clinical implementation, real-world human factors, and patient outcomes were not applicable and are reported in Supplementary Table 10. Additional methodological details are provided in Supplementary Section 1–8.

2.2. Edge Deployment Architecture

The pipeline was deployed on an NVIDIA Jetson AGX Orin Developer Kit (64 GB) [33] with a 4 TB external solid-state drive for model storage, preserving unified memory for inference. The software stack comprised Whisper large-v3 [34], nomic-embed-text [35], FAISS [36], and quantised GGUF models (27B–70B), orchestrated using CUDA, Docker, TensorRT, and llama.cpp. The 64 GB unified memory enabled local inference with high-parameter models that would otherwise require cloud-based GPU infrastructure. All processing was performed on device; audio, transcripts, embeddings, and model outputs remained within local hardware and institutional infrastructure, supporting data locality and compatibility with NHS information governance requirements.

2.3. Data Sources

Three datasets were used and are reported separately because they provide different levels of evidence.

Tier 1: Authentic simulated MDT recordings. Members of the Hull University Teaching Hospitals (HUTH) NHS breast MDT, including surgeons, oncologists, radiologists, pathologists, specialist nurses, and MDT coordinators, recorded discussions of previously encountered cases while performing their usual MDT roles. All recordings were fully anonymised, preserving authentic clinical terminology, conversational dynamics, and decision-making without processing identifiable patient information. Two recordings were obtained (8 min 21 s and 10 min 20 s); removal of introductory discussion yielded 7 min 44 s of analysable audio for Recording 1. Audio was captured and stored on a University of Hull-managed device without external transmission, and reference transcripts were verified by a consultant oncologist (FH).

Tier 2: Synthetic MDT discussions. Ten multi-speaker MDT scripts were generated from a consultant-authored template using a general-purpose LLM, reviewed for clinical plausibility by FH, and synthesised using Amazon Polly (16 kHz, 16-bit, mono) [37] with role-specific voices and accents (Fig. 1, Supplementary Table 2-3).

Tier 3: Acoustically augmented audio. Each Tier 2 recording was augmented using all combinations of background noise, room impulse response convolution, speech overlap, speed perturbation, pitch shifting, and signal dropout, generating 127 variants per recording and 1,270 audio files (Supplementary Table 4). The workflow used to generate the Tier 2 and Tier 3 datasets is shown in Fig. 1.

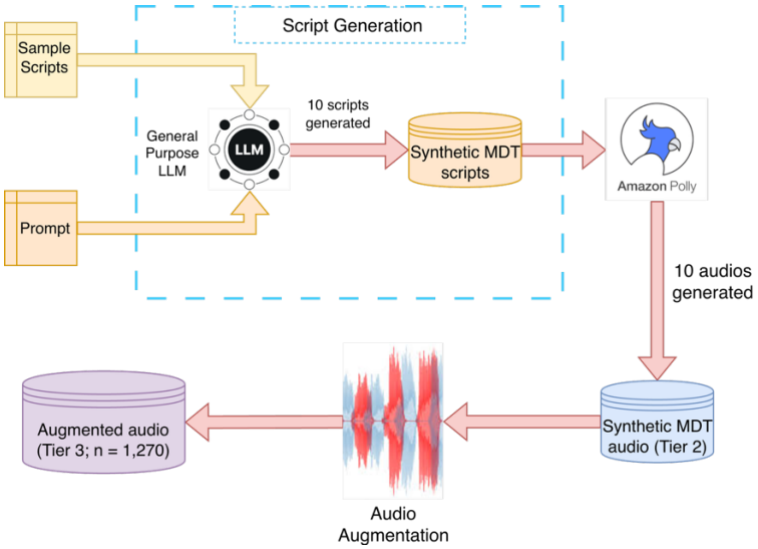

Fig. 1 Workflow for generation of synthetic and augmented datasets. Sample MDT scripts and a prompt template were used to generate synthetic MDT discussions with a general-purpose LLM. Discussions were synthesised using Amazon Polly to produce Tier 2 audio, which was subsequently augmented using acoustic transformations to generate the Tier 3 evaluation dataset.

Synthetic and augmented datasets (Tier 2 and Tier 3) were derived from text-to-speech audio and therefore did not reproduce spontaneous speech characteristics such as disfluencies, interruptions, false starts, or variation in microphone distance. These datasets were used for controlled development and benchmarking, whereas the Tier 1 recordings were reserved as a held-out evaluation set to assess performance on representative MDT conversations.

2.4 Transcription Pipeline and Evaluation

Four ASR systems were evaluated in their default configurations: Whisper large-v3 [34] , NVIDIA Parakeet [38], and WhisperX [39], all open-source and locally deployable, together with Amazon Transcribe Medical (AWS) [40], which was included only as a commercial clinical benchmark and was not part of the proposed pipeline (model details in Supplementary Table 5).

Transcription quality was assessed using word error rate (WER) and word information lost (WIL) following normalisation with the Whisper English text normaliser. As these metrics assign equal weight to all tokens, a domain-specific measures were additionally evaluated using recall of curated breast cancer diagnostic and treatment phrases using exact and fuzzy matching (Levenshtein similarity ≥80%) across 1-6-grams (Supplementary Section 3).

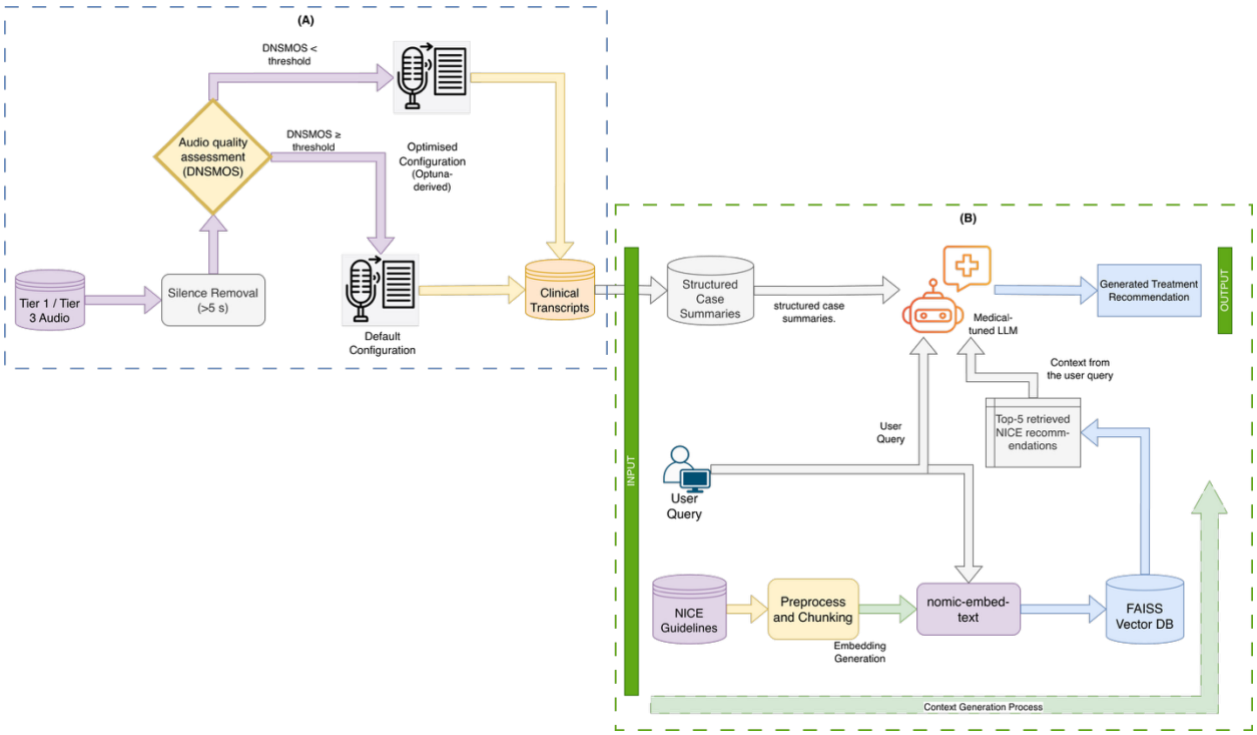

Fig. 2 Overview of the proposed edge-deployed AI pipeline. (A) Adaptive transcription pipeline. Audio recordings (Tier 1 and Tier 3) undergo silence removal followed by DNSMOS-based quality assessment. Depending on the estimated audio quality, recordings are transcribed using either the default or an optimised Whisper configuration to produce clinical transcripts. (B) Guideline-informed recommendation pipeline. Clinical transcripts are converted into structured case summaries, which are embedded and used to retrieve the five most relevant NICE guideline recommendations from a FAISS vector database. Retrieved guideline context and the structured case summary are provided to MedGemma to generate treatment recommendations.

The best-performing open-source ASR model was subsequently optimised using a three-stage pipeline (Fig. 2 (A)) comprising removal of silent segments longer than 5s with FFmpeg silencedetect to reduce repetition and hallucination during long-form decoding [41], Bayesian hyperparameter optimisation with Optuna [42] (100 trials, minimising WER), and noise-adaptive inference in which Deep Noise Suppression Mean Opinion Score (DNSMOS) [43] estimated audio quality and dynamically selected default or optimised decoding parameters using thresholds between 2.0 and 5.0 (Supplementary Table 6). Hyperparameter optimisation and threshold selection were performed exclusively on the Tier 3 augmented dataset.

2.5 LLM Processing and Recommendation Generation

Treatment prediction comprised three stages: extraction of structured case information from the MDT transcript, treatment recommendation generation without guideline retrieval, and recommendation generation with retrieval-augmented generation (RAG). Two domain-adapted LLMs were evaluated for case information extraction: MedGemma 27B (text-only, 8-bit quantised GGUF, ~31.8 GB, 128k context) [44] and Palmyra-Med 70B (3-bit quantised IQ3_M GGUF, ~31.9 GB, 32k context) [45], with decoding parameters provided in Supplementary Table 8. One-shot prompting guided the model to standardise clinical terminology, resolve contextually evident

transcription errors, and extract structured case summaries and MDT treatment recommendations without introducing new clinical information. System prompts used are provided in Supplementary Section 4.

The RAG knowledge base was constructed from NICE guidelines CG81 [46], CG164 [47], and NG101 [48] (total 177 pages). The LLM and retrieval-augmented generation workflow is illustrated in Fig. 2 (B). Guidelines were segmented into individual recommendation-level chunks while preserving section headings, recommendation identifiers, source guideline, and page number as metadata. Tables were extracted using Camelot [49] and stored in the same format. Chunks were embedded using nomic-embed-text [35] and indexed in FAISS [36]. During inference, each case summary was embedded, and the five most relevant guideline chunks were retrieved and incorporated into the generation prompt.

RAG based treatment recommendations were evaluated using two output formats: *free-text recommendation generation* and *structured intervention prediction*. For the former, two prompt variants were evaluated to assess the effect of prompt formulation. One generated treatment recommendation directly (named FT1, Supplementary Section 4.3.1), whereas the other first extracted MDT reasoning before recommendation generation (named FT2, Supplementary Section 4.3.2). For the latter, the model classified each intervention within a predefined taxonomy (Supplementary Section 4.4) as Yes, No, or Not Enough Information (NEI). Two oncology experts independently assigned reference labels for all six evaluable cases, with disagreements resolved by consensus before analysis (Supplementary Section 6).

2.6 Expert Evaluation and Statistical Analysis

Open-ended outputs were independently assessed by clinical experts against MDT-derived reference plans using a structured evaluation instrument covering clinical correctness, adherence to NICE guidance, completeness, and patient safety (Supplementary Section 7.1). ChatGPT-5.2 was evaluated using prompts identical to those used for MedGemma, serving as a proprietary cloud-based comparator. From the expert annotations, we calculated precision, recall, the Jaccard coefficient, overgeneration rate, miss rate, hallucination rate, and the proportion of additional interventions judged clinically appropriate (Supplementary Section 7.1). The Jaccard coefficient was included because precision and recall can each be maximised independently, whereas Jaccard provides a single measure of overall agreement between predicted and reference interventions. Overgeneration and miss rates are reported separately because Jaccard penalises false positives and false negatives equally, despite their potentially different clinical consequences.

For structured prediction, NEI responses were evaluated using three scoring schemes: treating NEI as *No*, treating NEI as *Yes*, and treating NEI as a separate third class. Because the positive (*Yes*) class comprised approximately 10% of observations, three-class performance metrics were macro-averaged, and specificity was calculated using a one-versus-rest approach for the *Yes* class. Paired model comparisons used McNemar's test [50] for overall accuracy and paired bootstrap resampling ($B = 5{,}000$) for class-specific performance metrics. Bootstrap resamples were generated by sampling the 396 paired prediction items with replacement while preserving the pairing between models. The resulting dependence structure and its implications are described in Section 4.4.

2.7 Stakeholder Consultation

Two semi-structured group discussions were conducted with members of the HUTH breast MDT at Castle Hill Hospital: MDT coordinators (n = 2) and clinicians (n = 12; eight breast surgeons, two clinical oncologists, one radiologist, and one pathologist). Sessions were facilitated by WSJ, AD, and FH using a predefined topic guide covering MDT workflow, case preparation, documentation, and operational challenges. During the clinician session, the proposed system was demonstrated, and participants were invited to discuss its perceived benefits, risks, and implementation considerations. Notes were reviewed independently by WSJ, AD, and FH and synthesised to identify recurring themes relevant to workflow and system design, using an approach previously reported by this group [51, 52].

