

Assessing the Concordance Between AI and Oncologist Treatment Recommendations in Breast
Cancer Care: A Blinded Validation Study at Kenyatta National Hospital

Anya Quig

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Public Health

University of Washington
2025

Committee:

Stephen Hawes

Kingsley Ndoh

Program Authorized to Offer Degree:

Epidemiology

©Copyright 2025

Anya Quig

University of Washington

Abstract

Assessing the Concordance Between AI and Oncologist Treatment Recommendations in Breast
Cancer Care: A Blinded Validation Study at Kenyatta National Hospital

Anya Quig

Chair of the Supervisory Committee:

Stephen Hawes

Department of Epidemiology

Artificial intelligence (AI) has emerged as a promising tool in oncology, with the potential to support clinical decision-making and expand access to evidence-based care. This study examines the clinical validity of Gukiza, a proprietary AI platform developed by Hurone AI for cancer management. For this study, a total of 21 de-identified breast cancer patients from Kenyatta National Hospital were assessed. Treatment plans were created by Gukiza and a team of oncologists, and evaluated by an expert panel, as well as against National Comprehensive Cancer Network (NCCN) Clinical Guidelines. Results showed that AI-generated treatment plans were not significantly correlated with oncologist-generated plans, potentially reflecting limitations in the scoring framework, restricted variability in responses, or underlying differences in clinical reasoning. Although clinician-generated plans received a higher proportion of “Acceptable” and “Exceptional” ratings, this difference was not statistically significant. In

contrast, an AI-assisted analysis of AI-generated plans demonstrated strong concordance with NCCN guidelines, highlighting their potential alignment with evidence-based standards despite limited agreement with clinician ratings. Subgroup analyses indicated that postmenopausal status was associated with lower odds of treatment plan unacceptability, while the presence of comorbidities was associated with higher odds, although neither association reached statistical significance. These findings suggest that AI-generated treatment recommendations are largely NCCN guideline-adherent, and while the data from the blinded concordance assessment did not show significant correlation, the analysis highlights some of inherent challenges in comparing AI tools with clinical standards of care. Further validation in larger, more diverse patient populations is needed.

INTRODUCTION

Cancer is a leading cause of mortality worldwide, with a disproportionately high burden in low- and middle-income countries such as Kenya. Despite significant advancements in global oncology, disparities in outcomes persist.¹ In 2022, Kenya reported a total of 44,726 incident cancer cases. Breast cancer is the most frequently diagnosed cancer in Kenya, accounting for approximately 16.2% of all cancer cases annually. It is also the second leading cause of cancer-related mortality, contributing to 3,398 deaths each year, or approximately 11.6% of total cancer deaths.² The burden of cancer in Kenya is also notably higher than the global average; the age-standardized mortality rate for breast cancer in Kenya is 19.6 compared to the global average of 12.7.² This disparity is driven by systemic and structural factors such as inadequate healthcare infrastructure, a shortage of oncology specialists, and high treatment costs, which prevent many patients from receiving timely, effective, and affordable care.³⁻⁵

Artificial intelligence (AI) has emerged as a promising tool in oncology, offering the potential to improve cancer care through AI-supported screening, diagnosis, clinical decision-making, follow-up, patient navigation, and drug discovery.⁶⁻⁸ Recent developments in AI also present a unique opportunity to reduce systemic inequities across the globe, particularly in low-resource settings.^{9,10} Although many studies have explored clinical applications of AI, they are often limited in scope and scale, and few have been widely adopted in clinical settings. Notably, few studies assess the validity of these tools in settings outside of the U.S., Asia and Europe.^{11,12} Studies that have assessed AI performance in populations that are historically underserved and underrepresented in healthcare data have underscored the risk of algorithmic bias and the potential for harm when these tools are not culturally, geographically and contextually relevant.^{3,13-15}

Hurone AI is working to address this gap by building culturally sensitive AI solutions in the field of oncology. Hurone AI is a MedTech start-up company based in Seattle, WA that uses culturally sensitive AI technologies to enhance cancer care delivery and advance biopharma research, with a particular focus on historically underrepresented populations (<https://hurone.ai/>).¹⁶ Hurone's platform, Gukiza, is a web-based interface that integrates electronic medical records (EMR), remote patient monitoring, AI-supported care planning, and advanced analytics.¹⁶ Gukiza also uses the National

Comprehensive Cancer Network (NCCN) Harmonized Guidelines™ as published in the Journal of the National Comprehensive Cancer Network (JNCCN) 2023;21(6)).^{11,17–19} Designed as a co-pilot for oncologists, Gukiza aims to improve clinical decision-making by analyzing patient medical histories and clinical characteristics to suggest contextually relevant treatment options. It can support multidisciplinary tumor boards, automate referrals, and significantly reduce the administrative burden on clinicians, potentially saving up to 50% of their time.¹⁶ By leveraging culturally appropriate AI technologies, Gukiza seeks to augment oncologists' capabilities, reduce preventable hospitalizations, mitigate financial toxicity, and provide efficient remote care to populations that have been historically underserved.

While platforms like Gukiza have great potential to transform cancer care, clinical validation in diverse patient settings remains a critical step for effective implementation.¹² Clinical validation is essential to ensure that AI algorithms are both accurate and reliable when used with real patient data.²⁰ This is particularly critical in regions where healthcare systems face digital infrastructure and data limitations. In several parts of Africa, for example, low digitization and limited use of electronic medical records, along with the underrepresentation of African ancestry in genomic research and clinical trials more broadly result in a lack of representative health data. Not only that, but many AI models have been developed using clinical guidelines from high-income countries, which may not fully account for the region-specific treatment protocols and healthcare infrastructure challenges in sub-Saharan Africa.²¹ This further limits the generalizability of many existing models to historically underrepresented and underserved populations.^{12,22}

This study aims to evaluate the clinical validity of the Gukiza platform among breast cancer patients at Kenyatta National Hospital (KNH) in Nairobi, Kenya. Specifically, the study examines how closely AI-generated treatment plans align with those developed by oncologists, evaluates their consistency with internationally recognized clinical guidelines, and explores whether message quality varies based on patients' clinical or demographic characteristics. Insights from this analysis will be used to refine the platform's functionality and ensure that it is effective in real-world settings. Ultimately, the findings will contribute to the equitable integration of AI technologies into cancer care in resource-constrained settings, advancing global efforts to reduce cancer disparities.

METHODS

Study Design

This analysis examines de-identified medical records from 21 breast cancer patients who were receiving care at KNH during the study period. The purpose of this comparative study was to evaluate the validity of AI-generated guidance by assessing its concordance with oncologist-generated guidance, evaluating adherence of AI-generated treatment plans to NCCN Harmonized Guidelines™ for Sub-Saharan Africa, and determining clinical and sociodemographic correlates of appropriate treatment recommendations.

This sub-study is part of a larger feasibility study currently being conducted at Kenyatta National Hospital (KNH) in Nairobi, Kenya. The primary study uses a quasi-experimental, nonequivalent groups post-test design, enrolling 120 breast cancer patients over a six-month period. The primary aim of that study is to assess the feasibility, effectiveness and cost-effectiveness of Hurone AI's premier platform in enhancing cancer care at the Cancer Treatment Center in KNH. Appropriate joint institutional and ethical review boards at the University of Washington and the University of Nairobi/KNH approved this secondary analysis and the parent study protocol.

Study Setting

Patient data was collected at Kenyatta National Hospital (KNH) in Nairobi, Kenya. KNH is the largest tertiary hospital in Kenya, receiving 800 cancer service visits per month. Breast cancer patients comprise approximately 12% of cancer patients seen at KNH. KNH was selected for the primary study due to its status as a major national referral hospital with a comprehensive cancer treatment center that serves a large, diverse patient population.

Study Population

The study population for this concordance analysis consisted of de-identified medical records from 21 breast cancer patients with sociodemographic and clinical variables of interest who were receiving care at the Cancer Treatment Center in KNH between March and May 2025. Patients were purposively sampled by oncologists at KNH. Patients were selected to capture diversity in age, menopause status and cancer type, and receptor status.

Oncologists participating in the expert panel were recruited to represent a variety of healthcare settings across Kenya. Recruiting aimed to ensure representation from diverse geographic, institutional and areas of clinical expertise across the country.

