life

MDPI

Review

The Role of Artificial Intelligence in Multidisciplinary Tumor Boards for Breast Cancer: A Critical Review

Vittorio Gebbia [1,2,*], Daniela Sambataro [1,3], Calogero Cipolla [4], Dario Piazza [5] and Maria Rosaria Valerio [6]

1 Medical Oncology, Kore University of Enna, 94100 Enna, Italy; daniela.sambataro@unikore.it
2 Medical Oncology Unit, Casa di Cura Torina, 90100 Palermo, Italy
3 Medical Oncology Unit, Ospedale Umberto I, 94100 Enna, Italy
4 Breast Unit, Policlinico, Dipartimento di Biomedicina, Neuroscienze e Diagnostica Avanzata (Bi.N.D.), University of Palermo, 90100 Palermo, Italy; calogero.cipolla@unipa.it
5 Data Managing Service, Casa di Cura Torina, 90100 Palermo, Italy; dariopiazza@gmail.com
6 Medical Oncology Unit and Breast Cancer Unit, Policlinico, University of Palermo, 90100 Palermo, Italy; mariarosaria.valerio@unipa.it
* Correspondence: vittorio.gebbia@unikore.it

Abstract

Multidisciplinary tumor boards (MTBs) are the gold standard for breast cancer (BC) treatment planning, but rising caseloads and the growing number of clinical, pathological, and molecular factors under consideration are straining their capacity. AI, particularly large language models (LLMs), has been proposed as a decision support tool, and BC is the tumor type for which AI-MTB concordance has been studied most. We conducted a narrative critical review of studies comparing AI-generated treatment recommendations with breast MTB recommendations, along with evidence on MTB performance, earlier rule-based decision support systems, and the relevant regulatory and medico-legal environment. We evaluated 17 LLM studies published between 2023 and 2026. Overall concordance ranged from 16% to 94%, increasing with model generation, structured input, and guideline grounding, and decreasing with disease stage and case complexity. Concordance was consistently high for endocrine therapy and standard indications for systemic therapy, variable for surgery, radiotherapy, and genomic testing, and low or inconsistent for genetic counseling, anti-HER2 regimen selection, and the management of metastatic disease. Reproducibility was poor, and no study examined patient outcomes. Concordance with the MTB is not equivalent to clinical correctness, and hallucinations, the handling of incomplete information, and automation bias are distinct safety dimensions that concordance does not capture; we therefore propose a multidimensional evaluation framework. The current, limited evidence base supports using AI as a supervised, guideline-grounded pre-MTB tool in treatment planning rather than as a replacement for the MTB itself.

Keywords: artificial intelligence; breast cancer; clinical decision support; concordance; large language models; multidisciplinary tumor board; precision oncology

Check for updates

Academic Editors: Alejandro Martin Sanchez, Flavia de Lauretis and Chiara Valeria Pirrottina

Received: 18 October 2026
Revised: 17 November 2026
Accepted: 21 November 2026
Published: 25 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Over the past decade, breast cancer (BC) management has exemplified precision oncology principles [1–5]. As the most prevalent malignancy among women, BC requires optimal diagnosis and treatment that depend on a growing array of clinical, pathological, and molecular variables [6,7]. These variables include primary tumor size and location, nodal involvement, presence or absence of metastases, menopausal status, histological grade, estrogen and progesterone receptor (ER/PR) status, human epidermal growth

factor receptor 2 (HER2) expression, proliferation index (Ki-67), genomic recurrence-risk signatures, and germline or somatic mutation status [1–7]. Each factor can influence the recommended sequence and combination of surgery, radiotherapy (RT), chemotherapy (CT), endocrine therapy (ET), immunotherapy, and HER2-targeted agents.

Multidisciplinary tumor boards (MTBs) synthesize this complexity into individualized management plans. The breast MTB typically comprises breast surgeons, medical oncologists, radiation oncologists, breast radiologists, pathologists, molecular biologists, and clinical geneticists [8]. Additional professionals, such as nurses, psycho-oncologists, nutritionists, and physiotherapists, may participate routinely or as needed. Core functions include confirming or refining staging based on imaging and pathology; sequencing decisions, such as neoadjuvant versus adjuvant systemic therapy and the timing of surgery and radiotherapy; selecting systemic therapy based on receptor status and genomic profile, including indications for genomic signature testing (e.g., Oncotype DX, MammaPrint), comorbidities, and concomitant medications [7,9]; deciding the type, intensity, and duration of anti-HER2 therapy; referring for germline genetic testing based on family history and tumor characteristics [10]; and considering clinical trial eligibility and geriatric assessment for older patients [11].

Concurrently, BC has become the most extensively studied tumor type in the emerging literature on artificial intelligence (AI), particularly regarding large language models (LLMs) applied to oncological decision-making [12–14]. Since 2023, multiple research groups have compared LLM-generated treatment recommendations with those of breast MTBs, resulting in a substantial body of concordance evidence for this tumor type [15,16]. A recent systematic review of LLMs in tumor boards reported that BC accounted for 11 of 31 included studies, with concordance rates ranging from 16% to 94% [16].

This rapid accumulation of evidence prompts consideration of whether AI could feasibly substitute for part or all the deliberative function of breast MTBs, and under what conditions such substitution might be appropriate. This critical review describes the evolution of AI-based decision support, from rule-based systems to LLMs and autonomous agents, and evaluates disease-specific evidence on AI–MTB concordance in BC, both overall and by treatment domain. The review also examines methodological, medico-legal, ethical, and relational factors that influence whether and in which contexts AI can safely augment or replace MTB deliberation.

Several systematic reviews have recently addressed this field, examining LLMs in BC management in general [15] and LLMs or AI in tumor boards across tumor types (Table 1) [16,17]. The present review differs from them in four respects. First, it is restricted to the decisions of the breast MTB and analyzes concordance by treatment domain, distinguishing agreement on treatment indications from agreement on detailed treatment strategies (Tables 2 and 3). Second, it separates concordance from other safety-relevant dimensions, namely clinical appropriateness, guideline traceability, hallucinations, and the handling of incomplete information, and proposes a multidimensional framework for their evaluation (Table 4). Third, it relates the evidence to the regulatory and medico-legal requirements in force in the EU, the United States, and Italy. Fourth, it translates the domain-level findings into a supervised, AI-augmented MTB workflow, which is presented as a framework for prospective testing rather than as an established clinical approach.

2. Materials and Methods

2.1. Design and Search Strategy

This is a structured, critical, narrative review. The search was conducted in PubMed/MEDLINE, Scopus, Web of Science, and the Scite.ai citation database for papers published between January 2015 and September 2026, using the following term combinations: “breast

cancer" or "breast neoplasm"; "multidisciplinary tumor board", "tumor board", "multidisciplinary team", or "MDT"; and "artificial intelligence", "large language model", "LLM", "ChatGPT", "GPT- 4", "Claude", "Gemini", "Llama", "clinical decision support", or "Watson for Oncology". Reference lists of the included studies and the most recent systematic reviews were hand-searched, and citation networks were explored in both directions using Scite.ai Smart Citations [15–17]. All bibliographic details and editorial notices (retractions, corrections) for the cited records were confirmed in the Scite.ai database. Each reference was additionally checked against the publisher's record of the article (DOI landing page or journal website).

2.2. Eligibility Criteria

Primary studies were eligible if they compared AI-generated treatment recommendations (from a rule-based clinical decision support system, an LLM, a small language model, or an AI agent) with recommendations from a breast MTB, a guideline-based expert panel, or treating oncologists, using real or simulated BC cases. We prioritized peer-reviewed work and included conference abstracts and preprints when they provided data not otherwise available, and we made this clear. We excluded studies focused solely on image interpretation, information extraction, or patient-facing question answering from the concordance synthesis but used them to inform the discussion. We obtained evidence on MTBs' effectiveness, AI-assisted imaging, automation bias, and the regulatory and medico-legal framework through targeted searches.

2.3. Data Extraction and Critical Appraisal

For each concordance study, we extracted country/setting, number and type of cases, AI model(s), reference standard, outcome metric, overall concordance, and domain-specific results. Given the diversity of study designs and metrics (percent agreement, Likert scale scores, Cohen's or Fleiss' kappa, F1 score), we did not conduct a meta-analysis. The studies were assessed qualitatively with respect to the domains most pertinent to the substitution question adapted from the QUADAS-2 framework used in a previous systematic review of LLMs in BC [15]. The framework included case selection (consecutive real-world cases versus curated or fictional cases), reference standard (MTB decision versus guideline or expert panel), rater blinding, reporting of prompts and model versions, assessment of reproducibility, and clinical outcomes. We also recorded, where reported, (i) the guideline source and version used as the reference or supplied to the model; (ii) whether reference decisions were independently adjudicated against guidelines or by an external panel; (iii) how incomplete or ambiguous case information was handled; and (iv) whether hallucinations, unsupported recommendations, or citation accuracy were assessed. Agreement statistics were interpreted using the Landis–Koch scale (κ 21–40 = fair; 41–60 = moderate; 61–80 = substantial; >80 = almost perfect).

2.4. Use of Generative Artificial Intelligence and Verification of Extracted Data

Generative artificial intelligence (Claude Opus 5.5; Anthropic, San Francisco, CA, USA) was used for three tasks only: (i) language editing of the manuscript; (ii) support for literature retrieval, by running the predefined search strings in the Scite.ai database and listing candidate records, which were then screened for eligibility by the authors; and (iii) pre-filling a predefined extraction template with study characteristics and numerical results (setting, number and type of cases, model, reference standard, overall and domain-specific concordance, and agreement statistics) from the full texts, abstracts, or preprints supplied by the authors. The model was not used for study design, eligibility decisions, critical appraisal, interpretation of the evidence, or formulation of the conclusions. Every

numerical value reported in the text, in Tables 1–3, and in Figure 1 was verified by two authors (V.G. and D.P.) against the primary record (full publication, conference abstract, or preprint); discrepancies were resolved by consensus, and the primary record was always taken as the source of truth. Bibliographic details and editorial notices were checked through the Scite.ai database and, for every reference, against the publisher's record. The authors take full responsibility for the content of this publication.

3. The Breast Cancer MTB: Functions, Value, and Limitations

Multidisciplinary care is mandatory for all certified breast units in Europe [8] and is supported by international guidelines [6,7]. Observational evidence suggests benefit. A comparative cohort study of 13,722 women in western Scotland found that after introducing multidisciplinary team working in one health board area, BC-specific mortality was significantly lower than in adjacent areas without multidisciplinary care [18]. A national matched cohort of 18,532 Taiwanese patients found that MTB care was associated with a reduced risk of recurrence (HR 0.84; 95% CI 0.70–0.99) [19]. The benefits of MTB care extend beyond survival to include improved adherence to guidelines, care coordination, documentation, recruitment into clinical trials, and team learning [20,21].