3. Results

3.1. ASR Robustness Under Acoustic Degradation

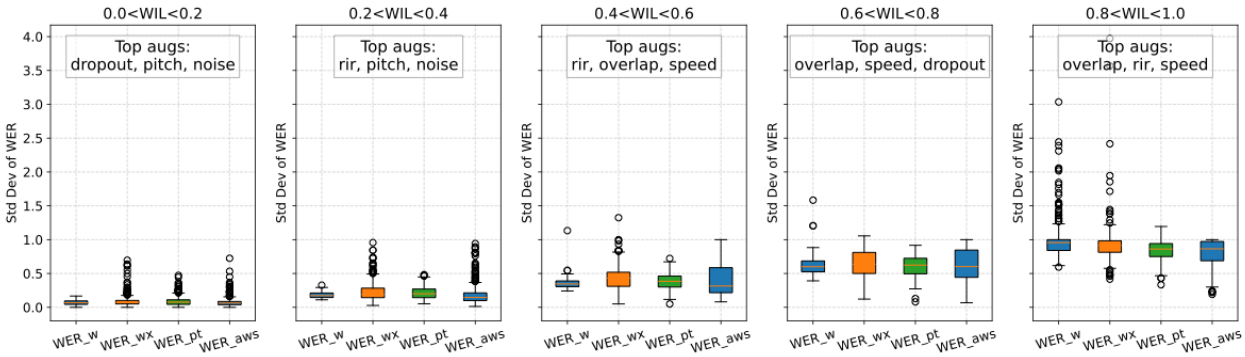

Fig. 3 Distribution of word error rate across word information lost ranges and augmentation types, illustrating model robustness under increasing acoustic degradation

Across the 1,270 augmented recordings (Fig. 3), WER remained consistently low for files with WIL below 0.4, where background noise and pitch perturbation predominated. Error variability increased markedly between WIL values of 0.4 and 0.6, corresponding mainly to reverberation and partial speaker overlap. At WIL values above 0.6, interquartile ranges widened substantially and extreme outliers became more frequent, indicating intermittent transcription failure rather than gradual performance decline. Recordings with WIL ≥ 0.8 were dominated by combinations of overlapping speech, speed perturbation, and reverberation. These findings suggest that ASR performance in MDT-like recordings is affected more by interacting temporal and multi-speaker distortions than by individual acoustic perturbations alone [53].

3.2. Model Selection and the Synthetic-to-Authentic Gap

Table 1 Baseline automatic speech recognition performance on the augmented dataset (Tier 3), with clinical phrase recall. Amazon Transcribe Medical is a commercial benchmark, not a deployable component of the pipeline

Metric	Whisper large-v3	WhisperX	Parakeet	Amazon Transcribe Medical
Mean WER	0.39	0.41	0.37	0.36
Mean WIL	0.44	0.47	0.46	0.42
Worst 3 WER (of 10 cases)	3.03, 2.44, 2.05	3.97, 3.56, 1.94	1.19, 0.98, 0.96	1.0, 1.0, 1.0
Clinical phrase recall	0.56	0.60	0.52	0.59

Baseline performance on Tier 3 dataset is summarised in Table 1. Amazon Transcribe Medical achieved the lowest WER and WIL, consistent with its optimisation for clinical speech. Among the open-source models, NVIDIA Parakeet achieved the lowest WER, whereas Whisper large-v3 achieved the lowest WIL. This ordering differed for clinically weighted evaluation, with Whisper large-v3 recovering more curated breast cancer diagnostic & treatment phrases than Parakeet (recall 0.56 vs 0.52). Although WhisperX achieved the highest phrase recall, it produced the highest WER and WIL and is based on the superseded Whisper large-v2 architecture; therefore, Whisper large-v3 was selected for subsequent optimisation.

Optimising Whisper large-v3 resulted in measurable improvements in transcription performance (optimised decoding parameters in Supplementary Table 6). A DNSMOS threshold of 3.5 produced the lowest WER (0.3635) and WIL (0.4237), representing improvements of 7.3% and 4.5%, respectively, compared with the default configuration and reducing the performance gap with the commercial benchmark to 0.58% WER and 1.58% WIL. As optimisation was performed on the augmented dataset, these results represent in-sample performance.

Table 2 Effect of silence removal, hyperparameter optimisation, and DNSMOS-based adaptive decoding on transcription performance, repetition artefacts, and clinical terminology recall for the held-out authentic MDT recordings (Tier 1). Ground-truth repetition severity was 3.16 for Recording 1 (R1) and 0 for Recording 2 (R2). The complete configuration sweep is available in Supplementary Table 7.

Configuration	Transcription Performance				Repetition Artefact Severity		Clinical Term Recall	
	R1		R2		R1	R2	R1	R2
	WER	WIL	WER	WIL				
Default	0.62	0.72	0.27	0.34	9.11	0.80	0.31	0.63
Silence removed	0.59	0.72	0.26	0.36	15.83	0.88	0.22	0.52
Tuned parameters	0.58	0.74	0.24	0.38	1.82	0.32	0.25	0.66
Silence removed + tuned	0.53	0.69	0.22	0.33	3.10	0.00	0.25	0.64
Tuned + threshold 2.5	0.70	0.52	0.33	0.26	4.09	0.67	0.32	0.61
Silence removed + tuned + threshold 3.0	0.58	0.75	0.19	0.29	4.09	0.67	0.32	0.61

Silence removed + tuned + threshold 3.5	0.49	0.66	0.21	0.33	0.62	0.15	0.28	0.58

Table 2 presents results on the held-out Tier 1 recordings. Silence removal combined with tuned decoding parameters reduced WER from 0.62 to 0.53 for Recording 1 and from 0.27 to 0.22 for Recording 2. Applying DNSMOS-based threshold selection further reduced WER to 0.49 for Recording 1 (20.7% reduction from the default configuration) and 0.2070 for Recording 2 (24.4% reduction). The lowest WER for Recording 2 (0.19; 31.4% reduction) was achieved with a threshold of 3.0. Although the optimal threshold differed between recordings, values between 3.0 and 3.5 consistently produced the best performance.

Performance differed between the augmented (Tier 3) and authentic (Tier 1) datasets. The optimised model achieved a WER of 0.36 on the augmented audio compared with 0.49 and 0.21 on the two authentic MDT recordings, indicating greater variability in performance on representative MDT discussions. Owing to the limited number of authentic recordings, these observations should be interpreted cautiously.

Clinical terminology recall (Table 2) demonstrated that improvements in overall transcription accuracy did not consistently translate into improved preservation of clinically important terms. For Recording 1, the highest terminology recall (0.32) was 4.8% higher than the default configuration, whereas silence removal alone reduced recall from 0.31 to 0.22 despite improving WER. This divergence indicates that aggregate transcription metrics and clinical terminology preservation do not necessarily improve in parallel.

3.3. Case Extraction and Model Selection for Generation

MedGemma produced more complete and consistently structured case summaries than Palmyra-Med, which frequently summarised rather than extracted clinical information and omitted content from longer MDT discussions. Residual errors in MedGemma outputs were primarily attributable to transcription inaccuracies rather than unsupported content generation. Under the edge deployment constraints, MedGemma was deployed as an 8-bit quantised 27B model, whereas Palmyra-Med required 3-bit quantisation to accommodate its 70B parameters within the available memory. In addition, Palmyra-Med was limited to a 4,096-token output compared with 8,192 tokens for MedGemma, reducing its ability to process longer transcripts. Consequently, MedGemma was selected for subsequent recommendation generation.

3.4. Open-Ended Treatment Generation

Table 3 Performance comparison of LLM-based treatment prediction using retrieval-augmented generation (RAG). MR = MedGemma-RAG (local model); CR = ChatGPT-5.2 (cloud-based comparator). Prompt variants FT1 and FT2 correspond to the prompting strategies described in Supplementary Section 4.3.

Model	Precision	Recall	Jaccard	Overgeneration	Miss rate	Hallucination	Proportion appropriate
MR - FT1	0.27	0.33	0.21	0.73	0.33	0.00	0.78
CR - FT1	0.29	0.61	0.27	0.71	0.22	0.00	0.68

MR 7 - FT2	0.44	0.44	0.36	0.56	0.28	0.06	0.53
CR 7 - FT2	0.42	0.56	0.34	0.58	0.33	0.12	0.47

Table 3 summarises the performance of model-generated treatment interventions against MDT-derived reference plans under two prompt variants. Across all conditions, precision ranged from 0.27 to 0.44 and recall from 0.33 to 0.61. ChatGPT-5.2 achieved higher recall under both prompt variants but at the expense of greater overgeneration. Under prompt variant FT2, MedGemma-RAG achieved higher precision (0.44 vs 0.42), Jaccard coefficient (0.36 vs 0.34), and proportion of clinically appropriate additional interventions (0.53 vs 0.47), while reducing overgeneration (0.56 vs 0.58) and hallucination (0.06 vs 0.12). Under prompt variant FT1, MedGemma-RAG produced the highest proportion of clinically appropriate additional interventions (0.78 vs 0.68), with no hallucinated interventions generated by either model. Overgeneration remained common across both models (0.56–0.73), whereas hallucination rates were low (0.00–0.12). Between 47% and 78% of additional interventions generated by the models were judged clinically appropriate by expert reviewers.

3.5. Intervention-based Structured Prediction

Table 4. Performance of intervention-based structured prediction across 396 paired predictions under three NEI scoring schemes. Scenario A treats NEI as No, Scenario B treats NEI as Yes, and the three-class analysis evaluates NEI as a separate class. Precision, recall, and F1 are reported for the Yes class in Scenarios A and B; the three-class analysis reports macro-averaged metrics with one-versus-rest specificity for the Yes class. Full paired-comparison statistics are provided in Supplementary Table 9.

Scenario	Model	Accuracy	Precision	Recall	F1	Specificity	Cohen's κ
A (NEI = No)	MR	83.6%	0.29	0.32	0.30	0.90	0.21
A (NEI = No)	CR	80.8%	0.14	0.14	0.14	0.89	0.03
B (NEI = Yes)	MR	60.1%	0.23	0.83	0.36	0.57	0.19
B (NEI = Yes)	CR	61.4%	0.24	0.91	0.39	0.57	0.22
Three-class	MR	54.0%	0.43	0.52	0.36	0.90	0.14
Three-class	CR	52.5%	0.39	0.49	0.31	0.89	0.12

Across 396 paired intervention predictions (Table 4), overall accuracy did not differ significantly between MedGemma-RAG and ChatGPT-5.2 under any scoring scheme (Scenario A: $p = 0.11$; Scenario B: $p = 0.63$; three-class: $p = 0.56$). Under Scenario A (NEI = No), MedGemma-RAG achieved significantly higher precision (+0.15, 95% CI 0.03–0.27; $p = 0.02$), recall (+0.18, 95% CI 0.04–0.33; $p = 0.020$), and F1 score (+0.16, 95% CI 0.04–0.29; $p = 0.01$) for the positive (**Yes**) class, identifying more than twice as many MDT-concordant interventions than ChatGPT-5.2 (recall 0.32 vs 0.14). Cohen's κ indicated slight-to-fair agreement for MedGemma-RAG and slight agreement for ChatGPT-5.2. When NEI responses were scored as positive (Scenario B), differences between models were no longer statistically significant.

3.6. Stakeholder Consultation

Four themes emerged across both stakeholder groups: fragmented digital infrastructure, administrative burden and workflow inefficiency, the potential value of AI-assisted documentation and data integration, and barriers to implementation relating to trust, governance, and system integration.

MDT coordinators described preparation as a multi-day process requiring manual collation of information from multiple non-interoperable hospital systems, followed by repeated transfer and cross-referencing into the Somerset Cancer Register [54] as the official MDT record. This duplication was viewed as time-consuming and cognitively demanding given the weekly caseload. Following demonstration of the proposed system, participants identified automated transcription, structured summarisation, and decision documentation as potential means of reducing administrative workload and improving documentation completeness, particularly when key personnel were unavailable. Concerns focused on the reliability of AI-generated outputs, integration with existing hospital systems, and the governance implications of recording MDT discussions.

Clinicians similarly identified fragmented patient information as a major barrier to efficient MDT decision-making, with relevant data distributed across electronic patient records, radiology, pathology, and MDT documentation systems. They considered automated generation of structured case summaries, prognostic scoring, guideline-informed treatment recommendations, and clinical trial identification to be the most valuable applications. Concerns centred on interoperability with existing systems, interpretation of unstructured clinical reports, and support for multiple clinically acceptable management options. Participants consistently emphasised that any AI system should undergo robust clinical validation, operate within clearly defined clinical boundaries, and integrate seamlessly into established MDT workflows.

4. Discussion

We developed and evaluated an end-to-end, edge-deployed AI pipeline that combines MDT speech capture with guideline-informed treatment recommendation while keeping all identifiable patient data within institutional infrastructure. A quantised 27B open-source model running on a single edge device outperformed a proprietary cloud-based model on the positive class of the structured prediction task, while the optimised open-source ASR pipeline achieved WIL within 1.58% of a commercial clinical benchmark. These findings demonstrate the technical feasibility of privacy-preserving, on-device AI for MDT decision support.

4.1. What ASR Performance Means in this Setting?

Aggregate transcription accuracy did not necessarily reflect clinically meaningful performance. Although NVIDIA Parakeet achieved a marginally lower WER than Whisper large-v3 (0.3727 vs 0.3927), Whisper recovered more clinically relevant terminology. In breast cancer MDTs, where receptor status, laterality, and disease stage determine treatment recommendations, preserving these entities is more important than overall lexical accuracy [55]. Configurations that improved WER sometimes reduced clinical terminology recall, indicating that WER alone is an insufficient optimisation target for clinical ASR [56]. Silence-aware

preprocessing and DNSMOS-guided adaptive decoding substantially reduced long-form Whisper repetition artefacts without domain-specific retraining, highlighting the importance of preprocessing and inference strategy.