Data Collection

Upon study enrollment, data on clinical and demographic factors were collected through medical record abstraction. Key covariates of interest included sex, age, clinical history of chronic comorbidities, substance use, obstetrics and gynecology history, familial history of cancer, breast cancer status, diagnostic workup and treatment history (Appendix 1). Comprehensive case summaries were generated by Gukiza and adjudicated by the research team. AI-generated treatment plans were developed by prompting Gukiza to interpret the case summaries and recommend treatment options. To serve as a comparison group, independent oncologists from Kenya created comparable treatment plans. Oncologist treatment plans reflected the clinical standard of care in Kenya and were based on their professional expertise and clinical guidelines. Oncologists were provided a template to ensure that their recommendations followed the same format as those produced by Gukiza. A blinded expert panel of six oncologists then evaluated the quality and adherence of both AI-generated and oncologist-created treatment plans to clinical guidelines. Treatment plans were scored using a 4-point Likert scale which included the following categories:

- *Inappropriate:* Recommendations were not suitable for the patient's condition and could potentially harm the patient. Recommendations contradict established clinical guidelines or lack relevance to the reported case.
- *Needs Improvement:* Recommendations may require significant modification or lack sufficient justification to be considered appropriate. Recommendations were somewhat generic and lacked personalization based on patient-specific data. Important aspects of the patient's condition were not addressed adequately.
- *Appropriate:* Recommendations were generally sound and aligned with standard practices for the patient's condition. Recommendations were personalized, taking into account the patient's clinical and demographic data. Recommendations provided clear, appropriate and actionable treatment steps.
- *Exceptional:* Recommendations were exceptionally well-suited to the patient's condition, highly personalized, and based on the latest clinical guidelines. Recommendations were comprehensive and addressed all relevant aspects of the patient's case. Recommendations provided clear and detailed treatment guidance, demonstrating a high level of understanding of the patient's clinical history.

Additionally, Manus (version 1.6.3 - 1.8.5), an advanced large language model (LLM)-based AI agent, was used to systematically evaluate the concordance of breast cancer treatment plans with NCCN Guidelines for Breast Cancer (Version 4.2023) that are publicly available on the Journal of the National Comprehensive Cancer Network (JNCCN) 2023;21(6))¹⁹. A structured prompting methodology was developed consisting of the following components:

1. Initial Case Presentation Prompt: Patient cases were presented to Manus in a standardized format containing:
 - Patient demographics and risk factors
 - Disease characteristics (histology, biomarker status, staging)
 - Diagnostic workup results
 - Current and recommended treatment plans

2. Validation Instruction Prompt: Manus was instructed to:
 - Systematically evaluate each treatment recommendation against specific NCCN guidelines
 - Provide precise page references from JNCCN (Volume 21, Issue 6, 2023)
 - Include direct quotes from the guidelines for all assessments
 - Classify recommendations using a three-tier system:
 - a. *SUPPORTED*: Recommendations directly aligned with NCCN guidelines
 - b. *PARTIALLY SUPPORTED*: Recommendations with some supporting evidence but with caveats
 - c. *NOT SUPPORTED*: Recommendations that contradicted guidelines or lacked supporting evidence

3. Output Format Prompt: Manus was directed to generate structured validation reports containing:
 - Patient summary
 - Individual assessment of each recommendation with citations
 - Summary of overall guideline concordance

To ensure accuracy of the AI-generated assessments, several reports were randomly selected to cross-check the accuracy of the assessments. Discrepancies or unclear interpretations were resolved

through human expert review. The prompting methodology was iteratively refined based on validation accuracy.

Study Measures

For aim 1, our primary outcome of interest, concordance between AI- and oncologist-generated treatment plans, was measured by comparing expert panel scores (AI versus oncologist generated). Scores were converted to numeric (1 = Inappropriate, 2 = Needs Improvement, 3 = Appropriate and 4 = Exceptional) with higher scores indicating more concordant treatment recommendations. For aim 2, the adherence of AI-generated treatment recommendations with NCCN guidelines was measured and classified as “Supported”, “Partially Supported”, and “Not Supported”. For aim 3, We determined the odds of appropriate concordance scores, dichotomized as acceptable (appropriate or exceptional) or unacceptable (inappropriate or needs improvement) by treatment recommendation type (AI versus oncologist generated). Demographic covariates included age, menopause status, cancer stage, presence of comorbidities, and evidence of metastasis. Age (<40 years, 40+ years), menopause status (pre- vs. post-menopausal), presence of comorbidities (none, any), and evidence of metastasis (absent, present) were dichotomous variables, while cancer stage was a categorical variable (stage 2, stage 3, stage 4).

Data Analysis

Descriptive statistics summarized the baseline breast cancer case characteristics. Continuous and categorical variables were presented as counts using mean $\pm$ standard deviation (SD) for normally distributed data and by the median and interquartile range (IQR) for non-normally distributed variables. To address the first aim, we assessed the concordance between AI-generated guidance and guidance provided by oncologists. Scores were converted from ordinal to numeric values of 1-4. Descriptive statistics were used to summarize scores from both Gukiza and the oncologists. Intraclass correlation coefficient (ICC) and pairwise weighted Cohen’s kappa were used to examine intra-rater reliability between oncologists. ICC was appropriate for this situation as it can compare multiple (2+) oncologists and appropriately account for the ordinal nature of their ratings.²³ We also computed Pearson’s correlation

coefficient to assess inter-rater reliability, comparing scores for AI-generated vs. oncologist-generated plans. To formally test whether Gukiza scores differed significantly from oncologist scores, we used a Wilcoxon signed-rank test. Statistical significance was evaluated at a two-sided alpha level of 0.05.

To address the second aim, descriptive statistics were used to calculate the percentage of AI-generated treatment recommendations classified as “Supported,” “Partially Supported,” or “Not Supported” within each patient’s treatment plan, recognizing that the total number of recommendations varied between patients. These percentages were then averaged across all patients to provide mean proportions and corresponding standard deviations. Each recommendation was then assigned a numeric score (Supported = 1, Partially Supported = 0.5, Not Supported = 0). Scores were summed for each treatment plan and divided by the total number of recommendations in that plan to calculate a percentage of concordance with the NCCN guidelines. Finally, to better understand recommendations that did not fully align with NCCN guidelines, we examined cases with lower concordance scores as case examples.

The third aim was an exploratory analysis that examined the association between patient sociodemographic and clinical characteristics, such as age, menopause status, cancer stage, and receptor status, and the likelihood of receiving acceptable (appropriate or exceptional) versus unacceptable (inappropriate or needs improvement) AI-generated treatment recommendation. Mixed effects logistic regression clustered by patient was used to model the odds and corresponding 95% confidence interval of acceptable compared with unacceptable messaging for each given exposure.

RESULTS

A total of 21 de-identified patient profiles were examined. The mean age of patients was 51.9 years (SD: 12.1) and 12/21 (57.1%) were reported to be postmenopausal (Table 1). 6/21 (28.6%) had a family history of breast cancer and 7/21 (33.3%) reported at least one chronic condition including dementia, diabetes, HIV, hypertension and valvular heart disease. 9/21 (42.9%) were reported to have Stage 3 cancer, 7 (33.3%) were reported to have Stage 2, and 5/21 (23.8%) were reported to have Stage 4 (Table 2). 16/21 patients (76.2%) had a positive hormone receptor (HR) whereas 5/21 (23.8%) had triple negative breast cancer (TNBC). Participants were undergoing a variety of treatments including

chemotherapy (52.4%), surgery (28.6%), and palliative chemotherapy (9.5%). At the time of data collection, no participants were reported to be in remission.

The expert review panel consisted of six oncologists with varied areas of clinical practice that included medical oncology, radiation oncology, and clinical research working in academic, public or private institutions in urban, semi-urban and rural settings (Table 3). 4/6 (66.7%) panelists had 5-10 years of experience in clinical oncology, while 2/6 (33.3%) had 11-20 years of experience.

Overall, intra-rater reliability among oncologists was moderate (ICC = 0.71, 95% CI: 0.47, 0.86) (Table 4). Notably, reliability was stronger for oncologists' ratings of AI-generated treatment plans (ICC = 0.78, 95% CI: 0.49, 0.93), while reliability was lower for oncologist-generated treatment plans (ICC = 0.28, 95% CI: -0.67, 0.78). Pairwise weighted Cohen's kappa statistics were also calculated to provide more granular detail on intra-rater reliability between individual oncologists. Weighted kappas were computed separately for AI-generated and oncologist-generated treatment plans, as well as overall. These results are visualized in the heat maps presented in Figures 1–3. A moderate negative correlation was observed between the average ratings of AI-generated and oncologist-generated recommendations ($R = -0.35$, 95% CI: -0.68 to 0.10, $p = 0.12$). However, this association was not statistically significant. Distribution of scores between AI- and oncologist-generated treatment plans were examined (Figure 4). Overall, the oncologist-generated plans received a greater proportion of “Acceptable” and “Exceptional” scores (64.2%) than AI-generated plans (54.7%). The mean score of AI-generated plans (2.94) was slightly lower than oncologist generated plans (3.24). A Wilcoxon signed-rank test, however, indicated no statistically significant difference between the mean scores of the two groups (V statistic = 71, $p = 0.1$).