The benefits are real but come at a cost. MTBs are burdensome in terms of resources and finances, including human resources and time.. Validated quality indicators are not consistently applied in routine practice due to structural, organizational, and resource-related issues [21]. Case volumes are increasing, while the time available for each case is decreasing. Information on comorbidities and patient preferences is often omitted from discussions [21]. Many cases presented to the breast MTB follow a well-defined guideline algorithm, which consumes significant time that could otherwise be used on more complex cases. These pressures are the primary reason for introducing AI into the MTB workflow: to pre-populate and structure cases, check for completeness and adherence to guidelines, and triage the level of discussion required for each case [17,22].

4. From Rule-Based Decision Support to Large Language Models

4.1. Rule-Based Clinical Decision-Support Systems

The first generation of AI tools for MTB were knowledge-based clinical decision support systems. The best known of these was IBM Watson for Oncology, which generated a ranked list of therapeutic options from a curated database of the literature, guidelines, and expert training cases. In 638 cases discussed at an Indian tertiary center, the system's recommendations agreed with the MTB in 93% of cases, and concordance was influenced by stage and patient age [23]. In China, concordance varied dramatically by tumor type, which was ascribed to differences in disease epidemiology, drug availability, and local practice patterns [24]. These early experiences point to issues that remain in the LLM era: concordance depends on the local reference standard; a system trained on one institution's practice may not transfer to another; and agreement with the board is not necessarily the same as clinical correctness.

4.2. Task-Specific AI Feeding the MTB

A second class of AI tools operates upstream of the MTB by improving the quality of the data used to make decisions. In mammography screening, the Randomized MASAI trial found that AI-assisted screen reading increased cancer detection and reduced screen reading workload by 44% compared with standard double reading [25,26]. A decade of evidence, summarized in a meta-analysis, concluded that AI is best deployed as a second reader and triage tool rather than as a fully autonomous replacement for the radiologist [26]. Similar tools are in development for radiomics and digital pathology, such as biomarker

quantification and prediction of molecular alterations from whole slide images [22]. Recent examples include a network-based method that integrates gene expression with multiple protein-interaction networks to identify prognostic, subtype-specific breast cancer biomarkers [27], and a multicenter ultrasound model combining radiomics and deep learning features that predicted fluorescence in situ hybridization (FISH) results in HER2 2+ tumors in a cohort of 5888 patients, with areas under the curve (AUC) of 0.63–0.80 in four external validation cohorts [28]. Such tools could reduce unnecessary testing and enrich the information available to the board, but the variable performance across external cohorts shows that their outputs require local validation before they inform MTB decisions. These are clearly relevant to MTB, but they also highlight a potential problem: in an experimental study, incorrect AI suggestions were found to have a detrimental effect on radiologist performance across all levels of experience, a phenomenon known as automation bias [29].

4.3. Large Language Models and AI Agents

LLMs are general-purpose generative models trained on large text corpora and can therefore be used to summarize clinical notes, retrieve guideline content, and generate free-text treatment plans [13,14]. In the standard BC concordance study, the model receives a structured patient profile (age, menopausal status, tumor size and grade, receptor and HER2 status, nodal status, Ki-67, and comorbidities) and produces a treatment plan, which is then scored against the actual MTB recommendation for each treatment domain (surgery, RT, CT, ET, anti-HER2 therapy, genomic testing, and referral for genetic counseling) [30–32]. Three technical developments are relevant to the question of substitution. First, retrieval- or guideline-augmented systems add a layer of structured guideline content to a foundation model [33]. Second, small language models can be fine-tuned to a national guideline and run locally, thereby providing source control, explainability, and data security [34]. Third, autonomous AI agents comprise an LLM with a suite of specialist tools, such as histopathology classifiers, image segmentation, and literature and knowledge base search. In a validation study of 20 realistic multimodal oncology cases, the agent achieved correct clinical conclusions in 91.0% of cases and improved decision-making accuracy from 30.3% with GPT-4 alone to 87.2% [35].

5. Evidence from Breast Cancer Concordance Studies

5.1. Overview of the Evidence

Seventeen studies compared large language model (LLM) recommendations with MTB, expert panel, or clinician decisions in BC (Table 1) [30–34,36–47]. Most were retrospective and single- center, with sample sizes ranging from 10 to 844 cases. The majority evaluated successive ChatGPT versions (GPT-3.5, GPT-4, GPT-4 Turbo, and GPT-4o), using Llama, Bard/Gemini, and Claude as comparators. Figure 1 shows the overall concordance values reported as percentages.

Table 1. Studies comparing large language model (LLM) treatment recommendations with multidisciplinary tumor board (MTB), expert, or clinician decisions in breast cancer, in chronological order.

Study	Setting and Cases	AI Model(s)	Reference Standard	Main Findings
Sorin et al., 2023 [36]	Israel; 10 consecutive real cases	GPT-3.5	Breast MTB	Recommendations like MTB in 7/10 cases (70%); summarization, recommendation, and explanation rated 3.7–4.6/5

Table 1. *Cont.*

Study	Setting and Cases	AI Model(s)	Reference Standard	Main Findings
Lukac et al., 2023 [30]	Germany; 10 consecutive primary early BC	GPT-3.5	MTB	Consensus score 16.05%; mostly general answers; anti-HER2 antibody suggested for HER2 1+/2+ FISH-negative tumors
Griewing et al., 2023 [37]	Germany; 20 fictional profiles (incl. DCIS and metastatic)	GPT-3.5	MTB	Concordance 50% overall and 58.8% for invasive BC; CT 94.5%, RT 95.0%, ET 75.0%
Griewing et al., 2024 [31]	Germany; 20 complex fictional profiles	GPT-4, GPT-3.5 (two versions), Llama 2, Bard	MTB	Invasive BC: GPT-4 70.6%, GPT-3.5 58.8% and 41.2%, Llama 2 35.3%, Bard 23.5%; GPT-4 100% for RT; genetic testing lowest (55–85%)
Griewing et al., 2024 [34]	Germany; 20 fictional profiles, 100 binary decisions	BC-SLM (local, guideline-tailored), GPT-4, GPT-3.5	Guideline-based MTB	Concordance 86% ($\kappa = 0.72$), 90% ($\kappa = 0.82$), and 83% ($\kappa = 0.66$), respectively; BC-SLM provides referenced guideline sections
Stalp et al., 2024 [38]	Germany; 30 MTB cases	GPT-3.5	Oncologists (questionnaire)	Overall "sufficient with minor limitations"; HER2-directed therapy best rated; weaker in postoperative and complex cases and in treatment sequencing
Torres-Cisneros et al., 2024 (abstract) [32]	Mexico; 124 consecutive patients (stage 0–IV)	GPT-4	MTB	Overall, 80.65%; 93–100% for most domains; anti-HER2 agent 68%; stage IV management 50%
Ah-Thiane et al., 2025 [39]	France; 112 early BC	Claude 3 Opus, GPT-4 Turbo, Llama 3-70B	MTB	Appropriate suggestions 86.6%, 85.7%, and 75.0%; accurate for ET and targeted therapy; RT overestimated; variable for CT and genomic tests
Goh et al., 2025 [33]	Singapore; 50 postoperative non-metastatic	Claude 2 + NCCN (TheSerenityBot), Claude 2, GPT-4	MTB	Accuracy 0.89, 0.86, and 0.78; GPT-4 less accurate for genetic testing (OR 0.05); Claude 2 less accurate for RT (OR 0.41)
Liao et al., 2025 [40]	International webinar; 362 consecutive cases	GPT-4 (three runs)	Expert panel	Fully concordant 46%; lower in stage IV and for molecular/genetic testing and ADCs; fully reproducible 32%; inter-response similarity 39%
Umihanić et al., 2025 [41]	Bosnia and Herzegovina; 91 treatment-naïve	GPT-4	MDT; four raters, 4-point Likert	Mean score 3.31; Cronbach's α 0.86; Fleiss' κ 0.31 (fair); higher for HR-negative and standard CT
Büyükceran et al., 2025 [42]	33 MTB cases	GPT-4o (two sessions)	MTB; two blinded raters, 5-point Likert	Full concordance 93.9%; discordances limited to genetic counseling; identical outputs across sessions

Table 1. *Cont.*

Study	Setting and Cases	AI Model(s)	Reference Standard	Main Findings
Nazlı et al., 2026 [43]	286 recommendations of a radiology-led MTB	GPT-4o, Claude 3.7 Sonnet, Gemini 2.5 Pro	MTB	Overall 83.2%, 79.7%, and 79.4%; >90% in HER2-enriched and TNBC, 66% in luminal A; F1 <58 for mastectomy and ≤23.5 for ALND
Tokoçin et al., 2026 [44]	100 consecutive patients	GPT-4, GPT-3.5	MTB	Primary treatment weighted κ 0.712 and 0.600; genetic counseling κ 0.398 and 0.302; better in early-stage disease
Jiang et al., 2026 [45]	China; 50 radiology and guideline questions	Nine LLMs incl. GPT-4o, Claude 3 Opus, Claude 3.5 Sonnet	NCCN 2024, BI-RADS; radiologists	GPT-4o most accurate; Claude models most consistent; no significant difference vs. physician groups; cannot replicate MTB decision-making
Buonaiuto et al., 2026 [46]	Italy; 607 + 237 (PRO BONO) HR+/HER2− with Oncotype DX	GPT-4o	Treating clinicians	Agreement 68–70% before and 90–93% after the RS; poor prediction of high genomic risk
Karabuğa et al., 2026 [47]	411 HR+/HER2− without genomic assay	GPT-4o, o3	Medical oncologist	Clinician–model κ 0.60 and 0.55 (moderate); model–model κ 0.88

ADC, antibody–drug conjugate; ALND, axillary lymph node dissection; BC, breast cancer; BI-RADS, Breast Imaging Reporting and Data System; CT, chemotherapy; DCIS, ductal carcinoma in situ; ET, endocrine therapy; HER2, human epidermal growth factor receptor 2; HR, hormone receptor; MDT, multidisciplinary team; NCCN, National Comprehensive Cancer Network; OR, odds ratio; RS, Recurrence Score; RT, radiotherapy; SLM, small language model; TNBC, triple-negative breast cancer.