Unlike previous tumour board studies using curated case summaries, our pipeline processes spontaneous MDT discussions, introducing transcription uncertainty into downstream tasks. However, errors affecting negation, disease stage, and laterality are likely to have greater clinical impact than other transcription errors. Despite imperfect transcripts, the LLM frequently generated clinically coherent case summaries, suggesting partial compensation for ASR errors during information extraction. Future work should characterise error propagation across the ASR-LLM pipeline and develop methods to detect clinically significant transcription errors rather than relying solely on aggregate ASR metrics.

4.2. Guideline Grounding and Model Behaviour

Without RAG, both models generated incomplete treatment plans and unsupported guideline references. Retrieval of relevant NICE recommendations improved agreement with MDT decisions and reduced inappropriate recommendations, consistent with previous studies [25]. ChatGPT-5.2 achieved higher recall but generated more additional interventions, whereas MedGemma-RAG achieved higher precision, Jaccard agreement, and positive **(***Yes***)** class performance, with lower hallucination rates. These findings suggest that higher recall may favour option generation, whereas higher precision may be preferable for recommendation verification. Interpretation should remain cautious, as transcript-derived case summaries cannot capture the full clinical context available to MDT members, and some apparently overgenerated interventions may reflect clinically relevant information that was not verbalised during discussion, consistent with the high proportion judged appropriate by expert reviewers.

4.3. Implementation Implications

Stakeholders identified documentation support and case triage as the most credible near-term applications of the proposed pipeline. Automated transcription, structured summarisation, and pre-meeting case preparation were viewed as valuable for reducing administrative burden, particularly the duplication associated with the Somerset Cancer Register. These findings complement the analytical validation by identifying clinical workflows in which the proposed pipeline could provide immediate benefit while supporting clinician-led decision-making.

Interoperability was identified as the principal implementation challenge. Because information is distributed across electronic patient records, radiology, pathology, and MDT documentation systems, AI will provide value only if integrated into existing clinical workflows. Although on-device processing reduces data transmission risks, governance issues relating to consent, data retention, and access control remain essential for deployment. Participants also emphasised the need for clearly defined clinical scope, robust validation, and seamless workflow integration.

For AIaMD, post-deployment monitoring should be considered during system design. The variability observed across authentic MDT recordings and the sensitivity of clinical terminology recall to decoding configuration indicate that performance may change with acoustic conditions, speaker composition, or guideline updates.

Monitoring should therefore extend beyond aggregate ASR metrics to include preservation of clinically important entities and guideline currency.

4.4. Limitations

This study provides analytical validation and bench feasibility testing rather than clinical evaluation. The evaluation was based on six analysable cases and two authentic MDT recordings. Because multiple intervention predictions were generated from each case, the observations were clustered rather than independent, and the reported confidence intervals should be interpreted accordingly. All interventions were weighted equally, although missing chemotherapy is unlikely to have the same clinical consequence as omitting an axillary biopsy; future studies should incorporate consequence-weighted evaluation. Model calibration was not assessed, both models performed poorly on the NEI class, and interpretation of the reference labels should consider the reported inter-rater agreement (Section 3.5).

The RAG pipeline used a general-purpose embedding model, fixed top-*k* retrieval, and general-purpose ASR models without domain adaptation, all of which may limit performance. DNSMOS thresholding was applied at the recording level and may not reflect mixed acoustic conditions within recordings. Finally, the Tier 2 and Tier 3 datasets were generated from text-to-speech audio and did not reproduce the disfluencies and spontaneous turn-taking of authentic MDT discussions. To mitigate this, optimisation was validated on an independent held-out set of authentic recordings, and synthetic and authentic performance are reported separately.

4.5. Future Work

Future work will focus on improving technical performance through domain-specific ASR models and Graph RAG-based treatment recommendation, supported by larger datasets for model adaptation and evaluation. Integration with electronic health records will be essential for the documentation workflows identified by stakeholders. Before recommendation-facing deployment, further work should address consequence-weighted evaluation, calibration, uncertainty estimation, explainability, post-deployment monitoring, and safety assurance. Prospective evaluation in live MDT meetings, reported in accordance with the DECIDE-AI framework, will assess clinical performance, workflow integration, and user acceptance (See Supplementary Table 10).

5. Conclusions

We developed and evaluated an edge-deployed AI pipeline using open-source ASR and LLM models that integrates ASR, locally hosted LLMs, and guideline-informed retrieval to support breast cancer MDT documentation and treatment recommendation without transmitting identifiable patient data beyond institutional infrastructure. The optimised ASR pipeline achieved performance within 1.58% (WIL) of a commercial clinical benchmark on augmented audio, while MedGemma-RAG identified 2.3 times more MDT-concordant interventions than a proprietary cloud-based comparator with consistently low hallucination rates. Performance was more variable on authentic MDT conversations than on synthetic audio, highlighting the importance of evaluating conversational clinical data. Stakeholder consultation indicated that documentation support and case

preparation represent the most credible near-term applications, with workflow integration and governance emerging as greater barriers to implementation than model performance. Prospective clinical evaluation, consequence-weighted assessment, and post-deployment monitoring are required before routine clinical use.

Statements and Declarations

Acknowledgements: The authors thank the members of the Hull University Teaching Hospitals NHS Foundation Trust breast cancer multidisciplinary team who contributed simulated case discussions and participated in the stakeholder consultation sessions.

Competing interests: The authors have no financial or non-financial competing interests to declare that are relevant to the content of this article.

Funding: This study was supported by a grant from the Humber and North Yorkshire Cancer Alliance, Cancer Research and Innovation Award Funding 25-26.

Ethics approval: Ethical approval was not required for this study. The study involved simulated MDT discussions conducted by members of the HUTH breast cancer MDT using fully de-identified clinical scenarios. The exercise was undertaken specifically for the purposes of this study and was performative in nature, with clinicians discussing the presented scenarios and formulating corresponding management decisions. The stakeholder consultation comprised focus group discussions on the breast cancer MDT pathway and the proposed technology. Participants provided informed consent to participate and to audio recording of these discussions, and no sensitive personal or patient-identifiable information was collected.

Consent to participate: Informed consent was obtained from all MDT members who participated in the recorded simulated-case discussions and stakeholder consultation sessions.

Consent to publish: Not applicable, as no identifiable participant or patient information is included in this manuscript.

Data availability: The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request, subject to institutional governance and data protection requirements.

Code availability: The code used to develop and evaluate the pipeline is available from the corresponding author on reasonable request.

Author contributions: Conceptualisation: WSJ, FH, KG, AD; Methodology: AD, WSJ, LBS, FH; Software: AD, LBS; Analysis and Evaluation: AD, WSJ; Data curation: AD, WSJ, FH; Manuscript- original draft preparation: WSJ, AD; Manuscript - review and editing: all authors.

Use of AI tools: All AI and LLM tools used for analytic or generative tasks are disclosed in Methods and Supplementary Material

Clinical Trial Number: not applicable.

References

1. Patkar, V., D. Acosta, T. Davidson, A. Jones, J. Fox, and M. Keshtgar, *Cancer multidisciplinary team meetings: evidence, challenges, and the role of clinical decision support technology.* International journal of breast cancer, 2011. **2011**(1): p. 831605. https://doi.org/10.4061/2011/831605
2. *Breast cancer statistics | Cancer Research UK*. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.
3. Winters, D.A., T. Soukup, N. Sevdalis, J.S. Green, and B.W. Lamb, *The cancer multidisciplinary team meeting: in need of change? History, challenges and future perspectives.* BJU international, 2021. **128**(3): p. 271-279.https://doi.org/10.1111/bju.15495Digital
4. Rajan, S., J. Foreman, M. Wallis, C. Caldas, and P. Britton, *Multidisciplinary decisions in breast cancer: does the patient receive what the team has recommended?* British journal of cancer, 2013. **108**(12): p. 2442-2447.https://doi.org/10.1038/bjc.2013.267
5. *Best Practice diagnostic guidelines for patients presenting with breast cancer symptoms | Association of Breast Surgery*. Available from: https://associationofbreastsurgery.org.uk/professionals/information-hub/guidelines/2010/best-practice-diagnostic-guidelines-for-patients-presenting-with-breast-cancer-symptoms.
6. Soukup, T., T.A. Gandamihardja, S. McInerney, J.S. Green, and N. Sevdalis, *Do multidisciplinary cancer care teams suffer decision-making fatigue: an observational, longitudinal team improvement study.* BMJ open, 2019. **9**(5): p. e027303.https://doi.org/10.1136/bmjopen-2018-027303
7. Gandamihardja, T.A., T. Soukup, S. McInerney, J. Green, and N. Sevdalis, *Analysing breast cancer multidisciplinary patient management: a prospective observational evaluation of team clinical decision-making.* World journal of surgery, 2019. **43**(2): p. 559-566. https://doi.org/10.1007/s00268-018-4815-3
8. Improvement, N.H.S. *Streamlining Multi-Disciplinary Team Meetings Guidance for Cancer Alliances*. Available from: https://www.england.nhs.uk/publication/streamlining-mdt-meetings-guidance-cancer-alliances/.
9. Soukup, T., B.W. Lamb, A. Morbi, N.J. Shah, A. Bali, V. Asher, T. Gandamihardja, P. Giordano, A. Darzi, and N. Sevdalis, *Cancer multidisciplinary team meetings: impact of logistical challenges on communication and decision-making.* BJS open, 2022. **6**(4): p. zrac093.https://doi.org/10.1093/bjsopen/zrac093
10. Gommers, J., V. Hernström, V. Josefsson, H. Sartor, D. Schmidt, A. Hjelmgren, A.-M. Larsson, S. Hofvind, I. Andersson, and A. Rosso, *Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial.* The Lancet, 2026. **407**(10527): p. 505-514.https://doi.org/10.1016/S0140-6736(25)02464-X
11. Javaeed, A. and A. Schuh, *Artificial intelligence in breast cancer diagnosis: A systematic literature review.* Cambridge Prisms: Precision Medicine, 2025. **3**: p. e7.https://doi.org/10.1017/pcm.2025.10006
12. Hammer, R.D., D. Fowler, L.R. Sheets, A. Siadimas, C. Guo, and M.S. Prime, *Digital Tumor Board Solutions Have Significant Impact on Case Preparation.* JCO Clinical Cancer Informatics, 2020(4): p. 757-768.https://doi.org/10.1200/CCI.20.00029
13. Kočo, L., C.C. Siebers, M. Schlooz, C. Meeuwis, H.S. Oldenburg, M. Prokop, and R.M. Mann, *The Facilitators and Barriers of the Implementation of a Clinical Decision Support System for Breast Cancer*

Multidisciplinary Team Meetings—An Interview Study. Cancers, 2024. **16**(2): p. 401.https://doi.org/10.3390/cancers16020401

14. *10 Year Health Plan for England: fit for the future - GOV.UK*. Available from: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future.
15. *Software and artificial intelligence (AI) as a medical device - GOV.UK*. Available from: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device.
16. *The Medical Devices Regulations 2002*. Available from: https://www.legislation.gov.uk/uksi/2002/618/contents.
17. Ng, J.J.W., E. Wang, X. Zhou, K.X. Zhou, C.X.L. Goh, G.Z.N. Sim, H.K. Tan, S.S.N. Goh, and Q.X. Ng, *Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review.* BMC medical informatics and decision making, 2025. **25**(1): p. 236.https://doi.org/10.1186/s12911-025-03061-0
18. Van Buchem, M.M., H. Boosman, M.P. Bauer, I.M. Kant, S.A. Cammel, and E.W. Steyerberg, *The digital scribe in clinical practice: a scoping review and research agenda.* NPJ digital medicine, 2021. **4**(1): p. 57.https://doi.org/10.1038/s41746-021-00432-5
19. Tran, B.D., R. Mangu, M. Tai-Seale, J.E. Lafata, and K. Zheng. *Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations*. in *AMIA Annual Symposium Proceedings*. 2023.
20. O'Kane, R., D. Stonehouse-Smith, L. Ota, R. Patel, N. Johnson, C. Slipper, J. Seehra, S. Papageorgiou, and M. Cobourne, *Transcription Accuracy of Automatic Speech Recognition for Orthodontic Clinical Records.* Journal of Dental Research, 2025: p. 00220345251382452.https://doi.org/10.1177/00220345251382452
21. Miner, A.S., A. Haque, J.A. Fries, S.L. Fleming, D.E. Wilfley, G. Terence Wilson, A. Milstein, D. Jurafsky, B.A. Arnow, and W. Stewart Agras, *Assessing the accuracy of automatic speech recognition for psychotherapy.* NPJ digital medicine, 2020. **3**(1): p. 82.https://doi.org/10.1038/s41746-020-0285-8
22. Shool, S., S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, and M. Tara, *A systematic review of large language model (LLM) evaluations in clinical medicine.* BMC Medical Informatics and Decision Making, 2025. **25**(1): p. 117.https://doi.org/10.1186/s12911-025-02954-4
23. Sorin, V., E. Klang, M. Sklair-Levy, I. Cohen, D.B. Zippel, N. Balint Lahat, E. Konen, and Y. Barash, *Large language model (ChatGPT) as a support tool for breast tumor board.* NPJ Breast Cancer, 2023. **9**(1): p. 44.https://doi.org/10.1038/s41523-023-00557-8
24. Griewing, S., J. Knitza, J. Boekhoff, C. Hillen, F. Lechner, U. Wagner, M. Wallwiener, and S. Kuhn, *Evolution of publicly available large language models for complex decision-making in breast cancer care.* Arch Gynecol Obstet, 2024. **310**(1): p. 537-550.https://doi.org/10.1007/s00404-024-07565-4
25. Hager, P., F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, and D. Rueckert, *Evaluation and mitigation of the limitations of large language models in clinical decision-making.* Nature Medicine, 2024. **30**(9): p. 2613-2622.https://doi.org/10.1038/s41591-024-03097-1
26. Buyukceran, E.U., A. Seyfettin, A. Babaturk, M.B. Ozkan, D. Colak, I. Unal, E. Kaymaz, E. Ergun, M.O. Emer, and H.H. Mersin, *High Concordance Between GPT-4o and Multidisciplinary Tumor Board Decisions*

in Breast Cancer: A Retrospective Decision Support Analysis. J Med Syst, 2025. **49**(1): p. 179.https://doi.org/10.1007/s10916-025-02314-9

