original report

Artificial Intelligence Treatment Decision Support For Complex Breast Cancer Among Oncologists With Varying Expertise

Fengrui Xu, MD[1]; Martín-José Sepúlveda, MD, ScD[2]; Zefei Jiang, MD[3]; Haibo Wang, MD[4]; Jianbin Li, MD[5]; Yongmei Yin, MD[6]; Zhenzhen Liu, MD[7]; M. Christopher Roebuck, PhD[8]; Edward H. Shortliffe, MD, PhD[9]; Min Yan, MD[7]; Yuhua Song, MD[4]; Cuizhi Geng, MD, PhD[10]; Jinhai Tang, MD[6]; and Kyu Rhee, MD, MPP[11]

abstract

PURPOSE The aim of the current study was to assess treatment concordance and adherence to National Comprehensive Cancer Network breast cancer treatment guidelines between oncologists and an artificial intelligence advisory tool.

PATIENTS AND METHODS Study cases of patients (N = 1,977) who were at high risk for recurrence or who had metastatic disease and cell types for which the advisory tool was trained were obtained from the Chinese Society for Clinical Oncology cancer database (2012 to 2017). A cross-sectional observational study was performed to examine treatment concordance and guideline adherence among an artificial intelligence advisory tool and 10 oncologists with varying expertise—three fellows, four attending physicians, and three chief physicians. In a blinded fashion, each oncologist provided treatment advice on an average of 198 cases and the advisory tool on all cases (N = 1,977). Results are reported as rates and logistic regression odds ratios.

RESULTS Concordance for the recommended treatment was 0.56 for all physicians and higher for fellows compared with chief and attending physicians (0.68 *v* 0.54; 0.49; $P = .001$). Concordance differed by hormone receptor subtype–TNM stage, with the lowest for hormone receptor–positive human epidermal growth factor receptor 2/neu-positive cancers (0.48) and highest for triple-negative breast cancers (0.71) across most TNM stages. Adherence to National Comprehensive Cancer Network guidelines was higher for oncologists compared with the advisory tool (0.96 *v* 0.82; $P < .003$) and lower for fellows compared with attending physicians (0.93 *v* 0.98; 0.96; $P = .04$).

CONCLUSION Study findings reflect a complex breast cancer case mix, the limits of medical knowledge regarding optimum treatment, clinician practice patterns, and use of a tool that reflects expertise from one cancer center. Additional research in different practice settings is needed to understand the tool's scalability and its impact on treatment decisions and clinical and health services outcomes.

INTRODUCTION

Rapid advances in breast cancer research pose information management challenges for oncologists. Computing systems that provide clinical decision support have been designed to help address this challenge but have not been widely adopted in clinical practice, despite their continuing evolution and improvement. Knowledge-based artificial intelligence (AI) systems for cancer treatment, such as ONCOCIN and Kasimir/Case Based Reasoning, have existed for some time,[1,2] and, more recently, quantitative AI systems using machine learning and related computational approaches have been developed.[3] The Watson for Oncology treatment advisory tool (WfO) is an example of a quantitative oncology clinical decision support that leverages the clinical expertise of oncologists at Memorial Sloan Kettering Cancer Center (MSK).

WfO has been used in multidisciplinary tumor board settings and its recommendations have been compared with those rendered independently by such boards[4,5]; however, multidisciplinary tumor boards are not available routinely for many practicing oncologists[6] and little is known about the use of this technology at the level of individual oncologists. We therefore undertook studies to examine the relationship between treatment decisions of individual oncologists and the WfO advisory tool without assistance from a multidisciplinary tumor board. Here, we report results of the first study, to our knowledge, that focuses on concordance of treatment recommendations and adherence to clinical guidelines in a large population of patients with complex breast cancers among oncologists with three levels of expertise—fellows, attending physicians, and chief physicians.

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on June 14, 2019 and published at ascopubs.org/journal/cci on August 16, 2019: DOI https://doi.org/10.1200/CCI.18.00159

PATIENTS AND METHODS

Study Design and Patient Population

This cross-sectional observational study examined concordance (level of agreement) between oncologists (physician prescribers) and the IBM Watson for Oncology treatment advisory tool on breast cancer treatment and their respective adherence to National Comprehensive Cancer Network (NCCN) treatment guidelines.[7] Ten oncologists from five cancer centers with varying years of clinical experience participated in the study: three fellows (< 3 years), four attending physicians (3 to 8 years), and three chief physicians (> 8 years). Each clinician independently assessed an average of 198 different cases, and WfO assessed all 1,977 cases. The study protocol was approved by the institutional review boards of the 307th Hospital, Affiliated Hospital of Qingdao University, Jiangsu Province Hospital, Henan Cancer Hospital, and the Tumor Hospital of Hebei Province.