5.2. Overall Concordance

Concordance was not defined uniformly across studies, and two levels of agreement should be distinguished when interpreting the figures below. Agreement on treatment indications concerns whether a given modality (surgery, RT, CT, ET, anti-HER2 therapy, or genomic or genetic testing) is recommended; agreement on the detailed treatment strategy additionally requires a match in the specific agent or regimen, sequence, duration, and extent of surgery or RT. Most studies scored agreement at the level of indications, either as binary decisions per domain [34] or as domain-level indications [32,37]. Others required the full treatment plan to match the reference, including specific management details [40], rated the recommendations on Likert scales [36,41,42], or scored individual recommendations with F1 statistics [43]. Because indication-level agreement is easier to achieve, the overall rates reported below are not directly comparable: a high rate in a study scoring indications does not imply agreement on the specific regimen, and the lower rates in studies requiring full-plan agreement partly reflect a stricter definition. Where reported, we indicate the level at which agreement was assessed.

Overall agreement between LLM-generated treatment plans and BC MTB decisions varies considerably. Early studies using GPT-3.5 reported concordance scores ranging from 16% in 10 consecutive early BC cases [30] to 70% [36], and 58% for invasive cancer profiles in a German series of fictional cases [37]. For GPT-4, concordance reached 70% on a curated set of complex case profiles [31], whereas a real-world series of 124 consecutive patients with stage 0–IV disease, available only as a conference abstract, reported an overall concordance of 80.65% (corresponding to 100 of 124 cases) [32]. Performance on real-world cases was therefore not inferior to, and numerically higher than, performance on curated complex profiles, although the two series differ in case mix, reference standard, and level

of reporting and cannot be compared formally. More recent studies with GPT-4 class and Claude models on real cases reported concordance rates of 79–94% [39,42,43]. However, a large international series of 362 tumor board cases found only 46% full concordance with expert recommendations [40]. Three primary factors explain most of the observed variation. First, model generation: in direct head-to-head comparisons on identical cases, GPT-4 class models consistently outperformed GPT-3.5, Llama 2, and Bard [31]. Second, case complexity: concordance is higher in early-stage, guideline-typical cases than in advanced, atypical, or heavily comorbid cases [32,40,41,44]. Third, input structure and prompt design: studies using structured, guideline-referenced prompts reported higher agreement than those using open-ended prompts, which resemble how a resident might query a chatbot [30,38,42]. Differences in evaluation metrics also complicate direct comparison, as percent agreement, Likert scores, kappa statistics, and F1 scores represent distinct measures (see Section 6).

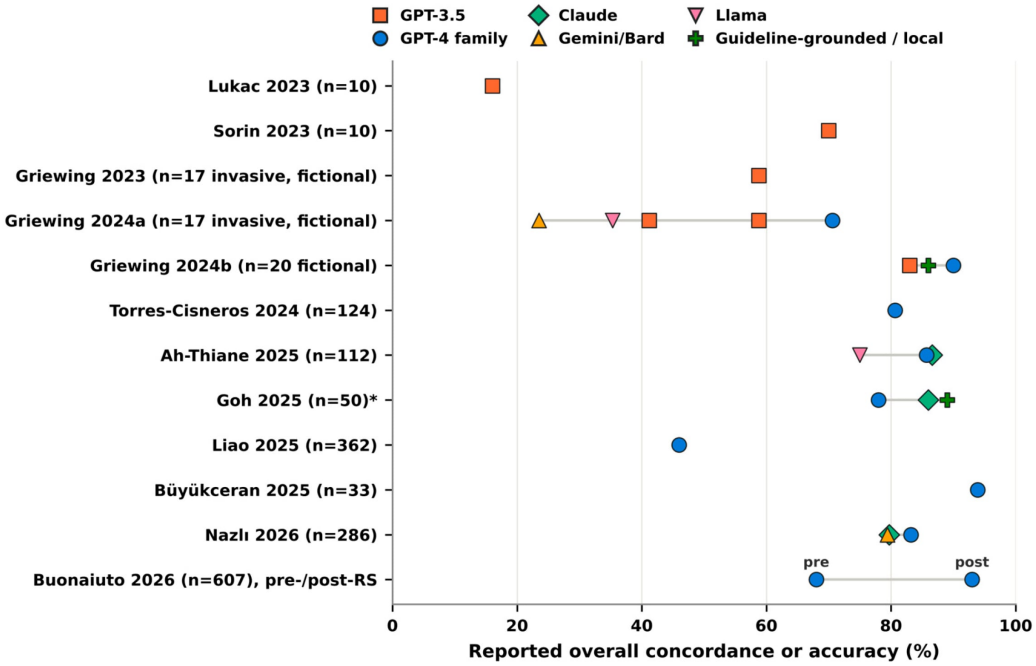


Figure 1. Overall concordance (or accuracy) between large language model recommendations and MTB, expert, or clinician decisions in breast cancer studies reporting a percentage-based metric [30–34,36,37,39,40,42,43,46]. Each marker represents one model; horizontal lines connect models evaluated on the same cases. Griewing 2024a and 2024b refer to references [31] and [34], respectively. Values are not directly comparable because of differences in case mix, reference standard, and metric (consensus score in Lukac et al.; accuracy in Goh et al. [*]; agreement before and after the Oncotype DX Recurrence Score in Buonaiuto et al., cohort 1). Studies reporting only Likert scores or κ statistics are not shown.

5.3. Domain-Specific Concordance

Concordance is typically reported for both the overall treatment strategy and individual treatment domains. Most domain-level figures refer to whether a modality is indicated rather than to the choice of a specific regimen: in the 124-patient series, for example, most domains, which largely concerned indications, reached 93–100%, whereas the choice of anti-HER2 agent, a strategy-level decision, reached 68% [32]. This granularity reveals a broadly consistent pattern: agreement is high for standardized, guideline-driven decisions and lower for complex, individualized, or advanced-stage decisions. In the 124-patient GPT-4 series, which is available only as a conference abstract without domain-level denominators or confidence intervals, concordance was very high for ET (100%), RT (98%), CT (97%),

and surgery (96%), as well as for additional staging or geriatric assessment (99%) and for the indication for genomic signature testing (96%). It was lowest for choosing an anti-HER2 agent (68%) and for overall management of stage IV disease (50%) (Table 2) [32]. These high figures must be interpreted in the context of case complexity. In HR-positive disease, the indication for ET is largely binary and guideline-defined, as are standard indications for CT; the abstract did not report concordance for the more demanding components of these decisions (choice of endocrine agent, ovarian function suppression, treatment duration, or the addition of adjuvant CDK4/6 inhibitors), nor did it stratify results by comorbidity, contraindications, previous treatment exposure, or competing treatment priorities. Across studies, concordance fell as complexity increased: it was lower in stage IV and advanced disease [40,44], in luminal A tumors (66%) [43], in postoperative and complex cases and in treatment sequencing [38], and in cases requiring individualized decisions in a prospective pan-cancer series [48]. High concordance for ET and standard systemic therapy should therefore be read as performance on straightforward, guideline-typical decisions; it cannot be extrapolated to patients with comorbidities, contraindications, previous treatment exposure, or competing treatment priorities, for whom no study has reported domain-specific results. However, these domain-level figures vary across studies, and three domains warrant specific comment.

Surgery and radiotherapy. The high concordance for surgery and RT reported in the 124-patient series [32] and the 100% RT concordance of GPT-4 in complex fictional cases [31] contrast with other findings. In a study of 286 MTB recommendations, F1 scores for surgical decisions fell to <58 for mastectomy and ≤23.5 for axillary lymph node dissection [43]. In a French series of 112 early BC cases, the models tended to overestimate the need for adjuvant RT. They showed limitations in validating surgery [39], and in a Singaporean series, Claude 2 was significantly less accurate for RT recommendations (odds ratio [OR] 0.41) [33]. Lower agreement was also reported for uncertain surgical indications [41]. The extent of surgery and the omission or de-escalation of RT, which depend on imaging nuances, margin status, and patient preference, therefore, cannot be considered a solved domain.

Table 2. Domain-specific concordance between GPT-4 recommendations and breast MTB decisions in a real-world series of 124 consecutive patients [32]. Data are from a single conference abstract; domain-level denominators and confidence intervals were not reported, and the figures should be regarded as hypothesis-generating.

Treatment/Decision Domain	Concordance Rate
Endocrine therapy indication	100%
Additional staging (PET/CT) or geriatric assessment	99%
Radiotherapy indication	98%
Chemotherapy indication	97%
Genomic recurrence-risk signature indication	96%
Surgical approach	96%
Genetic counseling/testing referral	93%
Anti-HER2 therapy (type of agent)	68%
Overall management, stage IV (metastatic) subgroup	50%

Overall concordance in this series was 80.65% (100/124 cases). Data from a conference abstract that did not report domain-level denominators or confidence intervals; percentages are shown as published and are not graded, because no validated thresholds exist for classifying levels of concordance; domain-level results from other studies are summarized in Table 3. HER2, human epidermal growth factor receptor 2; PET/CT, positron emission tomography/computed tomography.

Anti-HER2 therapy. An early pilot study by Lukac et al. found that ChatGPT recommended antibody therapy for HER2 1+ and HER2 2+/FISH-negative tumors, where it is not indicated in early BC, and conflated ET with generic "hormonal treatment" [30]. This error is specific to the early-stage setting and to HER2-directed monoclonal antibodies such as trastuzumab and pertuzumab. In metastatic disease, HER2-low status (immunohistochemistry [IHC] 1+, or IHC 2+ with negative in situ hybridization [ISH]) has become a treatment-defining category since the antibody–drug conjugate trastuzumab deruxtecan improved progression-free and overall survival compared with chemotherapy of physician's choice [49]. Evaluations of AI recommendations in HER2-low disease must therefore specify the disease setting and distinguish monoclonal antibodies from antibody–drug conjugates, a distinction that is all the more relevant given the lower concordance observed in cases involving antibody–drug conjugates [40]. By contrast, using more complex, deliberately constructed prompts, Stalp et al. found HER2-directed treatment to be the best-rated category [38], and Ah-Thiane et al. reported high accuracy for targeted therapy indications [39]. A plausible reconciliation is that indications for anti-HER2 therapy in clearly HER2-positive disease are well codified. In contrast, the selection of agent, regimen, and duration, including escalation and de-escalation based on residual disease, is not well codified [32]. AI performance in this domain thus appears highly sensitive to how clinical information is presented, an important caveat for clinical deployment, because real-world documentation is rarely as structured as a research prompt.