27. Zakka, C., R. Shad, A. Chaurasia, A. Dalal, J. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, and E. Ashley, *Almanac—retrieval-augmented language models for clinical medicine. NEJM AI. 2024.* AIoa2300068. **10**.https://doi.org/10.1056/AIoa2300068
28. Kresevic, S., M. Giuffrè, M. Ajcevic, A. Accardo, L.S. Crocè, and D.L. Shung, *Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework.* NPJ digital medicine, 2024. **7**(1): p. 102.https://doi.org/10.1038/s41746-024-01091-y
29. Unlu, O., J. Shin, C.J. Mailly, M.F. Oates, M.R. Tucci, M. Varugheese, K. Wagholikar, F. Wang, B.M. Scirica, and A.J. Blood, *Retrieval augmented generation enabled generative pre-trained transformer 4 (GPT-4) performance for clinical trial screening.* medRxiv, 2024.https://doi.org/10.1101/2024.02.08.24302376
30. Abdullayev, N., J. Kottlors, H. Habibov, F. Yilmaz, C. Zimmer, N.G. Hokamp, S. Lennartz, V. Valiyev, C. Abbasli, and L. Goertz, *European guideline informed RAG-based GPT-4 decision support tool in tumor board meetings for breast cancer treatment.* European Journal of Surgical Oncology, 2025: p. 110384.https://doi.org/https://doi.org/10.1016/j.ejso.2025.110384
31. Dehdab, R., S. Afat, F. Mankertz, J.M. Brendel, N. Maalouf, S. Werner, A. Brendlin, J. Herrmann, K. Nikolaou, L.D. Kloker, B. Calukovic, K. Benzler, L. Zender, and C.K.W. Deinzer, *When AI joins the table: evaluating large language model performance in soft tissue sarcoma tumor board decisions.* Journal of Cancer Research and Clinical Oncology, 2026. **152**(2): p. 52.https://doi.org/10.1007/s00432-026-06432-w
32. Ahmed, M.I., B. Spooner, J. Isherwood, M. Lane, E. Orrock, and A. Dennison, *A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare.* Cureus, 2023. **15**(10): p. e46454.https://doi.org/10.7759/cureus.46454
33. *Jetson AGX Orin for Next-Gen Robotics | NVIDIA*. Available from: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/.
34. Radford, A., J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, in *Proceedings of the 40th International Conference on Machine Learning*. 2023, JMLR.org: Honolulu, Hawaii, USA. p. Article 1182.
35. Nussbaum, Z., J.X. Morris, B. Duderstadt, and A. Mulyar, *Nomic embed: Training a reproducible long context text embedder.* arXiv preprint arXiv:2402.01613, 2024.https://doi.org/10.48550/arXiv.2402.01613
36. Johnson, J., M. Douze, and H. Jégou, *Billion-scale similarity search with GPUs.* IEEE transactions on big data, 2019. **7**(3): p. 535-547.https://doi.org/10.1109/TBDATA.2019.2921572
37. *Amazon Polly*. Available from: https://aws.amazon.com/pm/polly/?trk=5735757a-030b-49bd-96de-519b2a8af735&sc_channel=ps&ef_id=Cj0KCQjws83OBhD4ARIsACblj1-VDFvW9JdDcmqbU5RhEDLEcUwfxYHDDeudvegjmTIdTuNZynlS4JsaAotnEALw_wcB:G:s&s_kwcid=AL!4422!3!795841353823!p!!g!!amazon%20free%20text%20to%20speech!23533257079!191998606999&gad_campaignid=23533257079&gbraid=0AAAADjHtp9g_-6bGryZi1ISluqGvVKxo&gclid=Cj0KCQjws83OBhD4ARIsACblj1-VDFvW9JdDcmqbU5RhEDLEcUwfxYHDDeudvegjmTIdTuNZynlS4JsaAotnEALw_wcB.
38. Galvez, D., V. Bataev, H. Xu, and T. Kaldewey, *Speed of Light Exact Greedy Decoding for RNN-T Speech Recognition Models on GPU.*https://doi.org/10.1007/s00432-026-06432-w

39. Bain, M., J. Huh, T. Han, and A. Zisserman, *WhisperX: Time-Accurate Speech Transcription of Long-Form Audio.*https://doi.org/10.48550/arXiv.2303.00747
40. *Amazon Transcribe Medical*. Available from: https://aws.amazon.com/transcribe/medical/.
41. Barański, M., J. Jasiński, J. Bartolewska, S. Kacprzak, M. Witkowski, and K. Kowalczyk. *Investigation of whisper asr hallucinations induced by non-speech audio*. in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025. IEEE.
42. Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. *Optuna: A next-generation hyperparameter optimization framework*. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019.
43. Reddy, C.K., V. Gopal, and R. Cutler. *DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors*. in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. IEEE.
44. *MedGemma: Our most capable open models for health AI development*. Available from: https://research.google/blog/medgemma-our-most-capable-open-models-for-health-ai-development/.
45. *Introducing Palmyra Med and Palmyra Fin - WRITER*. Available from: https://writer.com/blog/palmyra-med-fin-models/.
46. *Advanced breast cancer: diagnosis and treatment Clinical guideline*. 2009; Available from: www.nice.org.uk/guidance/cg81.
47. *Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer Clinical guideline*. 2013; Available from: www.nice.org.uk/guidance/cg164.
48. *Early and locally advanced breast cancer: diagnosis and management NICE guideline*. 2018; Available from: www.nice.org.uk/guidance/ng101.
49. *Camelot: PDF Table Extraction for Humans — Camelot 1.0.9 documentation*. Available from: https://camelot-py.readthedocs.io/en/master/.
50. Mc, N.Q., *Note on the sampling error of the difference between correlated proportions or percentages.* Psychometrika, 1947. **12**(2): p. 153-7.10.1007/BF02295996
51. Elliott, H., A.J. Allen, N.D. Forester, S. Graziadio, W. Jones, B.C. Lendrem, M.S. Pearce, T. Powell, J. Scott, and A. Bray, *Women's perspectives of molecular breast imaging: a qualitative study.* British journal of cancer, 2025. **132**(3): p. 276-282.https://doi.org/10.1038/s41416-024-02930-1
52. Jones, W.S., J. Suklan, A. Winter, K. Green, T. Craven, A. Bruce, J. Mair, K. Dhaliwal, T. Walsh, and A. Simpson, *Diagnosing ventilator-associated pneumonia (VAP) in UK NHS ICUs: the perceived value and role of a novel optical technology.* Diagnostic and Prognostic Research, 2022. **6**(1): p. 5.https://doi.org/10.1186/s41512-022-00117-x
53. Shi, M., Z. Jin, Y. Xu, Y. Xu, S.-X. Zhang, K. Wei, Y. Shao, C. Zhang, and D. Yu. *Advancing multi-talker ASR performance with large language models*. in *2024 IEEE Spoken Language Technology Workshop (SLT)*. 2024. IEEE.
54. *Somerset cancer register - Somerset Cancer Register*. Available from: https://www.somersetft.nhs.uk/somerset-cancer-register/.

55. Atiku, S., K. Owolanke, and O. Olakotan, *Assessing the Reliability, Accuracy, and Relevance of Artificial Intelligence Speech Recognition for Clinical Documentation: A Scoping Review.* Journal of Evaluation in Clinical Practice, 2026. **32**(4): p. e70460.https://doi.org/10.1111/jep.70460
56. Ellis, Z., J. Joselowitz, Y. Deo, Y.V. He, A. Kalygina, A. Higham, M. Rahimzadeh, Y. Jia, I. Habli, and E. Lim. *Wer is unaware: Assessing how asr errors distort clinical understanding in patient facing dialogue*. in *Proceedings of the 16th International Workshop on Spoken Dialogue System Technology*. 2026.

Supplementary Information

Article Title: Development and Feasibility Evaluation of an Edge AI as Medical Device System for Breast Cancer Multidisciplinary Team Meetings

Journal: Journal of Medical Systems

Authors: Aarzoo Dhiman[1] (A.Dhiman@hull.ac.uk), Farzana Haque[2] (Farzana.Haque4@nhs.net), Kartikae Grover[2] (kartikaegrover@nhs.net), Lydia Brian Smith[3] (L.Bryan-Smith-2014@hull.ac.uk), William Stephen Jones[1] (Will.Jones@hull.ac.uk)

Corresponding author: Aarzoo Dhiman, Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM), Faculty of Science and Engineering, University of Hull, Hull, United Kingdom. Email: a.dhiman@hull.ac.uk

Supplementary Section 1: Rapid Scoping Review: Search Strategy and Evidence Table

1.1 Search Strategy

Database: PubMed/MEDLINE.

Copy-paste executable search string:

("breast cancer" AND ("multidisciplinary team" OR MDT OR "tumor board" OR "tumour board")) AND ("artificial intelligence" OR "machine learning" OR "clinical decision support" OR "large language model" OR LLM OR NLP OR "retrieval augmented generation")

A parallel search was performed for automatic speech recognition studies in breast cancer MDT settings.

1.2 Evidence Table: LLM, RAG and AI Decision Support in Breast Cancer MDT Settings

Supplementary Table 1. Characteristics and reported performance of published AI systems for breast cancer MDT decision support, including system role, study design, evaluation approach, and key outcomes.

Author(s), year	AI system	Role in MDT	Design and evaluation	Outcome
Sorin et al., 2023 [23]	ChatGPT-3.5 using clinical case information	Treatment recommendation and case summarisation for breast tumour board	Proof-of-concept comparison with tumour board decisions (10 breast cancer cases)	70% agreement with tumour board decisions

Author(s), year	AI system	Role in MDT	Design and evaluation	Outcome
Griewing et al., 2023 [57]	ChatGPT-3.5 generating treatment recommendations	Compared with breast cancer tumour board decisions	Observational comparison using diverse breast cancer patient profiles	~50% overall concordance; 58.8% for invasive cases
Griewing et al., 2024 [24]	GPT-4, GPT-3.5, Llama2, Bard	Treatment recommendation support for MDT decisions	Comparison with MDT recommendations, 20 complex breast cancer cases	GPT-4 highest concordance (70.6%); other models lower
Xu et al., 2024 [58]	CSCO AI clinical decision support system	Treatment recommendation support for MDT	Retrospective study of 537 breast cancer patients	92.4% concordance with MDT recommendations
Liao et al., 2025 [59]	ChatGPT-4.0 for breast cancer case recommendations	Compared with expert tumour board recommendations	362 breast cancer cases evaluated by expert tumour board and ChatGPT	46% concordance with experts; 39% reproducibility across responses
Dogan et al., 2025 [60]	ChatGPT-4.0 using clinical case summaries	Treatment recommendation support	Prospective comparison with MDT decisions (100 cancer cases)	76.4% agreement with MDT decisions
Ah-Thiane et al., 2025 [61]	Claude 3 Opus, GPT-4 Turbo, LLaMA3-70B	Treatment recommendation support for MDT decisions	Retrospective comparison with expert MDT decisions (112 early breast cancer cases)	86.6% (Claude 3 Opus), 85.7% (GPT-4 Turbo), 75% (LLaMA3-70B)
Büyükceran et al., 2025 [26]	GPT-4o generating treatment plans from structured clinical data	Decision support for breast cancer MDT	Retrospective comparison with MDT decisions (33 patients)	93.9% full concordance with MDT decisions
Umihanic et al., 2025 [62]	ChatGPT-4.0 generating treatment recommendations from patient data	Decision support for breast cancer MDT	Retrospective comparison with MDT decisions (91 patients)	High agreement (mean score 3.31/4); better in standard cases
Schmutz et al., 2025 [63]	ChatGPT-4.0 generating therapy recommendations	Decision support for molecular tumour board	Retrospective comparison with expert MTB decisions (20 cancer cases)	More treatment suggestions and faster

Author(s), year	AI system	Role in MDT	Design and evaluation	Outcome
	from molecular case data			recommendations; moderate consistency ($\kappa \approx 0.51$)
Zhou et al., 2021 [17]	IBM Watson for Oncology	Treatment recommendation support for MDT	Meta-analysis of 9 studies comparing WFO and MDT decisions (2,463 patients)	81.5% overall concordance with MDT decisions
Moser & Narayan, 2020 [64]	AI-driven predictive care-coordination toolkit	Supports MDT care planning and patient management	Conceptual/implementation discussion	May improve care coordination, risk prediction and patient management; requires human oversight

Supplementary Section 2: Datasets, Models and Full Configuration Results

2.1 Tier 2 Synthetic Voice Configuration (Amazon Polly)

Supplementary Table 2 Amazon Polly voices used to generate synthetic MDT discussions, showing the assigned MDT role, voice profile, language variety, gender, and synthesis type.