For the AI-guided analysis of concordance with NCCN Clinical guidelines, 377 specific recommendations were examined across 21 patient profiles, with the mean number of recommendations being 18.0 per case summary (SD: 6.2, Range: 9–34). Across treatment plans, the mean percentage of recommendations classified as “Supported” was 77.2% (SD: 14.4%), while “Partially Supported” recommendations accounted for a mean of 16.7% (SD: 9.4%), and “Not Supported” recommendations accounted for a mean of 6.2% (SD: 7.7%). Composite concordance scores, summarized in Table 5,

demonstrated a high level of overall agreement with NCCN guidelines, with a mean score of 0.87. Scores ranged from 0.73 to 1.00.

To explore potential factors driving variation in scores, cases at the extremes were further analyzed as case studies. One such case was case #17, which received a composite concordance score of 0.73. This case involved a 56-year-old patient with stage 2 HER2-positive, hormone receptor-negative breast cancer. The AI-generated plan for this case made several recommendations that were either “Partially Supported” or “Not Supported” by NCCN guidelines. For example, Gukiza recommended considering extended HER2-targeted therapy beyond 1 year. NCCN guidelines acknowledge the potential benefit of extended HER2-targeted therapy in specific contexts, but it is not routinely recommended for all patients. It also recommended genomic assays to guide extended endocrine therapy, which NCCN guidelines discuss primarily in the context of HR+/HER2-negative disease, with no mention of their use to guide extended endocrine therapy in HER2-positive disease. Gukiza also recommends considering newer HER2-targeted agents in the neoadjuvant/adjuvant setting, such as tucatinib, trastuzumab deruxtecan, or margetuximab. These agents have not been recommended by NCCN guidelines outside of clinical trials despite showing promising results.¹⁹ Case #20, who was a 60-year-old postmenopausal female with metastatic triple-negative breast cancer (TNBC) and significant comorbidities, on the other hand, received a perfect concordance score. Recommendations, which were all supported by NCCN guidelines, included continued systemic therapy with second-line agents, monitoring of renal, hepatic and cardiac function, palliative and other supportive care, and involvement family and caregivers in end-of-life care and other decision making.¹⁹

Finally, the exploratory subgroup analysis examined the association between the unacceptability of AI-generated messages and patient age, menopause status, cancer stage, presence of comorbidities, and evidence of metastasis (Table 6). Patients who were postmenopausal were somewhat less likely to have a treatment plan scored by an oncologist as unacceptable compared with those who were premenopausal (OR: 0.25, 95% CI: 0.05, 1.19, $p = 0.08$). Patients who had any comorbidity were somewhat more likely to have a treatment plan scored by an oncologist as unacceptable compared with those who had no

comorbidities (OR: 3.95, 95% CI: 0.77, 20.2, $p = 0.10$). However, neither of these associations reached statistical significance in bivariate mixed effects logistic regression.

DISCUSSION

Our results showed that AI-generated treatment plan scores were not significantly correlated with oncologist-generated treatment plans. This lack of correlation may reflect limitations in the scoring system, limited variability in responses, or fundamental differences in clinical reasoning. Although oncologist-generated plans received a greater proportion of “Acceptable” and “Exceptional” ratings overall, the difference in scoring was not statistically significant. Together, these findings suggest that while oncologist generated plans scored slightly better overall, the limited variability in scores and the study design may have constrained the ability to detect meaningful associations or differences between the two rating sources. On the other hand, AI-generated plans showed strong concordance with NCCN guidelines. This outcome is unsurprising given that Gukiza utilizes NCCN guidelines.

Overall intra-rater reliability was moderate. While pairwise weighted Cohen’s kappa between oncologists varied greatly, at times reflecting substantial disagreement, including a situation where one rater assigned a 4 while another assigned a 1, the overall ICC examining inter-rater reliability was moderate.²⁴ Conversely, Pearson’s correlation coefficient showed that AI-generated treatment plan scores were not significantly correlated with oncologist-generated treatment plans. It is difficult to determine the cause of the lack of correlation between scores in this study. One possible explanation is that oncologists in Kenya utilize a variety of clinical guidelines in addition to NCCN guidelines. Previous studies, including one within Kenya, have demonstrated low concordance between NCCN guidelines and recommendations made in clinical practice.^{25,26} Treatment decisions in the region are often shaped by patient affordability, given the high burden of out-of-pocket costs, as well as limited availability of certain therapies.^{25,26} Additionally, variability in care may reflect differences in institutional protocols or the preference for alternative guidelines such as those from European Society for Medical Oncology (ESMO) or American Society for Medical Oncology (ASCO). Another possible explanation is variation in how oncologists approached the scoring process itself. As shown in Figures 1–3, some oncologists, such as

Oncologist #1, consistently had low weighted kappa scores, indicating limited agreement with the other reviewers. This could suggest differences in interpretation of the scoring criteria or in thresholds for what constitutes an appropriate plan, contributing to overall low concordance and substantial variability between raters. Additionally, low variability between patients may have biased measurements of correlation. Koo and Lee recommend including a minimum of 30 heterogeneous subjects and at least three raters when examining ICC.²⁴ With 21 subjects and significant clustering of scores between 3 and 4 as shown in Figure 4, this may not be satisfied. The Wilcoxon-signed rank test showed there was no statistically significant difference in the ratings of treatment plans generated by Gukiza compared with those generated by oncologists. However, the relatively small p-value and average higher scores for clinicians suggests a non-statistically significant trend favoring oncologist-generated treatment plans.

When comparing Gukiza-generated output with NCCN Clinical Guidelines, the average overall concordance across case summaries was 0.87, indicating a high level of agreement. In many cases, “Partially Supported” or “Not Supported” recommendations did not represent outright errors by Gukiza, as illustrated by case #17. Many of the recommendations that were labeled “Partially Supported” or “Not Supported” appeared to be experimental or alternative approaches that were not explicitly recommended by NCCN guidelines but may have shown effectiveness in clinical trials and been recommended for other patient subgroups. There was minimal evidence of hallucinations, or situations where information was completely falsified and not based on real data. Previous studies have shown that hallucinations are a common issue in AI-tools for clinical decision making.^{27–29} Gukiza was created specifically for clinical oncology applications, and unlike more general models like ChatGPT, which are trained on open-domain data, it may be less prone to hallucinations. This aligned with a recent study which showed that smaller fine-tuned models could outperform larger zero-source models on specific tasks.³⁰ This is a promising finding for the use of these specialized models in clinical practice.

Subgroup analysis using mixed effects logistic regression should be interpreted with caution due to limited statistical power; however, it may provide directions for future research. While not statistically significant, the presence of comorbidities was associated with higher odds of unacceptable AI-generated treatment plans. Comorbidities may complicate approaches to clinical decision making and require

greater nuance. Not only that, but individuals with comorbidities are often excluded from clinical trials, suggesting limited data on those with chronic conditions.^{31,32} Postmenopausal status was also associated with a lower odds of unacceptable scores from oncologists, although again this was not statistically significant. Early-onset breast cancer tends to exhibit more aggressive histopathological and genomic features compared to breast cancer in older women, which may require more complex treatment options.³³ Subgroup analysis is a critical component of validation, as it enables the identification of performance differences across patient sub-populations. Understanding this can optimize performance for all patients by detecting bias and that may lead to systemically worse guidance for certain subgroups.

This study contributes to the growing body of literature on applications of AI in oncology. Although AI tools in oncology are advancing rapidly, evidence of their efficacy in breast cancer care remains mixed. While some studies have demonstrated high accuracy in tasks such as information extraction and responding to structured questions, others have highlighted challenges related to tasks like complex clinical decision making.^{34–38} This study adds to the existing body of knowledge in key ways. Many previous studies focus on binary or discrete outcomes, such as whether an AI model can accurately detect cancer from imaging or predict survival based on clinical data.³⁹ Relatively few studies have examined the overall quality and appropriateness of AI-generated treatment plans. Evaluating complete patient treatment plans presents a more nuanced challenge with respect to assessing concordance or clinical agreement. This is one of few studies that evaluates the quality and appropriateness of AI-generated cancer treatment plans in direct comparison with local oncologists and evidence-based guidelines for oncology. Others include a 2018 study which compared a tool called Watson for Oncology (WFO) with a multidisciplinary tumor board in India. While this study found high overall concordance (93%), their measurement of concordance differed substantially; both “Recommended” and “For Consideration” treatment options were counted as concordant, leading to higher overall concordance scores.⁴⁰ It is difficult to determine whether differences in concordance are a result of the AI model used, or the methods used to evaluate it. Another recent study comparing ChatGPT-4o-generated treatment plans with a multidisciplinary tumor board reported low concordance, with statistically significant differences in rating, and raters agreeing more often with the tumor board recommendations.³⁸ Similar to

this study, treatment plans were rated using a Likert scale, but differing methodology make direct comparison difficult.