Genetic counseling and testing. This domain was initially described as relatively favorable, with 93% concordance in the 124-patient series [32] and ChatGPT's ability to identify hereditary risk factors in pilot work [30]. Across studies, however, it emerges as one of the weakest. Genetic testing was the domain with the lowest alignment in complex fictional cases (55–85% depending on the model) [31]; GPT-4 was significantly less accurate than the guideline-augmented model for genetic testing recommendations (OR 0.05; 95% CI 0.015–0.149) [33]; agreement on genetic counseling was only fair for both GPT-4 ($\kappa = 0.398$) and GPT-3.5 ($\kappa = 0.302$) [44]; the only discordances in an otherwise highly concordant GPT-4o series concerned genetic counseling [42]; and responses involving molecular markers and genetic testing were significantly less accurate in the largest series [40]. This domain depends heavily on family history and rapidly evolving testing criteria [10], information that structured data often captures incompletely.

5.4. Comparisons Between Model Families

Four of the BC studies reviewed here included Claude models (Anthropic). Ah-Thiane et al. compared Claude 3 Opus, GPT-4 Turbo, and Llama 3-70B on 112 early BC cases; suggestions were appropriate in 86.6%, 85.7%, and 75.0% of cases, respectively. There was no difference between Claude 3 Opus and GPT-4 Turbo ($p = 0.85$), and both outperformed Llama 3 [39]. Jiang et al. tested nine models, including Claude 3 Opus and Claude 3.5 Sonnet, on 50 radiology and guideline questions relevant to the breast MTB; GPT-4o was the most accurate, both Claude models scored highest for response consistency, and Claude 3 Opus had the highest confidence score; the authors concluded that the models cannot replicate complex MTB decision-making [45]. Nazlı et al. compared 286 MTB recommendations with those of GPT-4o, Claude 3.7 Sonnet, and Gemini 2.5 Pro, prompted with reference to ASCO, ESMO, and NCCN guidelines; overall concordance was 83.2%, 79.7%, and 79.4%, respectively. Agreement exceeded 90% in HER2-enriched and triple-negative BC. Still, it fell to 66% in luminal A tumors, and stage-based analyses were heterogeneous, with lower agreement where complex multimodal or individualized decisions were required [43]. Finally, Goh et al. used Claude 2 both as a stand-alone model and as the base of a guideline-augmented system (Section 5.5) [33]. Taken together, these data indicate that differences

among contemporary frontier models of similar generation are small relative to those driven by case complexity, input quality, and guideline grounding.

5.5. Guideline Augmentation, Small Language Models, and Agents

Goh et al. looked at the post-op adjuvant MTB recommendations for 50 non-metastatic BC cases and compared them against the output of three systems: TheSerenityBot (a Claude 2 model augmented with the 2023 NCCN guidelines), Claude 2, and GPT-4. The augmented system had the highest overall accuracy (0.89), followed by Claude 2 (0.86) and GPT-4 (0.78) [33]. The benefit of augmentation over the base model (without augmentation) was small, and the study wasn't designed to test that formally; the bigger difference was between the model families. The data thus lend support (though don't prove) to the idea that some of the performance gap is due to retrieval and grounding rather than the inherent reasoning capacity of the model. There is even stronger support for this from work outside of BC where the accuracy of GPT-4 on multimodal oncology cases went from 30.3% to 87.2% when it was augmented with retrieval and precision oncology tools [35].

In a preclinical proof of concept study, Griewing et al. developed BC-SLM, a locally hosted, source-controlled chatbot based on a small language model trained to the German BC guidelines. They compared it to GPT-3.5 and GPT-4 on 20 fictional patient cases using a guideline-based MTB as the reference standard [34]. Overall agreement was 86% for BC-SLM ($\kappa = 0.721$), 90% for GPT-4 ($\kappa = 0.820$), and 83% for GPT-3.5 ($\kappa = 0.661$), with modality-specific agreement ranging from 65–100%, 85–100%, and 55–95%, respectively. It is not more accurate but does go some way to overcoming the three major impediments to the clinical use of commercial LLMs: lack of source control, lack of explainability, and the need to transfer patient data to third-party servers.

5.6. Hormone Receptor-Positive/HER2-Negative Early Disease and Genomic Risk

Two recent studies examined the adjuvant CT decision in the HR+/HER2-negative early BC setting, which is by far the most common scenario discussed by breast MTBs. Buonaiuto et al. compared GPT-4o with clinicians in 607 patients with Oncotype DX testing in routine practice and in 237 patients from the prospective multicenter PRO BONO study [46]. Pre-test agreement was 68% and 70% (κ 0.38 and 0.40, respectively, i.e., fair), with clinicians recommending CT more often than the model. Post-test agreement was 93% and 90% (κ 0.81 and 0.74, respectively). The model performed well at identifying low genomic risk but was very poor at identifying high genomic risk (43–45% in postmenopausal patients) [46]. In 411 HR+/HER2-negative patients without genomic testing, agreement between a medical oncologist and either GPT-4o or o3 was moderate (κ 0.60 and 0.55, respectively), whereas agreement between the two models was almost perfect (κ 0.88) [47]. This shows that LLMs can reproduce the guidelines' logic well once the decisive biomarker is available; they cannot replace genomic testing, and models from the same developer are clearly prone to shared biases.

5.7. Reproducibility and Stability

Reproducibility is essential for any decision support tool, yet it has rarely been assessed. Liao et al. submitted 362 consecutive tumor board cases to GPT-4 three times, each time with a different user entering the case. For 68% of patients, the responses differed in content and format; for 32%, they were the same, and overall inter-response similarity was 39% [40]. Only 46% of responses matched the experts exactly, and concordance was significantly lower in stage IV disease ($p = 0.019$) and in cases involving antibody-drug conjugates; responses were often not wrong but simply omitted specific management details [40]. In contrast, in a smaller series of 33 cases with standardized prompts, the GPT-4o model produced the same recommendations in two separate sessions [42]. Outside of BC, stability

varied considerably across contemporary models when queried repeatedly on the same MTB cases [50], and in a multi-institutional evaluation, clinicians found the LLM-generated tumor board case summaries to be of good quality, but the recommendations to be of poor quality [51]. Stability is therefore model-dependent, user interface-dependent, and input standardization-dependent, and it should be reported routinely.

6. Critical Appraisal of the Evidence

The BC concordance literature has significant methodological limitations that must be considered before drawing conclusions about substitution.

The reference standard is flawed. All studies use the MTB decision (or the decision of an expert panel or a single oncologist) as the reference standard. MTB decisions vary by center and over time, depend on local resources and preferences, and are not always implemented. Thus, concordance measures agreement with one board rather than correctness. A discordant AI recommendation may, in some cases, be the better one, and a concordant one may simply reproduce a suboptimal local practice [21,24]. Importantly, in most studies the reference decisions were not themselves independently assessed against contemporaneous guidelines, the available evidence, and patient-specific factors: only a minority used a guideline-based reference [34] or an external expert panel [40], and none, to our knowledge, systematically adjudicated discordant cases to establish whether the AI or the board recommendation was more appropriate. Expert consensus is itself uncertain: boards differ in composition, resources, and practice, and many decisions in BC fall within areas where guidelines allow more than one acceptable option [9,21]. A more informative design would have an independent, blinded panel classify each discordant case as AI preferable, MTB preferable, both acceptable, or both unacceptable, and would report concordance with the MTB separately from clinical appropriateness. There is no outcome-based validation: no study has examined whether AI-concordant or AI-assisted decisions differ in survival, toxicity, or quality of life.

Case selection and input differ markedly from real-world practice. Several studies have used very small samples (10–50 cases) [30,33,36,42], fictional or curated profiles [31,34,37], or cases taken from public educational webinars that may have been part of the model training data [40]. Most studies provided the model with a clean, structured summary of the case created by the clinician (a process that represents a significant portion of the MTB's cognitive effort). Performance on raw, unstructured, incomplete EHR data is largely unknown.

Metrics, raters, and reporting are highly heterogeneous. Outcomes range from % agreement and consensus scores to Likert ratings, kappa statistics, F1 scores, and odds ratios. Raters were frequently not blinded to the source of the recommendation, and inter-rater agreement was inconsistent. In the Umihanić et al. study, internal consistency was high (Cronbach's alpha = 0.86), but inter-rater reliability was only fair (Fleiss' kappa = 0.31), even though the authors described this as moderate [41]. Several studies fail to report their prompts, model versions, and decoding parameters.

The evidence is highly time-sensitive. Model versions assessed in 2023–2024 are already out of date, and the content of the guidelines is in a constant state of flux. A positive result from one model version does not necessarily carry over to the next, and a negative result may significantly underestimate current capability. Conversely, part of the recent evidence is available only as a conference abstract [32], most series are small, and publication bias toward positive results is very likely.

Prospective evidence is very limited. We found no prospective BC-specific study in which AI recommendations were generated prior to and independently of the MTB discussion. In a prospective pan-cancer study of 100 patients, GPT-4 recommendations generated from

anonymized summaries agreed with the actual MTB decisions in 76.4% of cases, with most disagreements occurring in cases requiring an individualized decision [48].

Guideline sources and versions were rarely aligned. Recommendations were judged against different references: the 2023 NCCN guidelines supplied to an augmented system [33], ASCO, ESMO, and NCCN guidelines specified in the prompt [43], the 2024 NCCN guidelines and BI-RADS [45], and German national guidelines for a small language model [34], whereas several studies did not specify the guideline version against which the MTB or the model was assessed. Guidance changed substantially over the study period, for example with adjuvant CDK4/6 inhibition in high-risk HR+/HER2− disease, perioperative immunotherapy in early triple-negative BC, and trastuzumab deruxtecan for HER2-low metastatic disease [6,7,49], and the knowledge cutoff of a model may precede the guideline version used by the board. Drug availability and reimbursement also differ between the healthcare settings represented in this literature, which span Europe, Asia, and the Americas [24,32,41]. Part of the observed disagreement may therefore reflect mismatches in guideline version or healthcare setting rather than model failure. Future studies should report the guideline source and version in force at the time of each MTB decision, the knowledge cutoff of the model, and local treatment availability.

The handling of incomplete or ambiguous information was not evaluated. In nearly all studies, the model received a complete, clinician-curated case summary, so that missing data were, by design, rarely an issue. To our knowledge, none of the included studies assessed whether the model recognized missing or contradictory information (e.g., unknown HER2 or menopausal status, or an incomplete family history), requested clarification, or appropriately deferred a recommendation; the early observation that ChatGPT often gave general rather than case-specific answers [30] and the frequent omission of management details [40] indicate that such behavior cannot be assumed. Recognizing what is not known is a core safety property of clinical decision support and should be evaluated separately from concordance, for example by presenting cases with deliberately withheld key variables and scoring appropriate deferral or clarification requests. In this review, we did not reclassify cases with incomplete information and report each study's results as published.