MDT role	Polly voice	Accent / language variety	Gender	Neural/Standard
Radiologist	Arthur	British English	Male	Neural
Pathologist	Raveena	Indian English	Female	Standard
Surgeon	Emma	British English	Female	Neural
Oncologist	Aditi	Indian bilingual English	Female	Standard
Breast care nurse	Kajal	Indian bilingual English	Female	Neural
MDT coordinator	Brian	British English	Male	Neural

2.2 Tier 2 Dataset Composition

Supplementary Table 3 Characteristics of the ten synthetic MDT discussion scenarios, including discussion duration and multidisciplinary speaker roles represented in each case.

Case ID	Duration (s)	Speaker roles represented
1	29	Radiologist, pathologist, surgeon, MDT coordinator
2	38	Radiologist, pathologist, surgeon, oncologist, breast care nurse, MDT coordinator
3	60	Radiologist, pathologist, surgeon, oncologist, MDT chair (consultant surgeon), MDT coordinator
4	29	Oncologist, radiologist, pathologist, palliative care specialist, MDT coordinator
5	30	Radiologist, pathologist, surgeon, MDT coordinator, chair
6	44	Radiologist, pathologist, oncologist, surgeon, breast care nurse, MDT coordinator, chair
7	38	Radiologist, pathologist, surgeon, oncologist, chair, MDT coordinator
8	29	Radiologist, pathologist, surgeon, breast care nurse, MDT coordinator, chair
9	23	Surgeon, pathologist, oncologist, MDT coordinator, chair
10	32	Radiologist, pathologist, oncologist, palliative care specialist, breast care nurse, MDT coordinator, chair

2.3 Acoustic Augmentation Parameters

Each of the ten Tier 2 synthetic recordings was augmented across all combinations of the techniques below, producing 127 variants per recording and 1,270 augmented files in total. Severely degraded configurations were retained deliberately to permit stress testing at the limits of intelligibility.

Supplementary Table 4 Acoustic data augmentation techniques applied to synthetic MDT recordings, including implementation methods and parameter ranges used to generate the augmented evaluation dataset.

Technique	Implementation	Parameter values
Background noise	Mixing with environmental recordings (keyboard typing, office ambience, paper handling, small-room conversation, crowded meeting-room speech) scaled to target SNR	SNR 0, 5, 10, 15 dB
Room impulse response	Convolution with RIR filters to replicate meeting-room reverberation	audio files with RIR used
Speech overlap	Insertion of randomly selected segments from secondary audio files into the primary signal	Overlap ratio 0.1, 0.15, 0.3; segment length 1–20 s; 2–10 segments per

Technique	Implementation	Parameter values
		recording; mixing volume ratio 0.1, 0.4, 0.8
Speed perturbation	Resampling to produce faster/slower versions	Speed factors 0.95, 1.3, 1.5, 2.0
Pitch shifting	Pitch shift without duration change	−2, 0, +2 semitones
Audio dropout	Chunk-based zeroing of waveform segments to simulate recording interruption	Dropout probability 0.05, 0.1

All augmented audio was produced in the same format as the source synthetic audio (16 kHz, 16-bit, mono).

2.4 ASR Models Evaluated

Supplementary Table 5 Automatic speech recognition (ASR) systems evaluated in this study, summarising model characteristics, intended capabilities, and role within the evaluation pipeline.

Model	Description
Whisper large-v3	Open-source encoder–decoder ASR model (~1.55B parameters) supporting multilingual transcription and translation, using 128 Mel-frequency filter banks, with reported robustness to background noise, overlapping speech and domain-specific terminology
NVIDIA Parakeet	Family of GPU-optimised ASR models (0.6–1.1B parameters) designed for high-throughput inference, with competitive WER, faster inference than Whisper, and native punctuation and timestamp generation
WhisperX	Extension of Whisper adding phoneme-level alignment and speaker diarization, improving timestamp precision and enabling multi-speaker attribution, with batched inference and alignment correction via secondary models. Built on Whisper large-v2
Amazon Transcribe Medical	HIPAA-eligible cloud ASR service for clinical documentation, with strong medical vocabulary support. Requires cloud processing and incurs ongoing operational cost, and was therefore used as a benchmark only

2.5 Whisper Large-v3 Decoding Parameters Optimised

Optimised via Optuna (Tree-structured Parzen Estimator, 100 trials, parallel search, objective: minimise WER on Tier 3 augmented audio): temperature, beam_size, best_of, compression_ratio_threshold, logprob_threshold, length_penalty, hallucination_silence_threshold.

Supplementary Table 6 Optimised Whisper large-v3 decoding parameters identified through Bayesian hyperparameter optimisation and used for adaptive transcription.

Parameter name	Optimised Value
temperature	0.7596350092541306
beam_size	2

best_of	2
compression_ratio_threshold	1.5323007177495684
logprob_threshold	-1.0903724240294121
length_penalty	0.7594738864766823
hallucination_silence_threshold	0.22902235239020646

Silence detection: FFmpeg silencedetect filter, −23 dB threshold, segments longer than 5s removed with 2s of surrounding silence retained before concatenation.

2.6 Full DNSMOS Threshold Sweep on Authentic Recordings (Tier 1)

Supplementary Table 7 Word error rate (WER) and word information lost (WIL) of Whisper large-v3 across default and optimised transcription configurations on the two held-out authentic MDT recordings (Recording 1: R1 and Recording 2: R2).

Model configuration	R1 WIL	R1 WER	R2 WIL	R2 WER
Whisper default	0.7252	0.6181	0.3412	0.2738
Default + silence removed	0.7184	0.5938	0.3567	0.2649
Tuned parameters	0.7451	0.5761	0.3813	0.2445
Silence removed + tuned	0.6941	0.5342	0.3265	0.2210
Tuned + noise threshold 2.5	0.7005	0.5221	0.3293	0.2624
Tuned + noise threshold 3.0	0.7473	0.5651	0.3483	0.2337
Tuned + noise threshold 3.5	0.6971	0.5198	0.3467	0.2223
Silence removed + tuned + threshold 2.5	0.6900	0.5088	0.3457	0.2726
Silence removed + tuned + threshold 3.0	0.7474	0.5827	0.2953	0.1878
Silence removed + tuned + threshold 3.5	0.6571	0.4901	0.3281	0.2070

Supplementary Section 3: ASR Evaluations

3.1 Evaluation Metric

$$\mathrm{WER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}}$$

Where: S = number of substitution, D = number of deletions, I = number of insertions, and N = total number of words in the reference (ground truth).

$$\mathrm{WIL} = 1 - \frac{\mathrm{H}^2}{\mathrm{N} \cdot \mathrm{M}}$$

Where: H = number of correctly recognised words (hits), N = total number of words in the reference, M = total number of words in the hypothesis.

3.2 Curated Clinical Phrase List for ASR Terminology Evaluation

Phrases were extracted from the reference transcripts of each case and matched against ASR output using exact and fuzzy (Levenshtein similarity ≥ 80%, FuzzyWuzzy) comparison of 1–6-grams. Recall was computed as TP / (TP + FN). Precision was not computed because candidate terms were restricted to this predefined vocabulary.

Case 1: 56; routine screening; 8 mm; lesion; right upper outer quadrant; invasive ductal carcinoma; BI-RADS 5; core biopsy; grade 1; ER positive; HER2 negative; straightforward lumpectomy; nodal involvement; wide local excision; sentinel node biopsy

Case 2: 42; palpable lump; MRI; multifocal disease; 6 cm; axillary nodes; biopsy; triple negative invasive carcinoma; grade 3; nodal core also positive; too extensive; breast conserving surgery; neoadjuvant chemotherapy; anxious; downstage; reassess surgical options

Case 3: 68; 61; male breast cancer; imaging; 2 cm lesion; no nodes; ER positive; HER2 negative; simple mastectomy; endocrine therapy; surgery if fitness is confirmed; survival benefit; anaesthetic review; surgical referral

Case 4: 72; surgery; chemotherapy; liver metastases; follow-up scans; bilateral liver lesions; not resectable; HER2 positive; HER2 targeted therapy; prognosis is guarded; plan supportive care; referral to palliative team; close monitoring

Case 5: 56; routine screening; 8 mm lesion; right breast; upper outer quadrant; malignant; BI-RADS 5; core biopsy; invasive ductal carcinoma; grade 1; ER positive; HER2 negative; wide local excision; sentinel node biopsy; lumpectomy

Case 6: 42; palpable lump; symptomatic; MRI; multifocal disease; 6 cm span; abnormal axillary nodes; triple negative invasive carcinoma; grade 3; nodal core also positive; neoadjuvant chemotherapy; breast conserving surgery not feasible; anxious; reassess surgery after response; psychological support; breast care team

Case 7: 68; male breast cancer; 2 cm lesion; no nodes; ER positive; HER2 negative; simple mastectomy; endocrine therapy; COPD; cardiac history; surgery risk is high; offers better local control; refer to anaesthetics; surgical referral; endocrine therapy if unfit

Case 8: 33; family history; genetic screen positive; MRI; 6 mm lesion; right breast; biopsy; high grade DCIS; BRCA1 mutation; bilateral mastectomy; risk reducing surgery; refer to genetics; plan for bilateral mastectomy; reconstruction discussion

Case 9: 61; lumpectomy; tumour 1.2 cm; grade 2; ER positive; HER2 negative; margins clear; no nodal involvement; adjuvant endocrine therapy; no chemotherapy needed; endocrine therapy; routine follow-up

Case 10: 72; post surgery; chemotherapy; multiple liver metastases; HER2 positive; HER2 targeted therapy; prognosis is poor; needs early palliative involvement; symptom control; fatigue; mobility issues; palliative care referral; supportive measures

NOTE. Spellings have been corrected relative to the original working list ("gaurded" → "guarded", "reasses" → "reassess", "BIRADS" → "BI-RADS", "HER 2" → "HER2"). If the matching code used the original strings verbatim, either re-run matching with the corrected list or state explicitly that matching used the working spellings, since fuzzy matching at 80% similarity is sensitive to these variants.

Supplementary Section 4: Prompting Strategy Used

4.1 Prompt: Case Description and Treatment Separation

```
You are a clinical MDT transcription assistant specialized in breast oncology.
Your tasks:
1. Consider the given text as the MDT meeting discussion transcript of a single case.
2. For the given case, you may clean and normalise the language by correcting spelling, grammar and punctuations without including/excluding any information. with your breast cancer expertise, you have to clean the medical terminologies wherever possible, by correcting ASR phonetic drift into medically valid terms.
3. For the case, you have to extract the case description, you MUST NOT include or modify any information not found in the given text. You HAVE TO only use the information provided in the text with corrections.
4. The case description may include patient's features/conditions like, age, gender, demographics, diagnosis, pathology, imaging, nodal status, biomarkers etc.
5. For each case, you also have to extract the Treatment Decisions made during the MDT meeting, using this text. you MUST NOT include or modify any information not found in the given text. You HAVE TO only use the information provided in the text.
6. The treatment decisions may include agreed MDT plan, investigations, referrals, follow-up actions etc.
7. For both case description and treatment plan, you have to use the EXACT VERBATIM from the provided text, without any structural modifications.
8. IF the details are not clear, mark it as "UNCLEAR in transcript".
Output format (STRICT):
For each case, use the following structure:
CASE X:
Corrections made:
Case Description:
Treatment Decisions:
==============================
EXAMPLE:
!---------------Example not provided for data privacy-------------!
"""
```

A one-shot worked example was appended to this prompt, comprising a sample noisy transcript excerpt and the expected corrected output.

4.2 Prompt: Treatment Prediction without Guideline Context

```
SYSTEM ROLE:
You are a clinical decision-support assistant for multidisciplinary team (MDT) meetings.
You do NOT provide medical advice.
You summarize cases and map them to high-level NICE guideline pathways.
TASK:
From the transcript below:
1. Understand the case description carefully, ONLY consider information that is present.
2. Propose a TREATMENT PLAN that is CONSISTENT WITH NICE BREAST CANCER GUIDELINES
RULES:
- Use ONLY information explicitly stated in the transcript
- Do NOT assume staging, performance status, or comorbidities
- Do NOT invent biomarkers, imaging results, or pathology
- If required information is missing, write: INSUFFICIENT INFORMATION
- Treatment plans must be guideline-aligned, not personalized prescriptions
- Use conditional language: "may be considered", "is typically recommended", "depending on eligibility"
OUTPUT FORMAT:
  "Case Description":
  "Proposed Treatment Plan":
  "guideline basis":
  "confidence": "low | medium | high"
IMPORTANT:
- If multiple patients are present, output an ARRAY (one per case)
- If no clear treatment decision can be derived, populate treatment_plan with "INSUFFICIENT INFORMATION"
==========
EXAMPLE:
!---------------Example not provided for data privacy-------------!
"""
```

4.3 Prompt: Treatment Prediction with Retrieved Guideline Context (RAG)

4.3.1 Free-text Treatment Prediction 1 (FT1)

```
f"""
SYSTEM ROLE:
You are a clinical decision-support assistant for multidisciplinary team (MDT) meetings.
You do NOT provide medical advice.
You summarize cases and map them to high-level NICE guideline pathways.
TASK:
From the transcript below:
1. Understand the case description carefully, ONLY consider information that is present.
2. Propose a TREATMENT PLAN that is CONSISTENT WITH NICE BREAST CANCER GUIDELINES
RULES:
- Use ONLY information explicitly stated in the transcript
- Do NOT assume staging, performance status, or comorbidities
- Do NOT invent biomarkers, imaging results, or pathology
- If required information is missing, write: INSUFFICIENT INFORMATION
- Treatment plans must be guideline-aligned, not personalized prescriptions
- Use conditional language: "may be considered", "is typically recommended", "depending on eligibility"
OUTPUT FORMAT:
       "Case Description":
       "Proposed Treatment Plan":
       "guideline basis":
       confidence": "low | medium | high"

IMPORTANT:
- If multiple patients are present, output an ARRAY (one per case)
- If no clear treatment decision can be derived, populate treatment_plan with "INSUFFICIENT INFORMATION"
==========================================
EXAMPLE:
!---------------Example not provided for data privacy-------------!
"""
```

4.3.2 Free-text Treatment Prediction 2 (FT2)

```
SYSTEM ROLE:
```

You are a clinical decision-support assistant for multidisciplinary team (MDT) meetings.
You do NOT provide medical advice.
Extract the MDT's explicit clinical reasoning and decision nodes, and align each to relevant NICE guidance where applicable.
TASK:
From the transcript below:

1. Extract the MDT's explicit clinical reasoning and key decision drivers.
2. Reconstruct the proposed treatment strategy exactly as discussed (including sequencing, conditional plans, and referrals).
3. Then verify that the strategy is consistent with relevant NICE guidance.