Additionally, few studies to date have focused on the application of AI tools for cancer management in sub-Saharan Africa. Countries with limited healthcare coverage and low per capita health expenditure have demonstrated a greater willingness to adopt innovative technologies like AI.⁴¹ Despite the interest in implementation, AI-driven tools have been adopted more slowly in Africa than other parts of the world.¹¹ This study represents one of the few efforts to address these gaps by assessing clinical validity of AI models in an African context.

While this study is an important step towards validating Gukiza, several key limitations should be acknowledged. Although this study provides valuable insight into the implementation of Gukiza at Kenyatta National Hospital, the generalizability of these findings to other populations should be carefully considered. KNH was an appropriate site for this study because it provided a representative mix of breast cancer patients. Not only that, but this study was designed to assess the platform's performance within the KNH patient population ahead of broader integration. Although the findings are encouraging and suggest potential applicability beyond Kenya, further validation studies should be conducted.

As mentioned previously, there are methodological challenges to evaluating complex clinical decision-making systems. Analyzing correlation between the AI- and oncologist-generated treatment plans offers valuable insights on how Gukiza performs relative to oncologists; however, it does not fully capture the complexity of clinical decision-making. This is one of the challenges of evaluating unstructured data. The use of a Likert scale provides a structured approach to evaluating treatment plans, but it does not necessarily account for all the nuances involved. To address this, we added an AI-supported concordance analysis of the specific treatment recommendations, but future studies should consider more detailed content analyses, such as a mixed methods approach, to further assess the alignment of AI-generated guidance with established clinical standards. These methodological challenges are compounded by the inherent subjectivity of clinical decision-making. Guidance can vary not only across different countries, where guidelines, training, and resource availability differ, but even among oncologists within the same country. A strength of this study was the diverse panel of six oncologists.

Nonetheless, variability in experience among oncologists, and differences in the ways that they approach the scoring process could have contributed to inconsistencies in scoring, and as mentioned above, impacted the measurement of correlation.

With regard to the comparison to NCCN guidelines, NCCN represents one of many clinical guidelines in the field of oncology. While NCCN offers harmonized guidelines specific to sub-Saharan Africa, guidelines from other regional and international bodies such as the European Society for Medical Oncology (ESMO), American Society of Clinical Oncology (ASCO) are also widely used in Kenya. Future studies should examine concordance with these guidelines as well. Finally, the moderate sample size limited statistical power to detect small effect sizes and generalize findings from this study. This was of particular concern with the subgroup analysis using mixed effects logistic regression. Future studies should include larger sample sizes to improve statistical power.

CONCLUSION

Despite these limitations, the findings of this study demonstrate the promise of AI-powered platforms like Gukiza in augmenting oncology care in settings where oncologists experience a high workload and limited support infrastructure. AI tools should not be considered a replacement for medical professionals, but rather an integrative tool that can augment and improve diagnostic and therapeutic processes, providing critical support in clinical decision-making.⁶ By demonstrating concordance with NCCN guidelines, this study suggests that Gukiza and similar tools can play a meaningful role in supporting high-quality cancer care in resource limited settings. While the data from the blinded concordance assessment did not show significant correlation, it provides valuable insight into the methodology used to assess concordance, as well as the inherent challenges in comparing AI tools with clinical standards of care. Validation studies are an important benchmark for successful implementation of AI tools, but further studies should examine whether these results translate to improved patients' outcomes, cost-effectiveness, patient-experience, and clinician workload.^{42,45} With continued refinement, clinical integration, and real-world evaluation, AI tools like Gukiza can reduce barriers to cancer care delivery, both within Kenya and in other resource-constrained settings globally.

TABLES AND FIGURES

Table 1. Baseline demographic and general clinical characteristics of breast cancer cases stratified by menopausal status for a sample of breast cancer patients at Kenyatta National Hospital in Nairobi, Kenya, 2025

	Overall (N=21)	Postmenopausal (N=12)	Premenopausal (N=9)
Sex			
Female	21 (100%)	12 (100%)	9 (100%)
Age (years)			
Median [Min, Max]	54.0 [34.0, 76.0]	58.0 [49.0, 76.0]	39.5 [34.0, 48.0]
BMI category			
Underweight	0 (0%)	0 (0%)	0 (0%)
Normal weight	6 (28.6%)	3 (25.0%)	3 (33.3%)
Overweight	7 (33.3%)	4 (33.3%)	3 (33.3%)
Obese	8 (38.1%)	5 (41.7%)	3 (33.3%)
Parity			
Median [Min, Max]	2.00 [1.00, 6.00]	2.00 [1.00, 6.00]	2.00 [1.00, 3.00]
Missing	6 (28.6%)	5 (41.7%)	1 (11.1%)
Age at menarche (years)			
Median [Min, Max]	15.0 [12.0, 19.0]	15.0 [12.0, 16.0]	15.5 [12.0, 19.0]
Missing	9 (42.9%)	6 (50.0%)	3 (33.3%)
Family history			
Yes	6 (28.6%)	4 (33.3%)	2 (22.2%)
<i>Breast cancer, unknown degree</i>	2 (9.5%)	1 (8.3%)	1 (11.1%)
<i>Other, first degree</i>	2 (9.5%)	2 (16.7%)	0 (0%)
<i>Other unknown degree</i>	1 (4.8%)	0 (0%)	1 (11.1%)
<i>Unknown type, unknown degree</i>	1 (4.8%)	1 (8.3%)	0 (0%)
No	9 (42.9%)	4 (33.3%)	5 (55.6%)
Unknown	6 (28.6%)	4 (33.3%)	2 (22.2%)

Any chronic comorbidity

None	14 (66.7%)	6 (50.0%)	8 (88.9%)
Any	7 (33.3%)	6 (50.0%)	1 (11.1%)
<i>Dementia</i>	<i>1 (4.8%)</i>	<i>1 (8.3%)</i>	<i>0 (0%)</i>
<i>Diabetes</i>	<i>1 (4.8%)</i>	<i>1 (8.3%)</i>	<i>0 (0%)</i>
<i>HIV</i>	<i>1 (4.8%)</i>	<i>1 (8.3%)</i>	<i>0 (0%)</i>
<i>Hypertension</i>	<i>5 (23.8%)</i>	<i>4 (33.3%)</i>	<i>1 (11.1%)</i>
<i>Valvular heart disease</i>	<i>1 (4.8%)</i>	<i>1 (8.3%)</i>	<i>0 (0%)</i>

Table 2. Breast cancer clinical characteristics stratified by menopause status, for a sample of breast cancer patients at Kenyatta National Hospital in Nairobi, Kenya, 2025