Cases among women age 18 to 65 years—for which the advisory tool had been trained—were assessed in the study. Patients included had ductal, lobular, metaplastic, or mixed histology tumors. Histologic tissue types for which the advisory tool had not yet been trained—for example, adenocystic, tubular, and secretory breast cancers—were excluded. Furthermore, cases were restricted to those that required complex therapeutic decisions so that the performance of the technology could be assessed in this challenging subset of breast cancers. These were patients with disease at high risk of recurrence—that is, they met one or more of the following criteria: age 35 years or younger, tumors larger than 2 cm, axillary lymph node positive, human epidermal growth factor receptor 2 (HER2)/neu positive, or triple negative tumors[8] —and patients with established metastatic disease that required first- or second-line therapy.

Data Collection and Concordance Determination

Data were collected for cases from the Chinese Society of Clinical Oncology breast cancer database which contains voluntarily registered, anonymized cases from 10 hospitals representing all regions of China. Demographic, medical, surgical, laboratory, pathology, and molecular data for each case were abstracted from the database and entered into an electronic data capture system (Medrio, San Francisco, CA)[9] by five oncologists who were trained to abstract the required elements from clinical records and to prepare a standard template for each case. As a result of the absence of Chinese standards regarding *BRCA* and multigene testing panels at the time of the study, these tumor attributes were unavailable for the cases. Data abstracters were trained on the advisory tool's user interface and secured and entered the attributes required by the tool to generate treatment recommendations. Case templates were numbered and allocated to physician prescribers in a systematic but nonrandom fashion according to the following rules: cases originating from a prescriber's hospital would not be eligible for assignment to that individual; only early breast cancer cases (stages I to III) would be assigned to surgical oncologists as they do not prescribe treatment of metastatic tumors; and allocation would occur upon prescriber enrollment and seek to provide equal numbers of early and metastatic breast cancer cases to medical oncologists. All prescribers independently reviewed their assigned cases and were blinded to the recommendations of the advisory tool.

Concordance was defined as a prescriber treatment recommendation falling into the advisory tool's categories of Recommended or For consideration. The tool's Recommended therapies are those preferred by MSK experts and supported by strong scientific evidence. For consideration therapies are considered reasonable alternatives by MSK experts. Nonconcordance was defined as a prescriber treatment recommendation falling into the advisory tool's category of Not recommended or not on the list of the tool's recommendations. This category contains treatment regimens that MSK experts advise against for such reasons as a high risk of harm from a patient's existing comorbidities or strong evidence of inferior outcomes. It also contains treatments that are available in China but not contained in the advisory tool's knowledge base.

WfO Advisory Tool

The WfO advisory tool is designed to assist oncologists in making treatment decisions. It is an AI advisory tool composed of a reasoning engine, a knowledge base, and data drawn from clinical records. The reasoning engine employs machine learning software that uses scoring features of patient attributes that influence breast cancer treatment decisions—for example, organ function laboratory values, tumor stage, hormone receptor (HR) status, and prior response to therapy—and MSK breast cancer training cases. The resulting breast cancer module, when approved by MSK collaborators, is then made available for use in the advisory tool. The knowledge base contained in WfO provides a curated body of knowledge that also includes documents from more than 300 medical journals and textbooks, MSK treatment guidelines (developed on the basis of national treatment guidelines and the center's clinical experience), and literature hand selected by MSK experts. Data abstracted from clinical records comprise the third component of the system and include demographic, medical, surgical, laboratory, pathology, genomic, and molecular attributes for each case. Detailed descriptions of how the WfO advisory tool works and has been evaluated during development are contained in an online supplement published in *Annals of Oncology*.[10] WfO version 17.6 was used in the current study.

Statistical Analysis

Distributions of patient age, tumor stage (I, IIA, IIB, IIIA, IIIB, IIIC, stage IV first line, and stage IV second line), and

receptor subtype (HR positive HER2/neu negative, HR positive HER2/neu positive, HR negative HER2/neu positive, or triple-negative breast cancer [TNBC]) were examined to assess whether there were differences in the physician-reviewed cases. Statistical significance of these potential differences was derived using the Kruskal-Wallis equality of populations test.[11] Mean values of concordance and adherence were calculated overall, as well as by physician level, physician type, stage, and receptor subtype. Statistical significance of pairwise differences in means were assessed using the Wilcoxon signed rank test.[12] All means (as proportions), standard deviations, and *P* values are reported. Multivariable logistic regression models of concordance and NCCN guidelines adherence were also estimated and resulting odds ratios and 95% CIs are reported in Appendix Tables A1 and A2.