MDT REASONING REQUIREMENTS:

In your Proposed Treatment Plan:

- Explicitly identify timing decisions (e.g., neoadjuvant vs adjuvant)
- Explicitly identify biomarker-dependent branching (e.g., PD-L1, BRCA)
- Explicitly identify surgical morbidity trade-offs if discussed
- Explicitly identify when genetic results influence surgical planning
- Explicitly identify MDT referrals (e.g., bone MDT)
- Explicitly identify when staging is NOT required
- Preserve uncertainty or conditional logic expressed by the MDT
- Do NOT collapse strategic reasoning into generic pathway summaries.

RULES:
- Use ONLY information explicitly stated in the transcript
- If the MDT explicitly rejects a standard step (e.g., "no staging required"), preserve that decision and do not reintroduce it.
- Do NOT invent biomarkers, imaging results, or pathology
- If required information is missing, write: INSUFFICIENT INFORMATION
- Treatment plans must be guideline-aligned, not personalized prescriptions
- Use conditional language: "may be considered", "is typically recommended", "depending on eligibility"
- If the transcript contains biomarker testing or pending results, structure the treatment plan with explicit IF/THEN branches.
- Do NOT default to surgery-first or adjuvant-first pathways unless explicitly supported by the transcript.

OUTPUT FORMAT:
"Case Description":
"Proposed Treatment Plan":

```
 "Key Decision Drivers":
- (bullet points listing what determined the plan)
  "guideline basis":
  "confidence": "low | medium | high"

IMPORTANT:
- If multiple patients are present, output an ARRAY (one per case)
- If no clear treatment decision can be derived, populate treatment_plan with "INSUFFICIENT INFORMATION"

---

EXAMPLE:
!---------------Example not provided for data privacy-------------!
"""
```

Retrieved context: the top five guideline chunks returned by FAISS similarity search over the embedded NICE corpus were inserted into the prompt. Generation used fixed decoding parameters (temperature 0.1, top_p 0.9, repeat_penalty 1.1, top_k 40, max_tokens 1000).

4.4 Prompt: Intervention-based Structured Prediction

```
SYSTEM ROLE:
You are a clinical decision-support assistant for multidisciplinary team (MDT) meetings.
Extract the MDT's explicit clinical reasoning and decision nodes, and align each to relevant NICE guidance where applicable.
TASK:
From the transcript below:
1. Understand the case description carefully and the context from the NICE guidelines, ONLY consider information that is present.
2. You will be given a list of possible treatment options that may or may not be applicable to the case description. Your task is to label the treatment option as (i) YES (ii) NO, and (iii) NEI (Not Enough Information) with a brief outline for the reason of the label made.
3. Then verify that the strategy is consistent with relevant NICE guidance.
Below is the list of treatment options that you have to label. Use the full list, do not add  or remove any option from the list while doing the labelling.
Treatment options (label each as YES/NO/NEI):
1.        Surgical Procedures:
1.1.  Diagnostic & Image-Guided Procedures
1.1.1.      Biopsies (core, vacuum, excisional, punch)
1.1.2.      Localisation techniques (wire, seed, magnetic, ultrasound-guided)
```

1.2. Breast-Conserving Surgery (BCS)

1.2.1. Lumpectomy / Wide Local Excision

1.2.2. Margin re-excision

1.2.3. Oncoplastic Breast-Conserving Surgery

1.3. Mastectomy

1.3.1. Simple, skin-sparing, nipple-sparing

1.3.2. Modified radical

1.3.3. Skin-reducing mastectomy

1.4. Risk-Reducing & Prophylactic Surgery

1.4.1. Bilateral risk-reducing mastectomy

1.4.2. Contralateral prophylactic mastectomy

1.5. Axillary Staging, treatment

1.5.1. Sentinel lymph node biopsy (SLNB)

1.5.2. Axillary lymph node dissection (ALND)

1.5.3. Targeted axillary procedures

1.6. Breast Reconstruction (Oncoplastic & Post-Mastectomy)

1.6.1. Implant-based reconstruction

1.6.2. Autologous flaps (DIEP, LD)

1.6.3. Immediate & delayed reconstruction

1.7. Nipple/Areola Reconstruction

1.7.1. Nipple reconstruction (local flaps, grafts)

1.7.2. Areola tattooing

1.8. Recurrent & Palliative Breast Surgery

1.8.1. Chest wall resection

1.8.2. Surgery for recurrence

1.8.3. Palliative procedures

2. Imaging

2.1. Mammography

2.2. Contrast-Enhanced Mammography (CEM)

2.3. Breast Ultrasound (US)

2.4. Breast MRI

2.5. FDG Positron Emission Tomography-CT (PET-CT)

2.6. Sodium Fluoride PET-CT

2.7. CT Scan of Chest/Abdomen/Pelvis

2.8. Bone Scan (Scintigraphy)

3. Radiotherapy

3.1. Adjuvant Whole Breast Irradiation/ Chest Wall Irradiation (post-mastectomy radiotherapy, PMRT)

3.2. Tumour Bed Boost

3.3. Partial Breast Irradiation (PBI / APBI)

3.4. Regional Nodal Irradiation (RNI)

3.4.1. Axillary irradiation

3.4.2. Supraclavicular nodal irradiation

3.4.3. Internal mammary node (IMN) irradiation

3.5. Stereotactic & Special Techniques- Stereotactic body radiotherapy (SBRT) for oligometastases

3.6. Palliative Radiotherapy

3.6.1. Palliative breast/chest wall irradiation

3.6.2. Radiotherapy for bone metastasis

3.6.3. Radiotherapy to brain metastasis (WBRT / SRS)

3.7. No need for radiotherapy

4. Systemic treatment

4.1. Chemotherapy- Neoadjuvant/adjuvant/palliative chemotherapy

4.2. Endocrine (Hormonal) Therapy

4.2.1. Anti-estrogen therapy- Tamoxifen, Aromatase inhibitors

4.2.2. Ovarian suppression/ablation

4.3. Targeted Therapy (Biologic Therapy)

4.3.1. HER2-targeted therapy

4.3.2. CDK4/6 inhibitors

4.3.3. PI3K / mTOR inhibitors

4.3.4. Antibody-drug conjugates (ADCs)

4.3.5. PARP inhibitors (For BRCA-positive patients)

4.3.6. NTRK targeted (NTRK fusion positive patients)

4.4. Immunotherapy

4.5. Bone-Targeted Therapy- Bisphosphonates, RANKL inhibitors (denosumab)

4.6. Electrochemotherapy

4.7. Supportive treatments

5. Useful Tools

5.1. Prognostic & Treatment Benefit Calculators

5.1.1. PREDICT tool

5.1.2. Nottingham Prognostic Index (NPI)

5.2. Biomarkers

5.2.1. ER / PR status

5.2.2. HER2 status

5.2.3. Ki-67 proliferation index

5.2.4. PD-L1 expression

5.3. Genomic / Molecular Assays- Oncotype Dx

5.4. Common terms used- TNM staging, Tumour grade (Nottingham grading system), Nodal status

6. Referrals

6.1. Referral to other MDT

6.2. Refer to Oncology Health

6.3. Refer for clinical trials

7. Appointments

7.1. Clinical Oncology appointment

7.2. Medical Oncology appointment

LABELING RULES:

Answer YES if the option is explicitly: should be (done/performed OR recommended/planned as next management, OR
requested/accepted as the intended plan) AND aligned with (the NICE guidelines, OR per your existing knowledge).

Answer NO if the option is:

should NOT be (done/performed OR recommended/planned as next management, OR
requested/accepted as the intended plan) OR NOT aligned with (the NICE guidelines, OR per your existing knowledge). OR
should be explicitly rejected/declined/ruled out.

Answer Not Enough Information if the option could be considered but you can not gather enough information from the case description or the NICE guidelines required to choose this option.

ANSWER RULES:

- Use ONLY information explicitly stated in the transcript
- If the MDT explicitly rejects a standard step (e.g., “no staging required”), preserve that decision and do not reintroduce it.
- Do NOT invent biomarkers, imaging results, or pathology
- If required information is missing, write: INSUFFICIENT INFORMATION
- Use conditional language: "may be considered", "is typically recommended", "depending on eligibility"
- If the transcript contains biomarker testing or pending results, structure the treatment plan with explicit IF/THEN branches.
- Do NOT default to surgery-first or adjuvant-first pathways unless explicitly supported by the transcript.

INTERNAL REASONING POLICY:

- Perform reasoning internally
- Do not output reasoning
-DO not output immediate steps
- Do not explain decisions
- Only output final labels.

OUTPUT REQUIREMENTS:

- DO NOT INCLUDE ANY explanations, reasoning, chain-of-thought in the response.
- Return lables only.

```
==========================================
EXAMPLE:
!---------------Example not provided for data privacy-------------!
"""
```

Supplementary Section 5: LLM Decoding Hyperparameters

Supplementary Table 8 Inference parameters used for MedGemma 27B and Palmyra-Med 70B during case extraction and treatment separation.

Model	Parameters (case extraction and treatment separation)
MedGemma 27B	temperature 0.05; top_p 0.7; min_p 0.05; repeat_penalty 1.2; frequency_penalty 0.3; typical_p 0.7; top_k 20; mirostat_mode 0; tfs_z 1.0; max_tokens 4000
Palmyra-Med 70B	temperature 0.02; top_p 0.6; min_p 0.15; repeat_penalty 1.15; frequency_penalty 0.1; typical_p 0.7; top_k 15; mirostat_mode 0; tfs_z 1.0

For guideline-grounded treatment generation, MedGemma was run with fixed parameters: temperature 0.1; top_p 0.9; repeat_penalty 1.1; top_k 40; max_tokens 1000.

Supplementary Section 6: Intervention Taxonomy for Structured Prediction

The taxonomy below was supplied to the model in full for every case. The model classified each leaf item as Yes, No or NEI.

1. Surgical Procedures:
1.1. Diagnostic & Image-Guided Procedures
1.1.1. Biopsies (core, vacuum, excisional, punch)
1.1.2. Localisation techniques (wire, seed, magnetic, ultrasound-guided)
1.2. Breast-Conserving Surgery (BCS)
1.2.1. Lumpectomy / Wide Local Excision
1.2.2. Margin re-excision
1.2.3. Oncoplastic Breast-Conserving Surgery
1.3. Mastectomy
1.3.1. Simple, skin-sparing, nipple-sparing
1.3.2. Modified radical
1.3.3. Skin-reducing mastectomy
1.4. Risk-Reducing & Prophylactic Surgery
1.4.1. Bilateral risk-reducing mastectomy
1.4.2. Contralateral prophylactic mastectomy
1.5. Axillary Staging, treatment
1.5.1. Sentinel lymph node biopsy (SLNB)

1.5.2. Axillary lymph node dissection (ALND)

1.5.3. Targeted axillary procedures

1.6. Breast Reconstruction (Oncoplastic & Post-Mastectomy)

1.6.1. Implant-based reconstruction

1.6.2. Autologous flaps (DIEP, LD)

1.6.3. Immediate & delayed reconstruction

1.7. Nipple/Areola Reconstruction

1.7.1. Nipple reconstruction (local flaps, grafts)

1.7.2. Areola tattooing

1.8. Recurrent & Palliative Breast Surgery

1.8.1. Chest wall resection

1.8.2. Surgery for recurrence

1.8.3. Palliative procedures

2. Imaging

2.1. Mammography

2.2. Contrast-Enhanced Mammography (CEM)

2.3. Breast Ultrasound (US)

2.4. Breast MRI

2.5. FDG Positron Emission Tomography-CT (PET-CT)

2.6. Sodium Fluoride PET-CT

2.7. CT Scan of Chest/Abdomen/Pelvis

2.8. Bone Scan (Scintigraphy)

3. Radiotherapy

3.1. Adjuvant Whole Breast Irradiation/ Chest Wall Irradiation (post-mastectomy radiotherapy, PMRT)

3.2. Tumour Bed Boost

3.3. Partial Breast Irradiation (PBI / APBI)

3.4. Regional Nodal Irradiation (RNI)

3.4.1. Axillary irradiation

3.4.2. Supraclavicular nodal irradiation

3.4.3. Internal mammary node (IMN) irradiation

3.5. Stereotactic & Special Techniques- Stereotactic body radiotherapy (SBRT) for oligometastases

3.6. Palliative Radiotherapy

3.6.1. Palliative breast/chest wall irradiation

3.6.2. Radiotherapy for bone metastasis

3.6.3. Radiotherapy to brain metastasis (WBRT / SRS)

3.7. No need for radiotherapy

4. Systemic treatment

4.1. Chemotherapy- Neoadjuvant/adjuvant/palliative chemotherapy

4.2. Endocrine (Hormonal) Therapy

4.2.1. Anti-estrogen therapy- Tamoxifen, Aromatase inhibitors

4.2.2. Ovarian suppression/ablation

4.3. Targeted Therapy (Biologic Therapy)

4.3.1. HER2-targeted therapy

4.3.2. CDK4/6 inhibitors

4.3.3. PI3K / mTOR inhibitors

4.3.4. Antibody-drug conjugates (ADCs)

4.3.5. PARP inhibitors (For BRCA-positive patients)

4.3.6. NTRK targeted (NTRK fusion positive patients)

4.4. Immunotherapy

4.5. Bone-Targeted Therapy- Bisphosphonates, RANKL inhibitors (denosumab)

4.6. Electrochemotherapy

4.7. Supportive treatments

5. Useful Tools

5.1. Prognostic & Treatment Benefit Calculators

5.1.1. PREDICT tool

5.1.2. Nottingham Prognostic Index (NPI)

5.2. Biomarkers

5.2.1. ER / PR status

5.2.2. HER2 status

5.2.3. Ki-67 proliferation index

5.2.4. PD-L1 expression

5.3. Genomic / Molecular Assays- Oncotype Dx

5.4. Common terms used- TNM staging, Tumour grade (Nottingham grading system), Nodal status

6. Referrals

6.1. Referral to other MDT

6.2. Refer to Oncology Health

6.3. Refer for clinical trials

7. Appointments

7.1. Clinical Oncology appointment

7.2. Medical Oncology appointment

Supplementary Section 7: LLM Prediction Expert Evaluation Instrument and Full Statistical Comparison