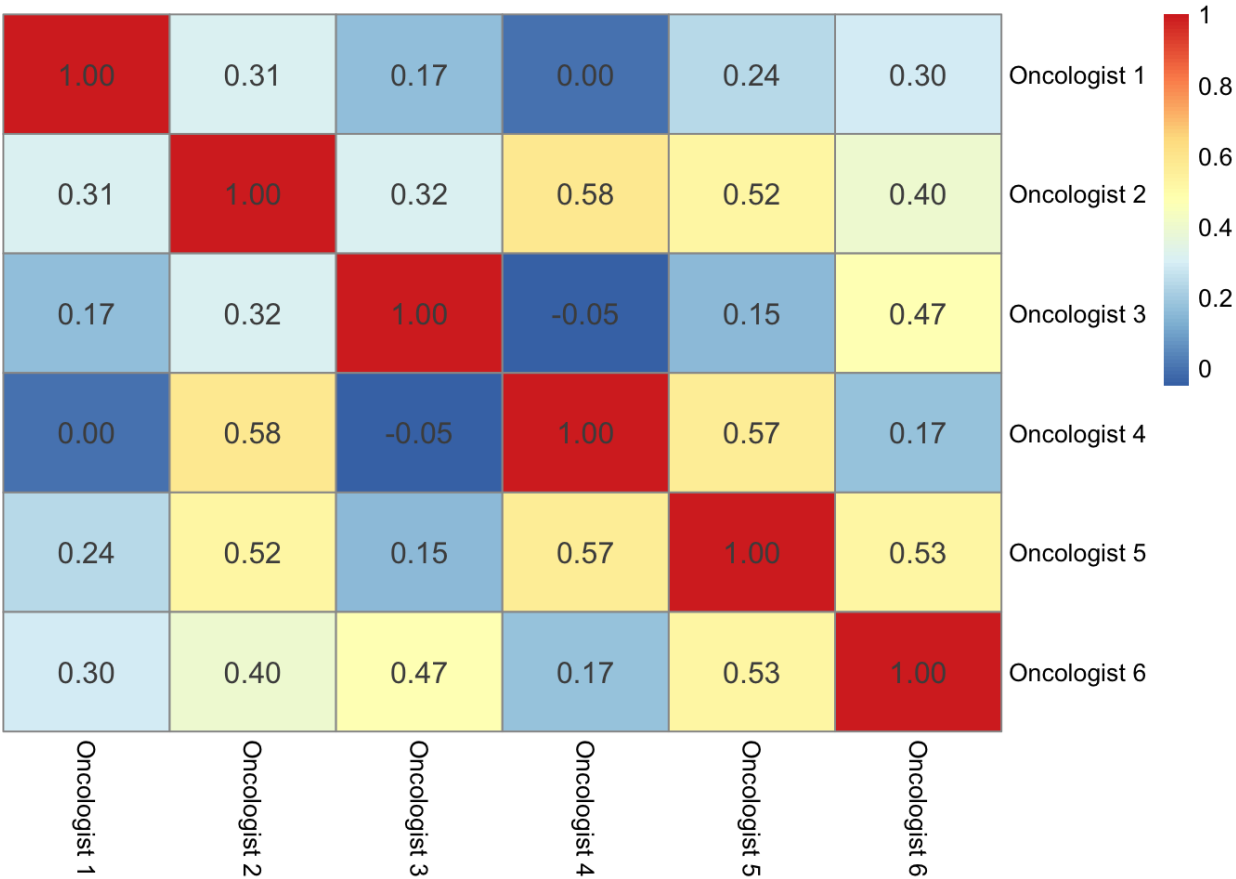
	Overall (N=21)	Postmenopausal (N=12)	Premenopausal (N=9)
Cancer stage			
Stage 2	7 (33.3%)	4 (33.3%)	3 (33.3%)
Stage 3	9 (42.9%)	5 (41.7%)	4 (44.4%)
Stage 4	5 (23.8%)	3 (25.0%)	2 (22.2%)
Breast side affected			
Left	13 (61.9%)	8 (66.7%)	5 (55.6%)
Right	8 (38.1%)	4 (33.3%)	4 (44.4%)
Tumor size			
2–5 cm	11 (52.4%)	7 (58.3%)	4 (44.4%)
> 5 cm	3 (14.3%)	2 (16.7%)	1 (11.1%)
> 5 cm with skin/muscle involvement	2 (9.5%)	1 (8.3%)	1 (11.1%)
Skin/muscle involvement	5 (23.8%)	2 (16.7%)	3 (33.3%)
Cancer location			
Central	4 (19.0%)	4 (33.3%)	0 (0%)
Lower inner quadrant	2 (9.5%)	2 (16.7%)	0 (0%)
Lower outer quadrant	2 (9.5%)	0 (0%)	2 (22.2%)

Upper inner quadrant	1 (4.8%)	1 (8.3%)	0 (0%)
Upper outer quadrant	8 (38.1%)	4 (33.3%)	4 (44.4%)
Multiple			
<i>Lower outer and central quadrant</i>	<i>1 (4.8%)</i>	<i>1 (8.3%)</i>	<i>0 (0%)</i>
<i>Upper and lower outer quadrant</i>	<i>2 (9.5%)</i>	<i>0 (0%)</i>	<i>2 (22.2%)</i>
<i>Upper inner and outer quadrant</i>	<i>1 (4.8%)</i>	<i>0 (0%)</i>	<i>1 (11.1%)</i>
Nodal involvement			
0–3 nodes	9 (42.9%)	6 (50.0%)	3 (33.3%)
4–9 nodes	5 (23.8%)	2 (16.7%)	3 (33.3%)
> 9 nodes	7 (33.3%)	4 (33.3%)	3 (33.3%)
Metastasis			
Absent	16 (76.2%)	9 (75.0%)	7 (77.8%)
Present	5 (23.8%)	3 (25.0%)	2 (22.2%)
<i>Bone</i>	<i>1 (4.8%)</i>	<i>0 (0%)</i>	<i>1 (11.1%)</i>
<i>Lung</i>	<i>2 (9.5%)</i>	<i>1 (8.3%)</i>	<i>1 (11.1%)</i>
<i>Lung and one</i>	<i>2 (9.5%)</i>	<i>2 (16.7%)</i>	<i>0 (0%)</i>
Hormone receptor status			
Hormone receptor positive	16 (76.2%)	8 (66.7%)	8 (88.9%)
Triple negative (TNBC)	5 (23.8%)	4 (33.3%)	1 (11.1%)
In remission			
Female	21 (100%)	12 (100%)	9 (100%)
Treatment received			
Chemotherapy	7 (33.3%)	4 (33.3%)	3 (33.3%)
Palliative chemotherapy	2 (9.5%)	2 (16.7%)	0 (0%)
Surgery (Mastectomy, lumpectomy), Chemotherapy	4 (19.0%)	3 (25.0%)	1 (11.1%)
Surgery (Mastectomy, lumpectomy)	2 (9.5%)	0 (0%)	2 (22.2%)
Missing	6 (28.6%)	3 (25.0%)	3 (33.3%)

Table 3. Demographic characteristics of oncologist expert panel

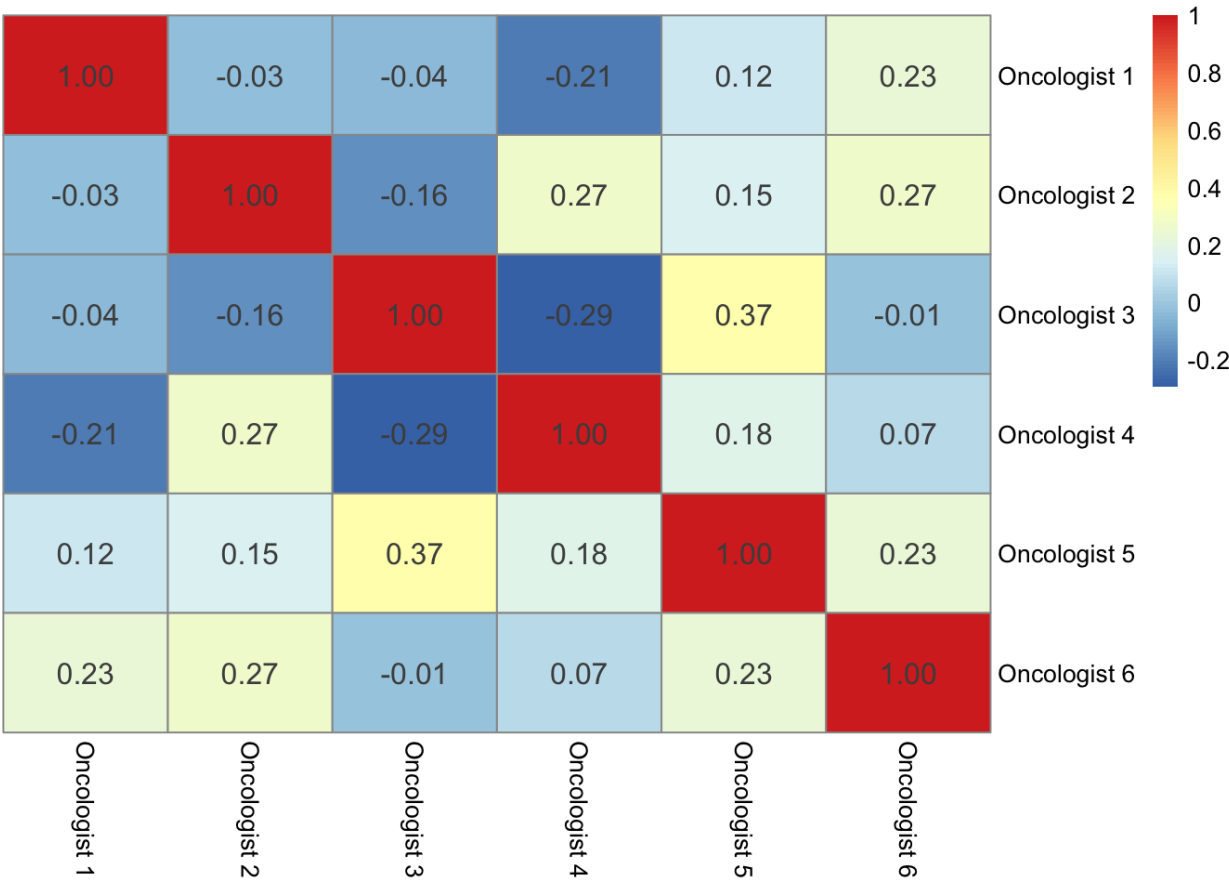
	Overall (N=6)
Area of oncology practice	
Medical Oncology	3 (50.0%)
Medical and Radiation Oncology	2 (33.3%)
Radiation Oncology	1 (16.6%)
Years of experience	
5-10 years	4 (66.6%)
11-20 years	2 (33.3%)
Primary work setting	
Academic institution	2 (33.3%)
Public institution	2 (33.3%)
Public institution; Private institution	1 (16.6%)
Academic institution; Public institution; Private institution	1 (16.6%)
Primary work setting location	
Rural	2 (33.3%)
Semi-urban	1 (16.6%)
Urban	3 (50.0%)