RESULTS

Table 1 lists the profiles of breast cancer cases by level of physician expertise. Case profiles are similar in terms of age, TNM stage distribution, and receptor subtype with one exception: Stage IV second-line therapy cases represented a significantly smaller proportion of chief physician–reviewed cases compared with other physician groups (0.08 *v* 0.24 to 0.25; $P < .001$).

Table 2 presents mean concordance rates between physician experience groups and the advisory tool. Overall agreement was 0.56 but differed among physicians. Fellows demonstrated a higher rate of concordance than either chief physicians or attending physicians (0.68 *v* 0.54 and 0.49, respectively; $P = .001$). Significant differences in concordance were also observed within HR subtypes by TNM stage (Table 2). Relative to stage IA, all physicians combined had significantly higher rates of concordance for five HER2/neu positive tumor types (HR-negative HER2/neu-positive stages IIA, IIB, IIIA, and IIIC; and HR-positive HER2/neu-positive stage IIB; 0.67 to 0.97; $P \leq .01$). Among physician groups, chief physicians had significantly higher concordance rates in four of these HER2/neu-positive tumor types (HR-negative stages IIA/B, IIIA/C; 0.90 to 1.00; $P < .001$), attending physicians in one (HR-negative stage IIIA; 1.00; $P = .04$), and fellows in four (HR-negative stages IIA/B, IIIA; 0.95 to 1.00; $P \leq .007$; HR-positive stages IIA/B; 0.94 to 1.00; $P \leq .003$). TNBC also exhibited significantly higher concordance relative to stage IA for physicians overall in stages IIA, IIB, and IIIA (0.96 to 1.00; $P = .000$). This pattern also occurred in the three physician groups. Finally, logistic regression modeling that accounted for physician experience group, patient age, TMN status, receptor subtype, and their interactions confirmed the bivariable findings presented in Table 2 and also showed that older age was associated with higher concordance rates (Appendix Table A1).

Figure 1 shows overall concordance by adjuvant therapy or metastatic therapy subtype. Concordance rates for individual components of adjuvant therapy were much higher than the rates for overall treatment regimens (0.56), ranging from 0.78 for adjuvant endocrine therapy to 1.00 for adjuvant targeted therapy. In contrast, treatment concordance for first- and second-line metastatic disease were modest and approximated the overall concordance rate (0.52 and 0.50, respectively *v* 0.56). Examples of adjuvant therapy nonconcordance are described in Appendix Table A3.

Table 3 presents reasons for nonconcordance overall and by prescriber level of experience. Three quarters of nonconcordant cases were a result of either disagreement on appropriate treatments, particularly for adjuvant endocrine or adjuvant chemotherapy (Fig 1), or instances in which the advisory tool's recommendations did not adhere to NCCN guidelines. The availability of treatments in China that were not contained in the advisory tool's knowledge base and the unavailability in China of treatments that were contained in the advisory tool accounted for the remaining nonconcordant cases.

Adherence to NCCN treatment guidelines was exceptional for all physicians combined (0.96) and for chief and attending physicians in particular (0.96 and 0.98, respectively; Table 4). Fellows exhibited a lower rate of adherence (0.93), which was because of lower rates of adherence for HR-positive HER2/neu-positive tumors. Logistic regression modeling further demonstrated that concordance among all physicians was significantly higher for stage IV first-line therapy tumors (odds ratio, 4.88; 1.22 to 19.51; $P < .01$; Appendix Table A1).

Overall adherence of the AI advisory tool to NCCN guidelines was high but significantly below the level of all physicians combined (0.82 *v* 0.96; $P < .001$). Adherence to NCCN guidelines was lowest for HR-positive HER2/neu-negative stages IIIA/C (0.43 and 0.42, respectively) and HR-positive HER2/neu-positive tumors (0.13 to 0.67 for five of eight stages). These findings were confirmed in odds ratios from multiple logistic regression modeling (Appendix Table A2).