7.1 Expert Evaluation Instrument (Open-Ended Generation)

Step 1. Identify the interventions

1. Identify all the interventions listed in the MDT GT
2. Identify all the interventions proposed by each LLM

Step 2. Compare the interventions

```
For each intervention proposed by the LLM, answer the following:
Q1. Does the LLM Intervention match the intervention in the MDT GT? (Yes / No)
Q2. If the intervention does not match the MDT plan:
      Is it a recognised, real clinical intervention (or made up / hallucinated)?
Q3. If the intervention does not match the MDT plan:
      Could it still be clinically appropriate for this specific case?
Q4. Are ALL MDT GT interventions present in the LLM response? (Yes / No)
      If 'No': list the interventions omitted
```

Metric definitions, where *Relevant* denotes interventions discussed in the MDT and *Retrieved* denotes interventions generated by the model:

- $\text{Precision} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|}$
- $\text{Recall} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Relevant}|}$
- $\text{Jaccard Coeff} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Relevant} \cup \text{Retrieved}|}$
- $\text{Overgeneration} = \frac{\text{FP}}{|\text{Retrieved}|}$ (from Q1)
- $\text{Miss rate} = \frac{|\text{Missed Interventions}|}{|\text{Relevant}|}$ (from Q4)
- $\text{Hallucination rate} = \frac{|\text{Hallucinated Interventions}|}{|\text{Retrieved}|}$ (from Q2)
- $\text{proportion of apt interventions} = \frac{|\text{nonMDT but appropriate}|}{|\text{clinically appropriate interventions}|}$ (from Q3)

7.2 Full Paired Statistical Comparison (Intervention-based Strategy)

McNemar's test (MN) was used for overall accuracy and paired bootstrap resampling (B = 5,000; BS) for class-specific precision, recall, and F1 score. Differences are reported as MedGemma-RAG (MR) minus ChatGPT-5.2 (CR); n = 396 paired predictions. Accuracy, precision, recall, and F1 score are reported using standard machine learning definitions.

Supplementary Table 9 Paired statistical comparison of MedGemma-RAG (MR) and ChatGPT-5.2 (CR) for intervention-based structured prediction under three NEI scoring schemes, reporting overall accuracy and class-specific performance with bootstrap confidence intervals and significance testing.

Scenario	Class	Metric	MR	CR	Diff	95% CI	p	Test	Sig.
A (NEI=No)	overall	Accuracy	83.6%	80.8%	+2.8 pp	—	0.1093	MN	No
B (NEI=Yes)	overall	Accuracy	60.1%	61.4%	−1.3 pp	—	0.6301	MN	No
3-class	overall	Accuracy	54.0%	52.5%	+1.5 pp	—	0.5611	MN	No
A (NEI=No)	Yes	Precision	0.286	0.136	+0.149	0.031–0.275	0.0156	BS	Yes
A (NEI=No)	Yes	Recall	0.318	0.136	+0.182	0.041–0.327	0.0200	BS	Yes
A (NEI=No)	Yes	F1	0.301	0.136	+0.165	0.038–0.292	0.0136	BS	Yes

Scenario	Class	Metric	MR	CR	Diff	95% CI	p	Test	Sig.
A (NEI=No)	No	Precision	0.914	0.892	+0.021	0.004–0.040	0.0136	BS	Yes
A (NEI=No)	No	Recall	0.901	0.892	+0.009	−0.020–0.037	0.6120	BS	No
A (NEI=No)	No	F1	0.907	0.892	+0.015	−0.003–0.033	0.0972	BS	No
B (NEI=Yes)	Yes	Precision	0.228	0.245	−0.017	−0.047–0.012	0.2336	BS	No
B (NEI=Yes)	Yes	Recall	0.830	0.906	−0.075	−0.180–0.023	0.1964	BS	No
B (NEI=Yes)	Yes	F1	0.358	0.386	−0.028	−0.072–0.016	0.1964	BS	No
B (NEI=Yes)	No	Precision	0.956	0.975	−0.019	−0.047–0.006	0.1496	BS	No
B (NEI=Yes)	No	Recall	0.566	0.569	−0.003	−0.049–0.041	0.9384	BS	No
B (NEI=Yes)	No	F1	0.711	0.718	−0.008	−0.046–0.030	0.6588	BS	No
3-class	Yes	Precision	0.286	0.136	+0.149	0.031–0.275	0.0156	BS	Yes
3-class	Yes	Recall	0.318	0.136	+0.182	0.041–0.327	0.0200	BS	Yes
3-class	Yes	F1	0.301	0.136	+0.165	0.038–0.292	0.0136	BS	Yes
3-class	No	Precision	0.956	0.975	−0.019	−0.047–0.006	0.1496	BS	No
3-class	No	Recall	0.566	0.569	−0.003	−0.049–0.041	0.9384	BS	No
3-class	No	F1	0.711	0.718	−0.008	−0.046–0.030	0.6588	BS	No
3-class	NEI	Precision	0.042	0.046	−0.004	−0.020–0.007	0.6532	BS	No
3-class	NEI	Recall	0.667	0.778	−0.111	−0.364–0.000	0.7400	BS	No
3-class	NEI	F1	0.078	0.087	−0.009	−0.038–0.011	0.6420	BS	No

Supplementary Section 8: DECIDE-AI Alignment and Deferred Items

DECIDE-AI is the reporting guideline for early live clinical evaluation of AI-based decision support. This study is bench (non-clinical) evaluation preceding that stage. The table records which items are addressed and which are deferred, so that reviewers and readers can locate the present work on the evaluation pathway.

Supplementary Table 10 Alignment of the present study with the DECIDE-AI framework, showing which reporting domains were addressed during analytical validation and bench feasibility testing and which are deferred to prospective clinical evaluation.

DECIDE-AI domain	Status in this study	Where addressed / why deferred
Intended use and clinical pathway position	Addressed	Introduction; Discussion §4.3. Intended as documentation support and case triage under clinician review, not autonomous recommendation

DECIDE-AI domain	Status in this study	Where addressed / why deferred
AI system description (architecture, versions, hardware)	Addressed	Methods §2.2, §2.5; Supplementary Section 3, 5, 6
Training and input data provenance	Addressed	Methods §2.3; Supplementary Section 2, 3
Algorithm performance (analytical validation)	Addressed	Results §3.1–§3.5
Comparator	Addressed	Amazon Transcribe Medical (ASR); ChatGPT-5.2 (generation)
Calibration	**Deferred**	Not assessed; stated as a limitation (§4.4). Required before recommendation-facing use
Users and training of users	**Deferred**	No users operated the system; stakeholder consultation was conceptual
Implementation and integration into the clinical pathway	**Deferred**	No live deployment; integration requirements identified in §4.3
Human factors and modification of user behaviour	**Deferred**	Requires live evaluation
Safety and errors in clinical use	Partially addressed	Hallucination and overgeneration measured on bench data; clinical-use safety requires prospective study
Patient-related outcomes	**Deferred**	No patients involved; no clinical outcomes measurable at this stage
Ethics and governance	Addressed	Declarations; §4.3 (recording governance, information governance rationale for on-device inference)

References

1. Patkar, V., D. Acosta, T. Davidson, A. Jones, J. Fox, and M. Keshtgar, *Cancer multidisciplinary team meetings: evidence, challenges, and the role of clinical decision support technology.* International journal of breast cancer, 2011. **2011**(1): p. 831605. https://doi.org/10.4061/2011/831605
2. *Breast cancer statistics | Cancer Research UK*. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.
3. Winters, D.A., T. Soukup, N. Sevdalis, J.S. Green, and B.W. Lamb, *The cancer multidisciplinary team meeting: in need of change? History, challenges and future perspectives.* BJU international, 2021. **128**(3): p. 271-279.https://doi.org/10.1111/bju.15495Digital

4. Rajan, S., J. Foreman, M. Wallis, C. Caldas, and P. Britton, *Multidisciplinary decisions in breast cancer: does the patient receive what the team has recommended?* British journal of cancer, 2013. **108**(12): p. 2442-2447.https://doi.org/10.1038/bjc.2013.267
5. *Best Practice diagnostic guidelines for patients presenting with breast cancer symptoms | Association of Breast Surgery*. Available from: https://associationofbreastsurgery.org.uk/professionals/information-hub/guidelines/2010/best-practice-diagnostic-guidelines-for-patients-presenting-with-breast-cancer-symptoms.
6. Soukup, T., T.A. Gandamihardja, S. McInerney, J.S. Green, and N. Sevdalis, *Do multidisciplinary cancer care teams suffer decision-making fatigue: an observational, longitudinal team improvement study.* BMJ open, 2019. **9**(5): p. e027303.https://doi.org/10.1136/bmjopen-2018-027303
7. Gandamihardja, T.A., T. Soukup, S. McInerney, J. Green, and N. Sevdalis, *Analysing breast cancer multidisciplinary patient management: a prospective observational evaluation of team clinical decision-making.* World journal of surgery, 2019. **43**(2): p. 559-566. https://doi.org/10.1007/s00268-018-4815-3
8. Improvement, N.H.S. *Streamlining Multi-Disciplinary Team Meetings Guidance for Cancer Alliances*. Available from: https://www.england.nhs.uk/publication/streamlining-mdt-meetings-guidance-cancer-alliances/.
9. Soukup, T., B.W. Lamb, A. Morbi, N.J. Shah, A. Bali, V. Asher, T. Gandamihardja, P. Giordano, A. Darzi, and N. Sevdalis, *Cancer multidisciplinary team meetings: impact of logistical challenges on communication and decision-making.* BJS open, 2022. **6**(4): p. zrac093.https://doi.org/10.1093/bjsopen/zrac093
10. Gommers, J., V. Hernström, V. Josefsson, H. Sartor, D. Schmidt, A. Hjelmgren, A.-M. Larsson, S. Hofvind, I. Andersson, and A. Rosso, *Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial.* The Lancet, 2026. **407**(10527): p. 505-514.https://doi.org/10.1016/S0140-6736(25)02464-X
11. Javaeed, A. and A. Schuh, *Artificial intelligence in breast cancer diagnosis: A systematic literature review.* Cambridge Prisms: Precision Medicine, 2025. **3**: p. e7.https://doi.org/10.1017/pcm.2025.10006
12. Hammer, R.D., D. Fowler, L.R. Sheets, A. Siadimas, C. Guo, and M.S. Prime, *Digital Tumor Board Solutions Have Significant Impact on Case Preparation.* JCO Clinical Cancer Informatics, 2020(4): p. 757-768.https://doi.org/10.1200/CCI.20.00029
13. Kočo, L., C.C. Siebers, M. Schlooz, C. Meeuwis, H.S. Oldenburg, M. Prokop, and R.M. Mann, *The Facilitators and Barriers of the Implementation of a Clinical Decision Support System for Breast Cancer Multidisciplinary Team Meetings—An Interview Study.* Cancers, 2024. **16**(2): p. 401.https://doi.org/10.3390/cancers16020401
14. *10 Year Health Plan for England: fit for the future - GOV.UK*. Available from: https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future.
15. *Software and artificial intelligence (AI) as a medical device - GOV.UK*. Available from: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device.
16. *The Medical Devices Regulations 2002*. Available from: https://www.legislation.gov.uk/uksi/2002/618/contents.