Hallucinations and unsupported recommendations are distinct from disagreement. LLMs can generate fluent but unsupported content, including fabricated clinical details, non-existent or misattributed citations, and internally inconsistent plans [52]. These errors differ in nature from disagreement with the MTB: a concordant recommendation can rest on a fabricated rationale, and a discordant one can be well grounded in current evidence. In the BC literature, unsupported recommendations were documented only qualitatively, for example, anti-HER2 antibody therapy for HER2-low early BC and the conflation of ET with generic hormonal treatment [30], and no study systematically measured citation accuracy, guideline traceability, internal consistency, or the presence of fabricated clinical details. Only systems with explicit retrieval or source control linked their recommendations to identifiable guideline sections [33,34]. These dimensions should be reported as separate safety outcomes alongside concordance.

Concordance is not clinical benefit. The absence of patient-outcome studies is the most important gap in this literature. Evidence on AI-assisted MTB decision-making can be ordered hierarchically: technical performance (accuracy, reproducibility, and absence of hallucination); recommendation concordance; effects on the decision process (treatment selection, guideline adherence, time to decision and to treatment, and MTB workload); patient outcomes (toxicity, recurrence, survival, and quality of life); and equity of care across settings and patient groups. All the BC studies reviewed here remain at the first two levels. High concordance does not demonstrate that AI assistance changes treatment selection, shortens time to treatment, reduces toxicity, or improves recurrence, survival,

quality of life, or equity, nor can it exclude harm arising from the minority of discordant or erroneous recommendations. Table 4 summarizes a multidimensional framework that separates these dimensions and could guide future evaluations.

Table 3. Proposed multidimensional framework for evaluating AI-generated treatment recommendations for the breast multidisciplinary tumor board (MTB), and status of each dimension in the breast cancer (BC) literature.

Dimension	What It Captures	Suggested Assessment	Status in BC Studies
Recommendation concordance	Agreement with the MTB, overall and by decision domain	Percent agreement and κ with 95% CI, by domain and case complexity	Reported in all studies; heterogeneous metrics
Clinical appropriateness	Whether the recommendation is acceptable against contemporaneous guidelines and patient factors, independently of the MTB	Blinded adjudication of discordant (and a sample of concordant) cases by an independent panel	Not systematically performed [34,40]
Guideline traceability and currency	Whether each element maps to an identifiable guideline section and version in force at the time of the decision	Proportion of recommendations with a verifiable source; guideline version and model knowledge cutoff reported	Only guideline-augmented or source-controlled systems [33,34]
Hallucination and unsupported content	Fabricated clinical details, non-existent or misattributed citations, unsupported or internally inconsistent recommendations	Predefined error taxonomy with severity grading; rate per case	Qualitative reports only [30]
Handling of incomplete information	Recognition of missing or ambiguous data, requests for clarification, appropriate deferral	Test cases with deliberately withheld key variables; rate of appropriate deferral	Not assessed
Completeness	Omission of necessary components of the treatment plan	Domain checklist (surgery, RT, systemic therapy, genetics, trials)	Omissions noted in one large series [40]
Reproducibility	Stability across repeated runs, users, interfaces, and model versions	Repeated submissions; agreement between runs	Assessed in few studies [40,42]
Human–AI interaction	Automation bias, anchoring, and influence of AI output on the board decision	Prospective designs with sealed AI output; override and decision-change rates	Not assessed in BC
Clinical impact	Effects on care processes and patient outcomes	Time to treatment, guideline adherence, toxicity, recurrence, survival, quality of life, equity	No study

AI, artificial intelligence; CI, confidence interval; RT, radiotherapy; κ, kappa statistic.

7. Discussion

MTBs integrate imaging, pathology, genomic, and patient data to make complex treatment decisions. ML, deep learning, NLP, and LLM-based tools are increasingly used to support both case preparation and discussion. A recent review found that AI recommendations aligned with MTB recommendations in 70–90% of cases across the cancers studied and that performance was strong for task-specific applications, such as screening and staging [17]. Reported benefits included faster synthesis of clinical information, greater adherence to guidelines, and clearer documentation of treatment options, although human review was still required [16,17].

7.1. Implications for the Substitution Question

The BC evidence base is quite detailed compared with that for most other tumor types, and the answer to the substitution question is clearly domain-specific rather than binary (Table 3). When AI recommendations align with MTB consensus across independent studies, a structured, guideline-based AI tool could plausibly pre-populate a recommendation for the clinician to review and sign off. MTB deliberation continues to provide clear value in instances with less concordance across studies or where clinically significant errors have been reported, such as recommendations of anti-HER2 monoclonal antibody therapy for HER2-low early BC, where it is not indicated [30], overestimation of RT [39], poor agreement on the extent of surgery [43], and incorrect genetic testing recommendations [33,44].

There's another implication for the path of model development. Given that newer models will outperform older ones on the same BC case sets [31], and that guideline grounding and tool use will lead to improved accuracy [33–35], we would expect concordance in standardized domains to continue to improve. There will still be a significant gap in stage IV management and in the selection of anti-HER2 regimens [32,40]. These domains are inherently complex and in flux, involving rapidly evolving evidence, combination regimens, sequencing after progression, and individualized risk/benefit considerations. They may remain difficult for AI even as overall concordance improves, as they require the integration of very recent trial data, toxicity history, and patient preferences, where specialist judgment and up-to-date knowledge of the literature are paramount. LLM performance also suffers similarly in other complex or rare tumor settings, described as a "complexity gap" [16].

Table 4. Synthesis of artificial intelligence (AI)–multidisciplinary tumor board (MTB) concordance by decision domain across breast cancer studies, and suggested role of AI.

Decision Domain	Representative Findings	Consistency Across Studies	Suggested Role of AI
Endocrine therapy indication (HR-positive disease)	100% [32]; high accuracy [39]; 75% with GPT-3.5 [37]	High (newer models) in guideline-typical cases	Draft recommendation; clinician sign-off
Chemotherapy indication, standard scenarios	97% [32]; higher agreement for standard regimens and HR-negative tumors [41]; variable [39]	High to moderate; complex cases not reported	Draft recommendation; MTB if borderline
Adjuvant chemotherapy in HR+/HER2− disease without genomic result	κ 0.38–0.40 [46]; κ 0.55–0.60 [47]	Moderate	Genomic testing and MTB judgment required
Indication for genomic signature testing	96% [32]; variable [39]; poor prediction of high genomic risk [46]	Moderate, variable	Flag candidates; clinician decision
Additional staging or geriatric assessment	99% [32] (single study)	High (limited data)	Checklist and reminders

Table 4. *Cont.*

Decision Domain	Representative Findings	Consistency Across Studies	Suggested Role of AI
Radiotherapy	98% [32]; 100% with GPT-4 [31]; overestimation [39]; lower accuracy with Claude 2 [33]	Variable	Draft for standard indications; MTB for omission or de-escalation
Surgery (type and axillary management)	96% [32]; F1 <58 for mastectomy and ≤23.5 for ALND [43]; uncertain indications [41]	Low to variable	MTB deliberation
Genetic counseling and testing	93% [32]; lowest domain [31]; OR 0.05 [33]; κ 0.30–0.40 [44]; less accurate [40]; only source of discordance [42]	Low to variable	MTB with genetics input; AI as checklist
Anti-HER2 regimen selection (HER2-positive disease)	68% [32]; antibody therapy recommended for HER2-low early BC [30]; best-rated category with structured prompts [38]	Inconsistent	MTB deliberation
Metastatic (stage IV) disease	50% [32]; lower concordance in stage IV [40] and advanced disease [44]	Low	MTB deliberation

ALND, axillary lymph node dissection; HER2, human epidermal growth factor receptor 2; HR, hormone receptor; OR, odds ratio. Values attributed to reference [32] derive from a single conference abstract without domain-level denominators or confidence intervals and should be interpreted with caution.

7.2. Medico-Legal, Ethical, and Regulatory Dimensions

Even with sufficient accuracy, substitution would be constrained by regulatory, legal, and ethical requirements. Two levels should be distinguished. General AI governance principles, namely transparency, explainability, human oversight, accountability, data protection, equity, and continuous monitoring, as set out, for example, in the World Health Organization guidance on large multimodal models [53], apply across jurisdictions but do not by themselves define the obligations attached to a specific product. Jurisdiction-specific requirements for clinical decision-support software, by contrast, depend on the intended use declared by the manufacturer and on the resulting risk classification, and no single regulatory framework applies universally. In the EU, software designed to assist in making a diagnosis or a treatment decision is a medical device under the Medical Device Regulation [54]. Under classification Rule 11, such software is at least class IIa, and class IIb or III when the decisions it informs may cause a serious deterioration of health, a surgical intervention, death, or an irreversible deterioration of health [54]. AI systems that are medical devices, or safety-critical components of medical devices, are classified as high risk under the AI Act, which requires risk management, data governance, transparency, logging, and effective human oversight [55]. These obligations, which also include post-market monitoring and the reporting of serious incidents, are being phased in and apply in addition to the requirements of the Medical Device Regulation [55]. General-purpose LLM chatbots used for clinical decision-making have not been approved as medical devices, and some argue they would need such approval [56]. In the United States, by contrast, certain clinical decision-support functions are excluded from the statutory definition of a device if they meet four criteria, including that they support rather than replace the judgment of a health care professional and allow that professional to review the basis of the recommendation independently; the Food and Drug Administration (FDA) revised its interpretation of these criteria in January 2026 [57]. For AI-enabled functions that do fall under device oversight, a predetermined change control plan can be used to manage planned model updates after authorization [58]. The same LLM-based tool could therefore be a regulated medical device in one jurisdiction and a non-device function in another, depending on its intended use, the transparency of its outputs, and the extent to which clinicians can verify them. In Italy,

Law No. 132/2025 states that AI can be used in healthcare to support prevention, diagnosis, treatment, and therapeutic choice, with the physician remaining ultimately responsible for the decision and the patient having the right to be informed of the use of AI [59]. Liability works on the same basis. Under current law, a doctor is likely to be liable if they follow an AI recommendation that is outside the standard of care and results in harm, and potentially also where they simply ignore a correct recommendation once AI use becomes the norm [60]. Delegating the MTB function to an AI would not remove responsibility from the clinician; it would only make it harder to exercise. Cognitive risks arising from human–AI interaction are discussed in Section 7.3. Data protection is another issue, as transferring identifiable patient data to external commercial servers under the GDPR and data security requirements [34,36] would clearly disincentivize the use of such systems and favor a locally hosted or source-controlled model. Data-protection requirements also differ between jurisdictions (e.g., the General Data Protection Regulation in the EU and the Health Insurance Portability and Accountability Act in the United States), and accountability for errors should be allocated in advance among manufacturers, deploying institutions, and clinicians. Finally, there is the issue of equity. AI tools have the potential to help close the gap in specialist expertise in LMICs [41], but models trained on data and guidelines from predominantly high-income countries may well recommend treatments that are not available or appropriate in the local context [24]. Finally, regardless of jurisdiction, any deployed system requires post-deployment monitoring of performance, of drift after model or guideline updates, and of errors, with a defined process for audit and, if necessary, withdrawal [55,58].