Figure 1. Heatmap displaying weighted pairwise kappa statistics assessing inter-rater agreement among an expert panel of oncologists evaluating AI-generated treatment plans using squared weights.¹



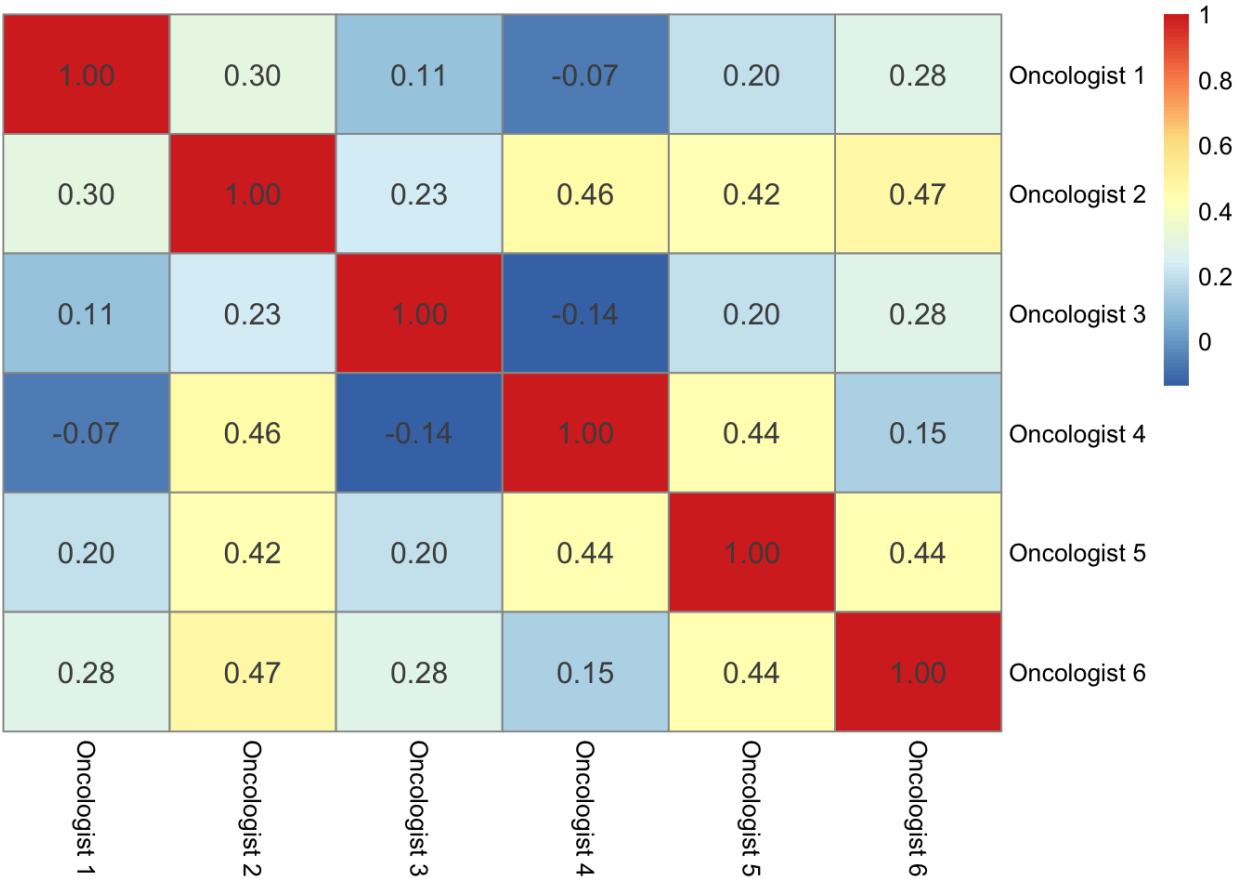
1. Squared weights were used to address the ordinal nature of the scoring data, meaning disagreements are weighted according to their squared distance from perfect agreement (<https://search.r-project.org/CRAN/refmans/irr/html/kappa2.html>).

Figure 2. Heatmap displaying weighted pairwise kappa statistics assessing inter-rater agreement among an expert panel of oncologists evaluating oncologist-generated treatment plans using squared weights.¹



1. Squared weights were used to address the ordinal nature of the scoring data, meaning disagreements are weighted according to their squared distance from perfect agreement (<https://search.r-project.org/CRAN/refmans/irr/html/kappa2.html>).

Figure 3. Heatmap displaying weighted pairwise kappa statistics assessing inter-rater agreement among an expert panel of oncologists evaluating AI-generated and oncologist-generated treatment plans using squared weights.¹



1. Squared weights were used to address the ordinal nature of the scoring data, meaning disagreements are weighted according to their squared distance from perfect agreement (<https://search.r-project.org/CRAN/refmans/irr/html/kappa2.html>).

Table 4. Intraclass Correlation Coefficient (ICC) as a measure of inter-rater reliability among expert panel of oncologists

<i>Type</i>	<i>ICC</i>	<i>95% CI</i>	<i>p-value</i>
AI-Generated	0.78	[0.49, 0.93]	< 0.001
Oncologist-Generated	0.28	[-0.67, 0.78]	0.21
Combined	0.71	[0.47, 0.86]	< 0.001

Figure 4. Violin plot comparing distribution of scores between AI-generated and oncologist-generated treatment plans

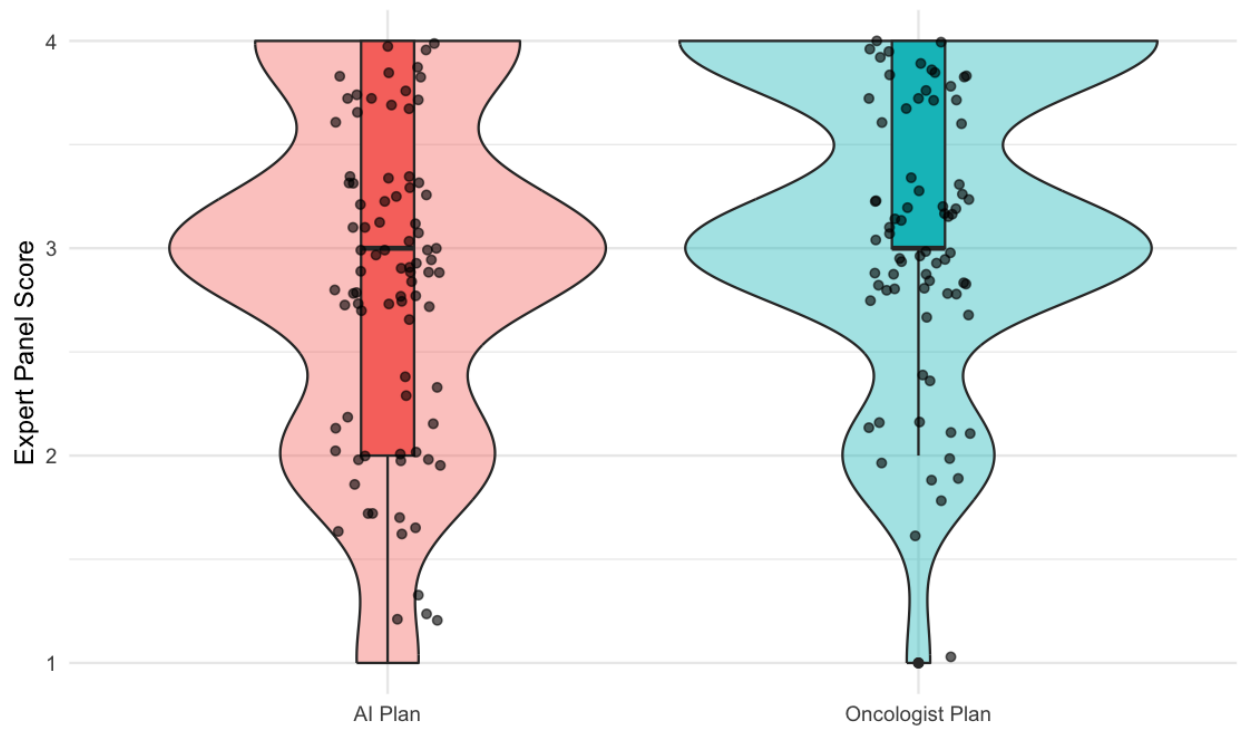


Table 5. Mean Percentage of Supported, Partially Supported, and Not Supported AI Recommendations Within Patient Treatment Plans¹