17. Ng, J.J.W., E. Wang, X. Zhou, K.X. Zhou, C.X.L. Goh, G.Z.N. Sim, H.K. Tan, S.S.N. Goh, and Q.X. Ng, *Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review.* BMC medical informatics and decision making, 2025. **25**(1): p. 236.https://doi.org/10.1186/s12911-025-03061-0
18. Van Buchem, M.M., H. Boosman, M.P. Bauer, I.M. Kant, S.A. Cammel, and E.W. Steyerberg, *The digital scribe in clinical practice: a scoping review and research agenda.* NPJ digital medicine, 2021. **4**(1): p. 57.https://doi.org/10.1038/s41746-021-00432-5
19. Tran, B.D., R. Mangu, M. Tai-Seale, J.E. Lafata, and K. Zheng. *Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations*. in *AMIA Annual Symposium Proceedings*. 2023.
20. O'Kane, R., D. Stonehouse-Smith, L. Ota, R. Patel, N. Johnson, C. Slipper, J. Seehra, S. Papageorgiou, and M. Cobourne, *Transcription Accuracy of Automatic Speech Recognition for Orthodontic Clinical Records.* Journal of Dental Research, 2025: p. 00220345251382452.https://doi.org/10.1177/00220345251382452
21. Miner, A.S., A. Haque, J.A. Fries, S.L. Fleming, D.E. Wilfley, G. Terence Wilson, A. Milstein, D. Jurafsky, B.A. Arnow, and W. Stewart Agras, *Assessing the accuracy of automatic speech recognition for psychotherapy.* NPJ digital medicine, 2020. **3**(1): p. 82.https://doi.org/10.1038/s41746-020-0285-8
22. Shool, S., S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, and M. Tara, *A systematic review of large language model (LLM) evaluations in clinical medicine.* BMC Medical Informatics and Decision Making, 2025. **25**(1): p. 117.https://doi.org/10.1186/s12911-025-02954-4
23. Sorin, V., E. Klang, M. Sklair-Levy, I. Cohen, D.B. Zippel, N. Balint Lahat, E. Konen, and Y. Barash, *Large language model (ChatGPT) as a support tool for breast tumor board.* NPJ Breast Cancer, 2023. **9**(1): p. 44.https://doi.org/10.1038/s41523-023-00557-8
24. Griewing, S., J. Knitza, J. Boekhoff, C. Hillen, F. Lechner, U. Wagner, M. Wallwiener, and S. Kuhn, *Evolution of publicly available large language models for complex decision-making in breast cancer care.* Arch Gynecol Obstet, 2024. **310**(1): p. 537-550.https://doi.org/10.1007/s00404-024-07565-4
25. Hager, P., F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, and D. Rueckert, *Evaluation and mitigation of the limitations of large language models in clinical decision-making.* Nature Medicine, 2024. **30**(9): p. 2613-2622.https://doi.org/10.1038/s41591-024-03097-1
26. Buyukceran, E.U., A. Seyfettin, A. Babaturk, M.B. Ozkan, D. Colak, I. Unal, E. Kaymaz, E. Ergun, M.O. Emer, and H.H. Mersin, *High Concordance Between GPT-4o and Multidisciplinary Tumor Board Decisions in Breast Cancer: A Retrospective Decision Support Analysis.* J Med Syst, 2025. **49**(1): p. 179.https://doi.org/10.1007/s10916-025-02314-9
27. Zakka, C., R. Shad, A. Chaurasia, A. Dalal, J. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, and E. Ashley, *Almanac—retrieval-augmented language models for clinical medicine. NEJM AI. 2024.* AIoa2300068. **10**.https://doi.org/10.1056/AIoa2300068
28. Kresevic, S., M. Giuffrè, M. Ajcevic, A. Accardo, L.S. Crocè, and D.L. Shung, *Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework.* NPJ digital medicine, 2024. **7**(1): p. 102.https://doi.org/10.1038/s41746-024-01091-y

29. Unlu, O., J. Shin, C.J. Mailly, M.F. Oates, M.R. Tucci, M. Varugheese, K. Wagholikar, F. Wang, B.M. Scirica, and A.J. Blood, *Retrieval augmented generation enabled generative pre-trained transformer 4 (GPT-4) performance for clinical trial screening.* medRxiv, 2024.https://doi.org/10.1101/2024.02.08.24302376
30. Abdullayev, N., J. Kottlors, H. Habibov, F. Yilmaz, C. Zimmer, N.G. Hokamp, S. Lennartz, V. Valiyev, C. Abbasli, and L. Goertz, *European guideline informed RAG-based GPT-4 decision support tool in tumor board meetings for breast cancer treatment.* European Journal of Surgical Oncology, 2025: p. 110384.https://doi.org/https://doi.org/10.1016/j.ejso.2025.110384
31. Dehdab, R., S. Afat, F. Mankertz, J.M. Brendel, N. Maalouf, S. Werner, A. Brendlin, J. Herrmann, K. Nikolaou, L.D. Kloker, B. Calukovic, K. Benzler, L. Zender, and C.K.W. Deinzer, *When AI joins the table: evaluating large language model performance in soft tissue sarcoma tumor board decisions.* Journal of Cancer Research and Clinical Oncology, 2026. **152**(2): p. 52.https://doi.org/10.1007/s00432-026-06432-w
32. Ahmed, M.I., B. Spooner, J. Isherwood, M. Lane, E. Orrock, and A. Dennison, *A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare.* Cureus, 2023. **15**(10): p. e46454.https://doi.org/10.7759/cureus.46454
33. *Jetson AGX Orin for Next-Gen Robotics | NVIDIA*. Available from: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/.
34. Radford, A., J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, in *Proceedings of the 40th International Conference on Machine Learning*. 2023, JMLR.org: Honolulu, Hawaii, USA. p. Article 1182.
35. Nussbaum, Z., J.X. Morris, B. Duderstadt, and A. Mulyar, *Nomic embed: Training a reproducible long context text embedder.* arXiv preprint arXiv:2402.01613, 2024.https://doi.org/10.48550/arXiv.2402.01613
36. Johnson, J., M. Douze, and H. Jégou, *Billion-scale similarity search with GPUs.* IEEE transactions on big data, 2019. **7**(3): p. 535-547.https://doi.org/10.1109/TBDATA.2019.2921572
37. *Amazon Polly*. Available from: https://aws.amazon.com/pm/polly/?trk=5735757a-030b-49bd-96de-519b2a8af735&sc_channel=ps&ef_id=Cj0KCQjws83OBhD4ARIsACblj1-VDFvW9JdDcmqbU5RhEDLEcUwfxYHDDeudvegjmTIdTuNZynlS4JsaAotnEALw_wcB:G:s&s_kwcid=AL!4422!3!795841353823!p!!g!!amazon%20free%20text%20to%20speech!23533257079!191998606999&gad_campaignid=23533257079&gbraid=0AAAADjHtp9g_-6bGryZi1ISluqGvVKxo&gclid=Cj0KCQjws83OBhD4ARIsACblj1-VDFvW9JdDcmqbU5RhEDLEcUwfxYHDDeudvegjmTIdTuNZynlS4JsaAotnEALw_wcB.
38. Galvez, D., V. Bataev, H. Xu, and T. Kaldewey, *Speed of Light Exact Greedy Decoding for RNN-T Speech Recognition Models on GPU.*https://doi.org/10.1007/s00432-026-06432-w
39. Bain, M., J. Huh, T. Han, and A. Zisserman, *WhisperX: Time-Accurate Speech Transcription of Long-Form Audio.*https://doi.org/10.48550/arXiv.2303.00747
40. *Amazon Transcribe Medical*. Available from: https://aws.amazon.com/transcribe/medical/.
41. Barański, M., J. Jasiński, J. Bartolewska, S. Kacprzak, M. Witkowski, and K. Kowalczyk. *Investigation of whisper asr hallucinations induced by non-speech audio*. in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025. IEEE.

42. Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. *Optuna: A next-generation hyperparameter optimization framework*. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019.
43. Reddy, C.K., V. Gopal, and R. Cutler. *DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors*. in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. IEEE.
44. *MedGemma: Our most capable open models for health AI development*. Available from: https://research.google/blog/medgemma-our-most-capable-open-models-for-health-ai-development/.
45. *Introducing Palmyra Med and Palmyra Fin - WRITER*. Available from: https://writer.com/blog/palmyra-med-fin-models/.
46. *Advanced breast cancer: diagnosis and treatment Clinical guideline*. 2009; Available from: www.nice.org.uk/guidance/cg81.
47. *Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer Clinical guideline*. 2013; Available from: www.nice.org.uk/guidance/cg164.
48. *Early and locally advanced breast cancer: diagnosis and management NICE guideline*. 2018; Available from: www.nice.org.uk/guidance/ng101.
49. *Camelot: PDF Table Extraction for Humans — Camelot 1.0.9 documentation*. Available from: https://camelot-py.readthedocs.io/en/master/.
50. Mc, N.Q., *Note on the sampling error of the difference between correlated proportions or percentages.* Psychometrika, 1947. **12**(2): p. 153-7.10.1007/BF02295996
51. Elliott, H., A.J. Allen, N.D. Forester, S. Graziadio, W. Jones, B.C. Lendrem, M.S. Pearce, T. Powell, J. Scott, and A. Bray, *Women's perspectives of molecular breast imaging: a qualitative study.* British journal of cancer, 2025. **132**(3): p. 276-282.https://doi.org/10.1038/s41416-024-02930-1
52. Jones, W.S., J. Suklan, A. Winter, K. Green, T. Craven, A. Bruce, J. Mair, K. Dhaliwal, T. Walsh, and A. Simpson, *Diagnosing ventilator-associated pneumonia (VAP) in UK NHS ICUs: the perceived value and role of a novel optical technology.* Diagnostic and Prognostic Research, 2022. **6**(1): p. 5.https://doi.org/10.1186/s41512-022-00117-x
53. Shi, M., Z. Jin, Y. Xu, Y. Xu, S.-X. Zhang, K. Wei, Y. Shao, C. Zhang, and D. Yu. *Advancing multi-talker ASR performance with large language models*. in *2024 IEEE Spoken Language Technology Workshop (SLT)*. 2024. IEEE.
54. *Somerset cancer register - Somerset Cancer Register*. Available from: https://www.somersetft.nhs.uk/somerset-cancer-register/.
55. Atiku, S., K. Owolanke, and O. Olakotan, *Assessing the Reliability, Accuracy, and Relevance of Artificial Intelligence Speech Recognition for Clinical Documentation: A Scoping Review.* Journal of Evaluation in Clinical Practice, 2026. **32**(4): p. e70460.https://doi.org/10.1111/jep.70460
56. Ellis, Z., J. Joselowitz, Y. Deo, Y.V. He, A. Kalygina, A. Higham, M. Rahimzadeh, Y. Jia, I. Habli, and E. Lim. *Wer is unaware: Assessing how asr errors distort clinical understanding in patient facing dialogue*. in *Proceedings of the 16th International Workshop on Spoken Dialogue System Technology*. 2026.

57. Griewing, S., N. Gremke, U. Wagner, M. Lingenfelder, S. Kuhn, and J. Boekhoff, *Challenging ChatGPT 3.5 in Senology-An Assessment of Concordance with Breast Cancer Tumor Board Decision Making.* J Pers Med, 2023. **13**(10).https://doi.org/10.3390/jpm13101502
58. Xu, W., X. Wang, L. Yang, M. Meng, C. Sun, W. Li, J. Li, L. Zheng, T. Tang, W. Jia, and X. Chen, *Consistency of CSCO AI with Multidisciplinary Clinical Decision-Making Teams in Breast Cancer: A Retrospective Study.* Breast Cancer (Dove Med Press), 2024. **16**: p. 413-422.https://doi.org/10.2147/BCTT.S419433
59. Liao, N., C. Li, W.J. Gradishar, V.S. Klimberg, J.A. Roshal, T. Yuan, S.S. Agarwala, V.K. Valero, S.M. Swain, J.A. Margenthaler, I.T. Rubio, S.A. Hurvitz, C.E. Geyer, Jr., N.U. Lin, H.S. Rugo, G. Zhang, N. Liu, and C.M. Balch, *Accuracy and Reproducibility of ChatGPT Responses to Breast Cancer Tumor Board Patients.* JCO Clin Cancer Inform, 2025. **9**: p. e2500001.https://doi.org/10.1200/CCI-25-00001
60. Dogan, I., M.K. Bartin, E. Sonmez, E. Seyran, H.A. Bozkurt, M. Yuksek, E.D. Serbes, G. Zalova, and S. Celik, *Chat GPT Performance in Multi-Disciplinary Boards-Should AI Be a Member of Cancer Boards?* Healthcare (Basel), 2025. **13**(18).https://doi.org/10.3390/healthcare13182254
61. Ah-Thiane, L., P.E. Heudel, M. Campone, M. Robert, V. Brillaud-Meflah, C. Rousseau, M. Le Blanc-Onfroy, F. Tomaszewski, S. Supiot, T. Perennec, A. Mervoyer, and J.S. Frenel, *Large Language Models as Decision-Making Tools in Oncology: Comparing Artificial Intelligence Suggestions and Expert Recommendations.* JCO Clin Cancer Inform, 2025. **9**: p. e2400230.https://doi.org/10.1200/CCI-24-00230
62. Umihanic, S., H. Osmanovic, N. Selak, D. Kopric, A. Huseinbasic, E. Sehic-Kozica, B. Babic, and F. Umihanic, *Evaluating the Concordance Between ChatGPT and Multidisciplinary Teams in Breast Cancer Treatment Planning: A Study from Bosnia and Herzegovina.* J Clin Med, 2025. **14**(18).https://doi.org/10.3390/jcm14186460
63. Schmutz, M., S. Sommer, J. Sander, D. Graumann, J. Raffler, I. Soto-Rey, S. Sheikhalishahi, L. Schmidt, L.P. Unkelbach, L. Ortak, T. Schaller, S. Dintner, K. Hildebrand, M. Kuhlen, F. Jordan, M. Trepel, C. Hinske, and R. Claus, *Large language model processing capabilities of ChatGPT 4.0 to generate molecular tumor board recommendations-a critical evaluation on real world data.* Oncologist, 2025. **30**(10).https://doi.org/10.1093/oncolo/oyaf293
64. Moser, E.C. and G. Narayan, *Improving breast cancer care coordination and symptom management by using AI driven predictive toolkits.* Breast, 2020. **50**: p. 25-29.https://doi.org/10.1016/j.breast.2019.12.006