7.3. Automation Bias, Anchoring, and Contamination of the Reference Standard

Automation bias, the tendency to over-rely on automated advice even when it is wrong, is well documented across clinical decision-support systems and is amplified by workload, time pressure, and task complexity, all of which characterize busy MTBs [61]. In breast imaging, incorrect AI suggestions degraded the performance of radiologists at all levels of experience [29]. A fluent, confident, text-based treatment plan presented before discussion may be even more persuasive than a numerical score. Related effects include anchoring, whereby the first recommendation presented frames the subsequent discussion; confirmation bias, whereby board members preferentially seek information that supports the AI draft; and complacency in checking AI outputs, which may also deskill trainees. These effects have a methodological consequence: once AI recommendations are shown to the board, the MTB decision is no longer an independent reference standard, and concordance becomes partly self-fulfilling. Retrospective studies, in which the board decision preceded the AI output, are protected from this bias, whereas prospective evaluations must generate and seal AI recommendations before, and independently of, the MTB discussion and should record how often the board changes or overrides its initial view after seeing the AI draft. Mitigation strategies include presenting AI output as information rather than as directives, displaying the guideline sources and uncertainty of each recommendation, requiring an independent clinician assessment before the AI draft is revealed for complex cases, and training staff to recognize these biases [61].

7.4. The Relational Dimension

The MTB is more than a decision engine. It is a forum where specialists negotiate uncertainty, take ownership of decisions, educate trainees, identify candidates for trials, and build the shared understanding that enables consistent communication with the patient [20,21]. In BC, treatment decisions are often preference-sensitive, such as breast conservation versus mastectomy, reconstruction, fertility preservation, the acceptability of CT toxicity given

a small absolute benefit, and treatment intensity in older patients [11]. These situations require a dialogue with the patient and have not been modeled by any current concordance study. The final step of shared decision-making with the patient is, of course, outside the scope of any AI tool and should be seen as supporting the clinician/patient relationship rather than replacing it.

7.5. Toward a Hybrid, AI-Augmented MTB

We propose a hybrid model in which AI is used as a supervised pre-meeting layer rather than a replacement (Figure 2). This workflow is a conceptual framework derived from the evidence reviewed here; it is not an established clinical approach and has not been validated. Prior to the meeting, the AI could: (i) compile a structured case summary from the EHR, imaging, pathology, and genomic reports; (ii) highlight any missing data (e.g., family history, comorbidities, patient preferences); (iii) generate a guideline-based recommendation with explicit reference to the guideline sections that were used; (iv) perform a clinical trial eligibility screen; and (v) assign a complexity score. Guideline-typical cases in high-concordance areas could then be subject to rapid review and sign-off by the appropriate specialist, with more complex/atypical/metastatic/HER2+ or flagged cases going to full MTB discussion. The triage of guideline-typical cases to expedited review, in particular, rests on retrospective concordance data and has not been prospectively validated; its safety, including the risk that relevant complexity is missed at triage and the effect of automation bias on expedited sign-off (Section 7.3), should be established in prospective studies before any clinical use (Section 7.6). In all cases, the AI draft, the final MTB decision, and the rationale for any differences would be documented to create an audit loop to monitor concordance and link to clinical registry outcomes.

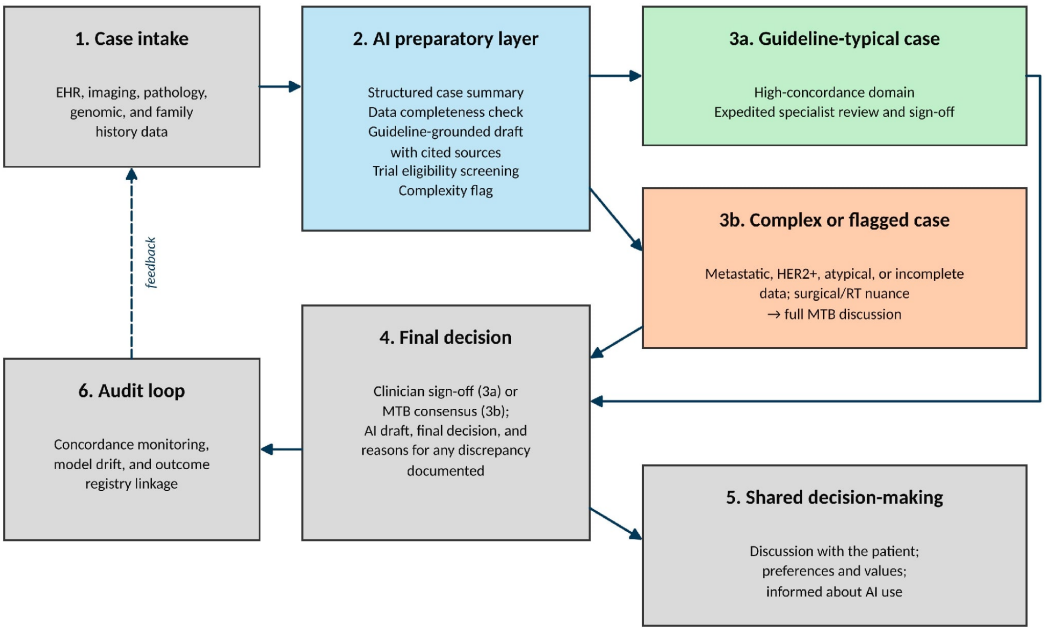


Figure 2. Proposed hybrid, AI-augmented workflow for the breast multidisciplinary tumor board (MTB). AI serves as a supervised preparatory layer, and clinicians remain responsible for the decision. The workflow is a conceptual framework proposed by the authors and has not been prospectively validated. Step 4 denotes the final decision (clinician sign-off of guideline-typical cases or MTB consensus for complex cases) and its documentation. EHR, electronic health record; HER2+, HER2-positive; RT, radiotherapy.

7.6. Research Priorities

Several priorities arise. We need prospective studies that generate AI recommendations before the MTB discussion and are completely blind to it, ideally in a multicenter randomized design comparing AI-assisted MTBs to standard MTBs. Outcomes should include concordance, time to decision, guideline adherence, implementation of the recommendation, clinician workload, patient-reported outcomes, and survival. They should also include time to treatment, toxicity, recurrence, quality of life, and equity of care, and safety endpoints (hallucination rate, guideline traceability, appropriate deferral when information is incomplete, and override rates) should be reported separately from concordance (Table 4). These studies should also test the triage step of the proposed workflow (Figure 2) explicitly, comparing expedited sign-off with full MTB discussion for cases classified as guideline-typical. Second, evaluations should be based on raw, real-world clinical documentation rather than curated summaries. Third, studies should report the model version, prompt, retrieval sources, and decoding parameters; assess reproducibility; and use harmonized metrics with blinded raters. Fourth, there is a clear need to further develop open, local, and guideline-based models to address data protection and explainability [34]. Finally, regulatory pathways for LLM-based decision support and post-market surveillance of model drift need to be defined [55,56].

7.7. Limitations of This Review

This review has inherent limitations. It is a narrative review (structured) rather than a systematic review with meta-analysis, and we did not perform study selection and appraisal in duplicate as per a registered protocol. Metric heterogeneity meant that we could not perform any quantitative pooling. We included one study available only as a conference abstract [32] to capture the rapidly moving field; its data may change upon full publication. In addition, concordance figures are reported as published: we could not independently re-adjudicate the reference decisions of the included studies against guidelines or patient-specific factors. Finally, given the rate of model development, the specific performance numbers we report here will be rapidly overtaken; the domain-level patterns and the methodological points are likely to be more enduring.

8. Conclusions

Researchers have examined whether AI could replace the BC MTB more broadly than in almost any other cancer type, but the answer is complicated. For more standard decisions (e.g., ET indication in HR+ disease, standard systemic therapy, and the need for genomic testing once a key biomarker is identified), AI recommendations are highly and consistently aligned with MTB decisions in guideline-typical cases and could be used to support, streamline, and pre-document MTB discussions. Agreement on the extent of surgery and on RT is less consistent across studies, and in more complex, higher-stakes scenarios (e.g., genetic counseling, anti-HER2 regimen choice, metastatic disease, and cases with incomplete or atypical clinical information), concordance is lower, and clinically significant errors have been reported. Overall, the current evidence base, which is largely retrospective, heterogeneous, and limited to concordance, does not support using AI to replace the BC MTB. On the basis of this limited evidence, AI currently appears most appropriate as a supervised, guideline-grounded pre-MTB support tool that summarizes the case, checks data completeness, and pre-populates the more algorithmic elements of the recommendation for clinician review. The hybrid workflow proposed here should be regarded as a framework for prospective evaluation, not as a validated clinical pathway. Full MTB discussion should remain for key decisions that depend on the individual patient (surgical and radiotherapeutic decisions, sequencing of systemic therapy in the HER2+ metastatic setting, and

any case with incomplete, atypical, or contested data). As newer models and guideline-augmented systems continue to close the concordance gap in standardized domains, the scope of what we can safely pre-populate is likely to grow. That said, the current evidence, largely retrospective and lacking outcome data, does not support its use in the areas where the BC MTB provides the most demonstrable value. Prospective, outcome-based trials of AI-assisted MTBs are now required. Safe implementation will require reproducibility testing, evaluation against clinically meaningful endpoints rather than concordance alone, explicit assessment of hallucinations and of the handling of incomplete information, robust human oversight with safeguards against automation bias, and prospective validation.