<i>ID</i>	<i>Composite Score</i>
Case 1	0.89
Case 2	0.95
Case 3	0.75
Case 4	0.95
Case 5	0.93
Case 6	0.74
Case 7	0.88
Case 8	0.86
Case 9	0.86
Case 10	0.93
Case 11	0.92
Case 12	0.87
Case 13	0.87
Case 14	0.90
Case 15	0.89
Case 16	0.80
Case 17	0.73
Case18	0.86
Case 19	0.75
Case 20	1.00
Case 21	0.95

1. Composite scores were calculated for each patient by scoring ‘Supported’ recommendations with 1.0, ‘Partially Supported’ recommendations with 0.5 and ‘Not Supported’ recommendations with 0. Total scores were divided by the total number of recommendations within each treatment plan to account for variation by treatment plan.

Table 6. Unadjusted odds ratios (and 95% confidence intervals) from mixed effects logistic regression models examining the association between unacceptability of AI-generated treatment plans and patient clinical and sociodemographic characteristics.

<i>Characteristic</i>	<i>OR</i>	<i>95% CI</i>	<i>p-value</i>
Age Category			
Under 40	—	—	
40 and Over	0.42	0.06, 3.07	0.4
Menopause Status			
Premenopausal	—	—	
Postmenopausal	0.25	0.05, 1.19	0.08
Stage			
Stage 2	—	—	
Stage 3	2.87	0.58, 14.1	0.2
Stage 4	2.54	0.38, 17.0	0.3
Comorbidity			
None	—	—	
Any	3.95	0.77, 20.2	0.1
Metastasis			
Absent	—	—	
Present	1.46	0.23, 9.16	0.7

REFERENCES

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229-263. doi:10.3322/caac.21834
2. Ferlay J, Ervik M, Lam F, et al. Global Cancer Observatory: Kenya. Global Cancer Observatory: Cancer Today. 2024. <https://gco.iarc.who.int/media/globocan/factsheets/populations/404-kenya-fact-sheet.pdf>
3. Omotoso O, Teibo JO, Atiba FA, et al. Addressing cancer care inequities in sub-Saharan Africa: current challenges and proposed solutions. *Int J Equity Health.* 2023;22(1):189. doi:10.1186/s12939-023-01962-y
4. Gakunga R, Kinyanjui A, Ali Z, et al. Identifying Barriers and Facilitators to Breast Cancer Early Detection and Subsequent Treatment Engagement in Kenya: A Qualitative Approach. *The Oncologist.* 2019;24(12):1549-1556. doi:10.1634/theoncologist.2019-0257
5. Ngwa W, Addai BW, Adewole I, et al. Cancer in sub-Saharan Africa: a Lancet Oncology Commission. *Lancet Oncol.* 2022;23(6):e251-e312. doi:10.1016/S1470-2045(21)00720-8
6. Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. *Br J Cancer.* 2022;126(1):4-9. doi:10.1038/s41416-021-01633-1
7. Kolla L, Parikh RB. Uses and limitations of artificial intelligence for oncology. *Cancer.* 2024;130(12):2101-2107. doi:10.1002/cncr.35307
8. Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov.* 2021;11(4):900-915. doi:10.1158/2159-8290.CD-21-0090
9. Bockarie MJ, Ansumana R, Machingaidze SG, et al. Transformative potential of artificial intelligence on health care and research in Africa. *Int J Infect Dis.* 2024;143:107011. doi:10.1016/j.ijid.2024.107011
10. Dankwa-Mullan I, Ndoh K, Akogo D, Rocha HAL, Juaçaba SF. Artificial Intelligence and Cancer Health Equity: Bridging the Divide or Widening the Gap. *Curr Oncol Rep.* 2025;27(2):95-111. doi:10.1007/s11912-024-01627-1
11. Adeoye J, Akinshipo A, Thomson P, Su YX. Artificial intelligence-based prediction for cancer-related outcomes in Africa: Status and potential refinements. *J Glob Health.* 2022;12:03017. doi:10.7189/jogh.12.03017
12. Owoyemi A, Owoyemi J, Osiyemi A, Boyd A. Artificial Intelligence for Healthcare in Africa. *Front Digit Health.* 2020;2:6. doi:10.3389/fdgth.2020.00006
13. Farlow A, Hoffmann A, Tadesse GA, et al. Rethinking global digital health and AI-for-health innovation challenges. Deo S, ed. *PLOS Glob Public Health.* 2023;3(4):e0001844. doi:10.1371/journal.pgph.0001844
14. Mohamed S, Png MT, Isaac W. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philos Technol.* 2020;33(4):659-684. doi:10.1007/s13347-020-00405-8

15. Dankwa-Mullan I. Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine. *Prev Chronic Dis*. 2024;21:240245. doi:10.5888/pcd21.240245
16. Hurone AI - Our Solutions. Hurone AI. <https://hurone.ai/our-solutions/>
17. NCCN Harmonized Guidelines™. National Comprehensive Cancer Network. 2025. <https://www.nccn.org/global/what-we-do/harmonized-guidelines>
18. Sussman L, Garcia-Robledo JE, Ordóñez-Reyes C, et al. Integration of artificial intelligence and precision oncology in Latin America. *Front Med Technol*. 2022;4:1007822. doi:10.3389/fmedt.2022.1007822
19. Gradishar WJ, Moran MS, Abraham J, et al. NCCN Guidelines® Insights: Breast Cancer, Version 4.2023: Featured Updates to the NCCN Guidelines. *J Natl Compr Canc Netw*. 2023;21(6):594-608. doi:10.6004/jnccn.2023.0031
20. Park SH, Choi J, Byeon JS. Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. *Korean J Radiol*. 2021;22(3):442. doi:10.3348/kjr.2021.0048
21. Akingbola A, Adegbesan A, Ojo O, Otumara JU, Alao UH. Artificial Intelligence And Cancer Care in Africa. *J Med Surg Public Health*. 2024;3:100132. doi:10.1016/j.glmedi.2024.100132
22. Bentley AR, Callier SL, Rotimi CN. Evaluating the promise of inclusion of African ancestry populations in genomics. *Npj Genomic Med*. 2020;5(1):5. doi:10.1038/s41525-019-0111-x
23. Sainani KL. Reliability Statistics. *PM&R*. 2017;9(6):622-628. doi:10.1016/j.pmrj.2017.05.001
24. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
25. Mezger NCS, Seraphin TP, Ballé R, et al. NCCN guideline–concordant cancer care in sub-Saharan Africa: a population-based multicountry study of 5 cancers. *JNCI J Natl Cancer Inst*. 2025;117(1):120-133. doi:10.1093/jnci/djae221
26. Mezger NCS, Hämmerl L, Griesel M, et al. Guideline Concordance of Treatment and Outcomes Among Adult Non-Hodgkin Lymphoma Patients in Sub-Saharan Africa: A Multinational, Population-Based Cohort. *The Oncologist*. 2023;28(11):e1017-e1030. doi:10.1093/oncolo/oyad157
27. Chen S, Kann BH, Foote MB, et al. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol*. 2023;9(10):1459. doi:10.1001/jamaoncol.2023.2954
28. Huang Y, Gomaa A, Semrau S, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. *Front Oncol*. 2023;13:1265024. doi:10.3389/fonc.2023.1265024
29. Zhou J, Zhang J, Wan R, et al. Integrating AI into clinical education: evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. *BMC Med Educ*. 2025;25(1):406. doi:10.1186/s12909-025-06916-2
30. Zhang G, Jin Q, Zhou Y, et al. Closing the gap between open source and commercial large language models for medical evidence summarization. *Npj Digit Med*. 2024;7(1):239. doi:10.1038/s41746-024-

31. Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatment. *CA Cancer J Clin*. 2016;66(4):337-350. doi:10.3322/caac.21342
32. Fortin M. Randomized Controlled Trials: Do They Have External Validity for Patients With Multiple Comorbidities? *Ann Fam Med*. 2006;4(2):104-108. doi:10.1370/afm.516
33. Franco I, Alshalalfa M, Hernandez A, et al. Genomic Characterization of Aggressive Breast Cancer in Younger Women. *Ann Surg Oncol*. 2023;30(12):7569-7578. doi:10.1245/s10434-023-14080-4
34. Sorin V, Glicksberg BS, Artsi Y, et al. Utilizing large language models in breast cancer management: systematic review. *J Cancer Res Clin Oncol*. 2024;150(3):140. doi:10.1007/s00432-024-05678-6
35. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol*. 2023;20(10):990-997. doi:10.1016/j.jacr.2023.05.003
36. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J*. 2023;41(3):209-216. doi:10.3857/roj.2023.00633
37. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *Npj Breast Cancer*. 2023;9(1):1-4. doi:10.1038/s41523-023-00557-8
38. Karabuğa B, Karaçin C, Büyükkör M, et al. The Role of Artificial Intelligence (ChatGPT-4o) in Supporting Tumor Board Decisions. *J Clin Med*. 2025;14(10):3535. doi:10.3390/jcm14103535
39. Zhang B, Vakanski A, Xian M. Bi-Rads-Net: An Explainable Multitask Learning Approach for Cancer Diagnosis in Breast Ultrasound Images. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE; 2021. doi:10.1109/mlsp52302.2021.9596314
40. Somashekhar SP, Sepúlveda MJ, Puglielli S, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol*. 2018;29(2):418-423. doi:10.1093/annonc/mdx781
41. *What Doctor? Why AI and Robotics Will Define New Health.*; 2017.
<https://www.pwc.com/gx/en/news-room/docs/what-doctor-why-ai-and-robotics-will-define-new-health.pdf>
42. Knop M, Weber S, Mueller M, Niehaves B. Human Factors and Technological Characteristics Influencing the Interaction of Medical Professionals With Artificial Intelligence–Enabled Clinical Decision Support Systems: Literature Review. *JMIR Hum Factors*. 2022;9(1):e28639. doi:10.2196/28639
43. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res*. 2020;22(6):e15154. doi:10.2196/15154
44. Hantel A, Walsh TP, Marron JM, et al. Perspectives of Oncologists on the Ethical Implications of Using Artificial Intelligence for Cancer Care. *JAMA Netw Open*. 2024;7(3):e244077. doi:10.1001/jamanetworkopen.2024.4077
45. Macheka S, Ng PY, Ginsburg O, Hope A, Sullivan R, Aggarwal A. Prospective evaluation of artificial

intelligence (AI) applications for use in cancer pathways following diagnosis: a systematic review.
BMJ Oncol. 2024;3(1):e000255. doi:10.1136/bmjonc-2023-000255

Appendix 1. Clinical and demographic variables used in the creation of de-identified breast cancer patient profiles.

Variable	Type	Values
Age	Continuous	<i>NA</i>
Sex	Categorical	Male Female
Weight (kg)	Continuous	<i>NA</i>
Height (cm)	Continuous	<i>NA</i>
Chronic disease history	Categorical	Diabetes Hypertension Kidney Disease Osteoporosis Autoimmune Disease e.g. Lupus, Rheumatoid Arthritis Chronic respiratory disease e.g. COPD, Asthma HIV None Unknown Other
Family history of cancer	Categorical	None Breast cancer Ovarian cancer BRCA mutation Unknown Other
Smoking status	Categorical	Current Former Never Unknown
Alcohol consumption history	Categorical	None Occasional Frequent Unknown
Body mass index [weight (kg) / height (m ²)]	Continuous	<i>NA</i>
Age at menarche (years)	Continuous	<i>NA</i>

Number of full term pregnancies	Continuous	<i>NA</i>
Contraceptive history	Categorical	None Oral contraceptive use Hormone replacement therapy Unknown Other
Menopausal status	Categorical	Premenopausal Postmenopausal Unknown
Breast side affected	Categorical	Left Right Bilateral
Location	Categorical	Upper outer quadrant Upper inner quadrant Lower outer quadrant Lower inner quadrant Central
Axillary lymph node involvement on physical examination	Free text	<i>NA</i>
Mammography	Free text	<i>NA</i>
Ultrasound	Free text	<i>NA</i>
CT Scan	Free text	<i>NA</i>
MRI	Free text	<i>NA</i>
PET Scan	Free text	<i>NA</i>
Lymph node involvement	Free text	<i>NA</i>
Biopsy type	Categorical	None Core needle Excisional
Histology	Categorical	Invasive ductal carcinoma Invasive lobular carcinoma Mixed Mucinous Lymphatic type Undifferentiated/anaplastic Ductal carcinoma insitu (DCIS)

		Other
Estrogen receptor (ER) status	Categorical	Positive Negative Not available
Progesterone receptor (PR) status	Categorical	Positive Negative Not available
HER2 status	Categorical	Positive Negative Not available
Lymph node involvement based on biopsy findings	Free text	<i>NA</i>
Tumor size	Categorical	< 2cm 2–5 cm > 5 cm Skin/muscle involvement
Nodal involvement (Based on clinical, imaging and/or histology/cytology)	Categorical	0–3 nodes 4–9 nodes > 9 nodes
Metastasis	Categorical	Present Absent
Distant metastasis	Categorical	Lung Liver Bone Brain None Other
Clinical stage	Categorical	Stage 1 Stage 2 Stage 3 Stage 4
High risk features	Categorical	Lymphovascular invasion Grade 3 Perineural invasion None Unknown Other
Breast cancer treatment	Categorical	Surgery (Mastectomy, lumpectomy)

received so far		Chemotherapy Radiation therapy Hormone therapy e.g Tamoxifen, Letrozole Targeted therapy e.g CDK 4/6 inhibitors Immunotherapy Other
Provide a treatment summary	Free text	<i>NA</i>
In remission	Categorical	Yes No

Appendix 2. Summary of scores across raters for AI- vs. oncologist-generated treatment plans using mean (SD) and median (IQR).¹

<i>Patient ID</i>	<i>AI-Generated</i>		<i>Oncologist-Generated</i>	
	<i>Mean (SD)</i>	<i>Median (IQR)</i>	<i>Mean (SD)</i>	<i>Median (IQR)</i>
Case 1	3.00 (0.89)	3.0 (1.50)	2.83 (0.75)	3.0 (0.75)
Case 2	2.83 (0.75)	3.0 (0.75)	3.67 (0.52)	4.0 (0.75)
Case 3	3.17 (0.75)	3.0 (0.75)	2.83 (0.41)	3.0 (0.00)
Case 4	3.33 (0.52)	3.0 (0.75)	3.50 (0.55)	3.5 (1.00)
Case 5	2.83 (0.98)	2.5 (1.75)	3.17 (0.75)	3.0 (0.75)
Case 6	3.50 (0.84)	4.0 (0.75)	3.17 (0.75)	3.0 (0.75)
Case 7	3.17 (0.41)	3.0 (0.00)	3.83 (0.41)	4.0 (0.00)
Case 8	3.17 (0.75)	3.0 (0.75)	3.33 (0.82)	3.5 (1.00)
Case 9	3.17 (0.41)	3.0 (0.00)	3.33 (0.52)	3.0 (0.75)
Case 10	2.50 (1.05)	2.5 (1.00)	3.00 (1.10)	3.0 (0.75)
Case 11	2.67 (0.58)	3.0 (0.50)	3.67 (0.58)	4.0 (0.50)
Case 12	3.00 (0.82)	3.0 (1.00)	4.00 (0.00)	4.0 (0.00)
Case 13	3.00 (0.00)	3.0 (0.00)	3.33 (1.15)	4.0 (1.00)
Case 14	3.67 (0.47)	4.0 (0.50)	2.33 (0.58)	2.0 (0.50)
Case 15	2.67 (0.47)	3.0 (0.50)	3.33 (1.15)	4.0 (1.00)
Case 16	2.00 (0.00)	2.0 (0.00)	3.67 (0.58)	4.0 (0.50)
Case 17	1.17 (0.37)	1.0 (0.00)	3.33 (0.52)	3.0 (0.75)
Case 18	3.00 (0.82)	3.0 (1.00)	3.33 (0.58)	3.0 (0.50)
Case 19	3.33 (0.47)	3.0 (0.50)	3.00 (1.00)	3.0 (1.00)
Case 20	3.33 (0.94)	4.0 (1.00)	3.00 (1.00)	3.0 (1.00)
Case 21	3.33 (0.47)	3.0 (0.50)	2.33 (1.53)	2.0 (1.50)