Author Contributions: Conceptualization, V.G. and M.R.V.; methodology, V.G. and D.S.; investigation, V.G., D.S., C.C., D.P. and M.R.V.; data curation and verification of extracted data, V.G. and D.P.; writing—original draft preparation, V.G.; writing—review and editing, D.S., C.C., D.P. and M.R.V.; supervision, V.G. and M.R.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The use of generative artificial intelligence in the preparation of this manuscript is described in Section 2.4 (language editing, support for literature retrieval, and pre-filling of the data-extraction template, with all extracted values and references verified by the authors against the primary sources). The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADC	antibody–drug conjugate
AI	artificial intelligence
ALND	axillary lymph node dissection
AUC	area under the curve
BC	breast cancer
CDK4/6	cyclin-dependent kinases 4 and 6
CT	chemotherapy
EHR	electronic health record
ER	estrogen receptor
ET	endocrine therapy
FDA	Food and Drug Administration
FISH	fluorescence in situ hybridization
GPT	Generative Pre-trained Transformer
HER2	human epidermal growth factor receptor 2
HR	hormone receptor
IHC	immunohistochemistry
ISH	in situ hybridization
LLM	large language model
MDR	Medical Device Regulation
MDT	multidisciplinary team
MTB	multidisciplinary tumor board
NCCN	National Comprehensive Cancer Network
OR	odds ratio
PR	progesterone receptor

RS	Recurrence Score
RT	radiotherapy
SLM	small language model
TNBC	triple-negative breast cancer
WHO	World Health Organization

References

1. Harbeck, N.; Penault-Llorca, F.; Cortes, J.; Gnant, M.; Houssami, N.; Poortmans, P.; Ruddy, K.; Tsang, J.; Cardoso, F. Breast cancer. *Nat. Rev. Dis. Primers* **2019**, *5*, 66. [PubMed]
2. Loibl, S.; Poortmans, P.; Morrow, M.; Denkert, C.; Curigliano, G. Breast cancer. *Lancet* **2021**, *397*, 1750–1769. [PubMed]
3. Perou, C.M.; Sørlie, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Rees, C.A.; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L.A.; et al. MolecUlar portraits of human breast tumours. *Nature* **2000**, *406*, 747–752. [CrossRef] [PubMed]
4. Cardoso, F.; van't Veer, L.J.; Bogaerts, J.; Slaets, L.; Viale, G.; Delaloge, S.; Pierga, J.-Y.; Brain, E.; Causeret, S.; DeLorenzi, M.; et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* **2016**, *375*, 717–729. [CrossRef] [PubMed]
5. Sparano, J.A.; Gray, R.J.; Makower, D.F.; Pritchard, K.I.; Albain, K.S.; Hayes, D.F.; Geyer, C.E., Jr.; Dees, E.C.; Goetz, M.P.; Olson, J.A., Jr.; et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med.* **2018**, *379*, 111–121. [CrossRef] [PubMed]
6. Cardoso, F.; Paluch-Shimon, S.; Schumacher-Wulf, E.; Matos, L.; Gelmon, K.; Aapro, M.S.; Bajpai, J.; Barrios, C.H.; Bergh, J.; Bergsten-Nordström, E.; et al. 6th and 7th International consensus guidelines for the management of advanced breast cancer (ABC guidelines 6 and 7). *Breast* **2024**, *76*, 103756. [CrossRef] [PubMed]
7. Loibl, S.; André, F.; Bachelot, T.; Barrios, C.H.; Bergh, J.; Burstein, H.J.; Cardoso, M.J.; Carey, L.A.; Dawood, S.; Del Mastro, L.; et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **2024**, *35*, 159–182. [CrossRef] [PubMed]
8. Biganzoli, L.; Cardoso, F.; Beishon, M.; Cameron, D.; Cataliotti, L.; Coles, C.E.; Delgado Bolton, R.C.; Trill, M.D.; Erdem, S.; Fjell, M.; et al. The requirements of a specialist breast centre. *Breast* **2020**, *51*, 65–84. [CrossRef] [PubMed]
9. Curigliano, G.; Burstein, H.J.; Gnant, M.; Loibl, S.; Cameron, D.; Regan, M.M.; Denkert, C.; Poortmans, P.; Weber, W.P.; Thürlimann, B.; et al. Understanding breast cancer complexity to improve patient outcomes: The St Gallen International Consensus Conference for the Primary Therapy of Individuals with Early Breast Cancer 2023. *Ann. Oncol.* **2023**, *34*, 970–986. [CrossRef] [PubMed]
10. Bedrosian, I.; Somerfield, M.R.; Achatz, M.I.; Boughey, J.C.; Curigliano, G.; Friedman, S.; Kohlmann, W.K.; Kurian, A.W.; Laronga, C.; Lynce, F.; et al. Germline testing in patients with breast cancer: ASCO–Society of Surgical Oncology guideline. *J. Clin. Oncol.* **2024**, *42*, 584–604. [CrossRef] [PubMed]
11. Biganzoli, L.; Battisti, N.M.L.; Wildiers, H.; McCartney, A.; Colloca, G.; Kunkler, I.H.; Cardoso, M.J.; Cheung, K.L.; de Glas, N.A.; Trimboli, R.M.; et al. Updated recommendations regarding the management of older patients with breast cancer: A joint paper from the European Society of Breast Cancer Specialists (EUSOMA) and the International Society of Geriatric Oncology (SIOG). *Lancet Oncol.* **2021**, *22*, e327–e340. [CrossRef] [PubMed]
12. Schlam, I.; Saad Menezes, M.C.; Corti, C.; Tan, A.; Abuali, I.; Tolaney, S.M. Artificial intelligence as an adjunct tool for breast oncologists—Are we there yet? *ESMO Open* **2023**, *8*, 101643. [CrossRef] [PubMed]
13. Chen, D.; Parsa, R.; Swanson, K.; Nunez, J.J.; Critch, A.; Bitterman, D.S.; Liu, F.F.; Raman, S. Large language models in oncology: A review. *BMJ Oncol.* **2025**, *4*, e000759. [CrossRef] [PubMed]
14. Chen, D.; Alnassar, S.A.; Avison, K.E.; Huang, R.S.; Raman, S. Large language model applications for health information extraction in oncology: Scoping review. *JMIR Cancer* **2025**, *11*, e65984. [CrossRef] [PubMed]
15. Sorin, V.; Glicksberg, B.S.; Artsi, Y.; Barash, Y.; Konen, E.; Nadkarni, G.N.; Klang, E. Utilizing large language models in breast cancer management: Systematic review. *J. Cancer Res. Clin. Oncol.* **2024**, *150*, 140. [CrossRef] [PubMed]
16. Konzmann, B.; Henkel, M.; Breit, C.; Wetterauer, C. Applications of large language models in tumor boards: A systematic review. *Oncol. Rev.* **2026**, *20*, 1757059. [CrossRef] [PubMed]
17. Wang, X.; Wang, Q.; Ding, G.; Wang, J.; Tang, Y.; Feng, Y. Artificial intelligence in multidisciplinary tumor boards enhancing decision making and clinical outcomes in oncology. *iScience* **2025**, *28*, 114082. [CrossRef] [PubMed]
18. Kesson, E.M.; Allardice, G.M.; George, W.D.; Burns, H.J.; Morrison, D.S. Effects of multidisciplinary team working on breast cancer survival: Retrospective, comparative, interventional cohort study of 13 722 women. *BMJ* **2012**, *344*, e2718. [CrossRef] [PubMed]
19. Tsai, C.H.; Hsieh, H.F.; Kung, P.T.; Kuo, W.Y.; Tsai, W.C. Effect of multidisciplinary team care on the risk of recurrence in breast cancer patients: A national matched cohort study. *Breast* **2020**, *53*, 68–76. [CrossRef] [PubMed]

20. Taylor, C.; Shewbridge, A.; Harris, J.; Green, J.S. Benefits of multidisciplinary teamwork in the management of breast cancer. *Breast Cancer (Dove Med. Press)* **2013**, *5*, 79–85. [CrossRef] [PubMed]
21. Schweiger, A.; Grössmann, N.; Offermanns, G.; Kratki, A.; Mittermayr, T. Tumour boards and their quality of structures, processes, and team performance in multidisciplinary cancer care: A systematic review. *BMC Health Serv. Res.* **2026**, *26*, 676. [CrossRef] [PubMed]
22. Elemento, O.; Khozin, S.; Sternberg, C.N. The use of artificial intelligence for cancer therapeutic decision-making. *NEJM AI* **2025**, *2*, AIra2401164. [CrossRef] [PubMed]
23. Somashekhar, S.P.; Sepúlveda, M.J.; Puglielli, S.; Norden, A.D.; Shortliffe, E.H.; Kumar, R.; Rauthan, A.; Kumar, N.; Patil, P.; Rhee, K. Watson for Oncology and breast cancer treatment recommendations: Agreement with an expert multidisciplinary tumor board. *Ann. Oncol.* **2018**, *29*, 418–423. [CrossRef] [PubMed]
24. Zhou, N.; Zhang, C.; Lv, H.; Hao, C.; Li, T.; Zhu, J.; Zhu, H.; Jiang, M.; Hou, H.; Liu, D.; et al. Concordance study between IBM Watson for Oncology and clinical practice for patients with cancer in China. *Oncologist* **2019**, *24*, 812–819. [CrossRef] [PubMed]
25. Lång, K.; Josefsson, V.; Larsson, A.M.; Larsson, S.; Högberg, C.; Sartor, H.; Hofvind, S.; Andersson, I.; Rosso, A. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **2023**, *24*, 936–944. [CrossRef] [PubMed]
26. Ciurescu, S.; Buciu, V.; Ilaș, D.-G.; Pârvănescu, R.; Șerban, D. Ten years of artificial intelligence in screening mammography: A systematic review and meta-analysis of diagnostic accuracy and clinical implementation (literature published 2015–2025). *Diagnostics* **2026**, *16*, 3045. [CrossRef] [PubMed]
27. Li, X.; Xiang, J.; Wang, J.; Li, J.; Wu, F.-X.; Li, M. FUNMarker: Fusion network-based method to identify prognostic and heterogeneous breast cancer biomarkers. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 2483–2491. [CrossRef] [PubMed]
28. Jiang, C.; Chen, D.; Yu, X.; Zhang, S.; Xiu, Y.; Luo, N.; Wu, J.; Zhang, X.; Chen, M.; Yang, D.; et al. The result prediction of fluorescence in situ hybridization for breast cancer patients based on machine learning and deep learning models: A multicenter study. *Cancer Imaging* **2026**, *26*, 52. [CrossRef] [PubMed]
29. Dratsch, T.; Chen, X.; Rezazade Mehrizi, M.; Kloeckner, R.; Mähringer-Kunz, A.; Püsken, M.; Baeßler, B.; Sauer, S.; Maintz, D.; Pinto dos Santos, D. Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* **2023**, *307*, e222176. [CrossRef] [PubMed]
30. Lukac, S.; Dayan, D.; Fink, V.; Leinert, E.; Hartkopf, A.; Veselinovic, K.; Janni, W.; Rack, B.; Pfister, K.; Heitmeir, B.; et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch. Gynecol. Obstet.* **2023**, *308*, 1831–1844. [CrossRef] [PubMed]
31. Griewing, S.; Knitza, J.; Boekhoff, J.; Hillen, C.; Lechner, F.; Wagner, U.; Wallwiener, M.; Kuhn, S. Evolution of publicly available large language models for complex decision-making in breast cancer care. *Arch. Gynecol. Obstet.* **2024**, *310*, 537–550. [CrossRef] [PubMed]
32. Torres-Cisneros, E.R.; Canavati-Marcos, M.; Tamez-Salazar, J.; Magallanes-Garza, G.I.; Guzmán-Huerta, E.A.; Mireles-Aguilar, T.; Gómez-Macías, G.S.; Pérez-Ramos, E.; de la Peña-Hinojosa, C.; Garza Montemayor, M.L.; et al. Decision-making concordance: A comparative analysis between a multidisciplinary tumor board and a large language model (ChatGPT) in breast cancer management. *J. Clin. Oncol.* **2024**, *42*, e13648. [CrossRef]
33. Goh, S.S.N.; Mariappan, R.; Yao, J.; Hew, F.M.; Ow, S.G.W.; Koh, W.Y.; Kumarakulasingh, N.B.; Tan, T.H.; Tai, B.C.; Hartman, M.; et al. Augmenting large language models with National Comprehensive Cancer Network guidelines for improved and standardized adjuvant therapy recommendations in postoperative breast cancer cases. *JCO Clin. Cancer Inform.* **2025**, *9*, e2400243. [CrossRef] [PubMed]
34. Griewing, S.; Lechner, F.; Gremke, N.; Lukac, S.; Janni, W.; Wallwiener, M.; Wagner, U.; Hirsch, M.; Kuhn, S. Proof-of-concept study of a small language model chatbot for breast cancer decision support—A transparent, source-controlled, explainable and data-secure approach. *J. Cancer Res. Clin. Oncol.* **2024**, *150*, 451. [CrossRef] [PubMed]
35. Ferber, D.; El Nahhas, O.S.M.; Wölflein, G.; Wiest, I.C.; Clusmann, J.; Leßmann, M.-E.; Foersch, S.; Lammert, J.; Tschochohei, M.; Jaeger, D.; et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat. Cancer* **2025**, *6*, 1337–1349. [CrossRef] [PubMed]
36. Sorin, V.; Klang, E.; Sklair-Levy, M.; Cohen, I.; Zippel, D.; Balint Lahat, N.; Konen, E.; Barash, Y. Large language model (ChatGPT) as a support tool for breast tumor board. *npj Breast Cancer* **2023**, *9*, 44. [CrossRef] [PubMed]
37. Griewing, S.; Gremke, N.; Wagner, U.; Lingenfelder, M.; Kuhn, S.; Boekhoff, J. Challenging ChatGPT 3.5 in senology—An assessment of concordance with breast cancer tumor board decision making. *J. Pers. Med.* **2023**, *13*, 1502. [CrossRef] [PubMed]
38. Stalp, J.L.; Denecke, A.; Jentschke, M.; Hillemanns, P.; Klapdor, R. Quality of ChatGPT-generated therapy recommendations for breast cancer treatment in gynecology. *Curr. Oncol.* **2024**, *31*, 3845–3854. [CrossRef] [PubMed]

39. Ah-Thiane, L.; Heudel, P.-E.; Campone, M.; Robert, M.; Brillaud-Meflah, V.; Rousseau, C.; Le Blanc-Onfroy, M.; Supiot, S.; Perennec, T.; Mervoyer, A.; et al. Large language models as decision-making tools in oncology: Comparing artificial intelligence suggestions and expert recommendations. *JCO Clin. Cancer Inform.* **2025**, *9*, e2400230. [CrossRef] [PubMed]
40. Liao, N.; Li, C.; Gradishar, W.J.; Klimberg, V.S.; Roshal, J.A.; Yuan, T.; Agarwala, S.S.; Valero, V.K.; Swain, S.M.; Margenthaler, J.A.; et al. Accuracy and reproducibility of ChatGPT responses to breast cancer tumor board patients. *JCO Clin. Cancer Inform.* **2025**, *9*, e2500001. [CrossRef] [PubMed]
41. Umihanić, Š.; Osmanović, H.; Selak, N.; Koprić, D.; Huseinbašić, A.; Šehić-Kozica, E.; Babić, B.; Umihanić, F. Evaluating the concordance between ChatGPT and multidisciplinary teams in breast cancer treatment planning: A study from Bosnia and Herzegovina. *J. Clin. Med.* **2025**, *14*, 6460. [CrossRef] [PubMed]
42. Büyükceran, E.U.; Seyfettin, A.; Babatürk, A.; Özkan, M.B.; Çolak, D.; Ünal, İ.; Kaymaz, E.; Ergün, E.; Emer, Ö. High concordance between GPT-4o and multidisciplinary tumor board decisions in breast cancer: A retrospective decision support analysis. *J. Med. Syst.* **2025**, *49*, 179. [CrossRef] [PubMed]
43. Nazlı, M.A.; Esmerer, E.; Keleş, A. Benchmarking large language models in breast cancer care: Agreement with radiology-led multidisciplinary tumor board decisions. *BMC Med. Inform. Decis. Mak.* **2026**, *26*, 243. [CrossRef] [PubMed]
44. Tokoçin, M.; Pehlivan, T.; Cin, S.; Toksöz, B.; Tokoçin, O.; Cingöz, E.; Erkoç, N.; Özen, A.; Sünnetçi Arıkan, N.; Bedir, Ş.; et al. Evaluating the role of artificial intelligence in enhancing multidisciplinary team decisions for breast cancer management. *Eur. J. Breast Health* **2026**, *22*, 184–189. [CrossRef] [PubMed]
45. Jiang, H.; Yang, C.; Zhou, W.; Yin, C.; Zhou, S.; He, R.; Ran, G.; Wang, W.; Wu, M.; Yu, J. Evaluation of large language models for radiologists' support in multidisciplinary breast cancer teams: Comparative study. *JMIR Med. Inform.* **2026**, *14*, e68182. [CrossRef] [PubMed]
46. Buonaiuto, R.; Caltavituro, A.; Di Rienzo, R.; Grieco, A.; Mangiacotti, F.P.; Longobardi, A.; Cantile, V.; Molinaro, V.; Pagliuca, M.; Buono, G.; et al. Analysis of large language model decision making in hormone receptor-positive/human epidermal growth factor receptor 2-negative early breast cancer. *JCO Clin. Cancer Inform.* **2026**, *10*, e2500230. [CrossRef] [PubMed]
47. Karabuğa, B.; Büyükkör, M.; Karabuğa, E.K.; Aydemir, E.; Kaya, O.B.; Yılmaz, M.E.; Kaptan, E.; Ateş, Ö.; Yıldız, F. Large language models as decision-support tools for adjuvant therapy planning in early-stage hormone receptor-positive breast cancer. *Cancer Diagn. Progn.* **2026**, *6*, 548–558. [CrossRef] [PubMed]
48. Doğan, İ.; Bartın, M.K.; Sönmez, E.; Seyran, E.; Bozkurt, H.A.; Yüksek, M.; Serbes, E.D.; Zalova, G.; Çelik, S. Chat GPT performance in multi-disciplinary boards—Should AI be a member of cancer boards? *Healthcare* **2025**, *13*, 2254. [CrossRef] [PubMed]
49. Modi, S.; Jacot, W.; Yamashita, T.; Sohn, J.; Vidal, M.; Tokunaga, E.; Tsurutani, J.; Ueno, N.T.; Prat, A.; Chae, Y.S.; et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N. Engl. J. Med.* **2022**, *387*, 9–20. [CrossRef] [PubMed]
50. Jo, S.J.; Chong, E.H.; Kang, I.; Yang, S.J.; Kang, B.; Kim, J.S.; Chon, H.J.; Kwon, C.-I.; Sung, M.J.; Shin, S.P.; et al. Large language models in multidisciplinary decision-making for hepatopancreatobiliary oncology: Retrospective comparative feasibility study. *J. Med. Internet Res.* **2026**, *28*, e101137. [CrossRef] [PubMed]
51. Yim, W.W.; Damm, H.; Pakull, T.M.G.; Preston, S.; Keyes, T.; Ellis-Caleo, T.J.; Sun, Z.; Yetisgen, M.; Codella, N.; Wei, M.; et al. Challenges in AI Based Tumor Board Case Summarization and Recommendations. *Res. Sq.* **2026**. [CrossRef] [PubMed]
52. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
53. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*; World Health Organization: Geneva, Switzerland, 2024. Available online: https://www.who.int/publications/i/item/9789240084759 (accessed on 16 November 2026).
54. European Parliament and Council of the European Union. Regulation (EU) 2017/745 of 5 April 2017 on medical devices. *Off. J. Eur. Union* **2017**, *L 117*, 1–175.
55. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, *L 2024/1689*, 1–144. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed on 14 October 2026).
56. Gilbert, S.; Harvey, H.; Melvin, T.; Vollebregt, E.; Wicks, P. Large language model AI chatbots require approval as medical devices. *Nat. Med.* **2023**, *29*, 2396–2398. [CrossRef] [PubMed]
57. U.S. Food and Drug Administration. *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff*; FDA: Silver Spring, MD, USA, 2026. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software (accessed on 16 November 2026).
58. U.S. Food and Drug Administration. *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff*; FDA: Silver Spring, MD, USA, 2024. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence (accessed on 16 November 2026).

59. Repubblica Italiana. Legge 23 Settembre 2025, n. 132. Disposizioni e Deleghe al Governo in Materia di Intelligenza Artificiale. *Gazzetta Ufficiale della Repubblica Italiana, Serie Generale*. 2025. Available online: https://www.gazzettaufficiale.it (accessed on 14 October 2026).
60. Price, W.N., II; Gerke, S.; Cohen, I.G. Potential liability for physicians using artificial intelligence. *JAMA* **2019**, *322*, 1765–1766. [CrossRef] [PubMed]
61. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.