DISCUSSION

The current study was undertaken in complex breast cancers because these patients pose therapeutic challenges for oncologists. Treatment decisions in these cases are challenging because no gold standard therapy exists and disagreement may occur even among experts. Furthermore, treatment decisions are affected by local practice patterns, patient preferences, socioeconomic status, and variability in available therapeutic modalities that results from regulatory approval and economic factors (eg, insurance coverage). In this study, for example, concordance was affected by the uncertainty of optimal treatment of complex cases (eg, HR-positive HER2/neu-negative cancers) and local contextual factors (eg, underuse of multigene or *BRCA* testing as a result of the absence of national

TABLE 1. Sample Variable Means by Physician Experience Level

Variable	TNM	All Physicians (N = 1,977)		Chief Physicians (n = 595)		Attending Physicians (n = 790)		Fellows (n = 592)		*P*
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Age, years		46.92	9.08	46.67	8.98	47.00	8.95	47.07	9.36	.77
Stage										
IA	T1N0M0	0.14	0.35	0.14	0.35	0.15	0.35	0.13	0.34	.86
IIA	T0N1M0, T1N1M0, T2N0M0	0.22	0.41	0.24	0.43	0.19	0.39	0.23	0.42	.29
IIB	T2N1M0, T3N0M0	0.10	0.30	0.10	0.31	0.08	0.27	0.12	0.32	.43
IIIA	T3N1M0, T0N2M0, T1N2M0, T2N2M0, T3N2M0	0.10	0.30	0.12	0.32	0.09	0.29	0.08	0.27	.48
IIIB	T4N0M0, T4N1M0, T4N2M0	0.00	0.05	0.01	0.07	0.00	0.00	0.00	0.06	.99
IIIC	AnyTN3M0	0.05	0.21	0.06	0.24	0.05	0.22	0.03	0.17	.67
IV, first line	AnyTAnyNM1	0.20	0.40	0.25	0.43	0.19	0.39	0.17	0.37	.05
IV, second line	AnyTAnyNM1	0.20	0.40	0.08	0.28	0.25	0.43	0.24	0.43	< .001
Subtype										
HR positive HER2/neu negative		0.46	0.50	0.44	0.50	0.46	0.50	0.47	0.50	.68
HR positive HER2/neu positive		0.19	0.39	0.20	0.40	0.19	0.40	0.18	0.38	.77
HR negative HER2/neu positive		0.16	0.37	0.19	0.39	0.15	0.36	0.16	0.36	.47
TNBC		0.19	0.39	0.17	0.38	0.20	0.40	0.20	0.40	.62

NOTE. Values are proportions unless otherwise noted. *P* values are from Kruskal-Wallis equality of populations test of differences across the physician level.

Abbreviations: HR, hormone receptor; HER2/neu, human epidermal growth factor receptor 2; SD, standard deviation; TNBC, triple-negative breast cancer (HR and HER2/neu negative-tumors); WfO = Watson for Oncology.

TABLE 2. Concordance Means by Physician Experience Level

	All Physician–WfO Concordance			Chief Physician–WfO Concordance			Attending Physician–WfO Concordance			Fellow–WfO Concordance		
Variable	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***
Overall	0.55	0.50										
By prescriber level												
Chief	0.53	0.50	Ref									
Attending	0.48	0.50	.03									
Fellow	0.68	0.47	< .001									
Within subtype: HR positive HER2 negative												
By stage												
IA	0.54	0.50	Ref	0.57	0.50	Ref	0.45	0.51	Ref	0.61	0.50	Ref
IIA	0.49	0.50	.41	0.41	0.50	.16	0.23	0.43	.03	0.81	0.39	.04
IIB	0.49	0.50	.49	0.31	0.47	.05	0.33	0.48	.38	0.77	0.43	.17
IIIA	0.41	0.49	.07	0.52	0.51	.67	0.26	0.44	.09	0.54	0.51	.64
IIIB	1.00	0.00	.20	1.00	0.00	.40	—	—	—	1.00	0.00	.43
IIIC	0.42	0.50	.16	0.55	0.51	.88	0.35	0.49	.44	0.36	0.50	.18
IV, first line	0.67	0.47	.05	0.53	0.50	.70	0.84	0.37	< .001	0.63	0.49	.82
IV, second line	0.58	0.50	.56	0.74	0.45	.22	0.38	0.49	.53	0.79	0.41	.07
Within subtype: HR positive HER2 positive												
By stage												
IA	0.36	0.48	Ref	0.41	0.51	Ref	0.25	0.44	Ref	0.56	0.51	Ref
IIA	0.58	0.50	.01	0.48	0.51	.70	0.40	0.50	.18	0.95	0.22	.005
IIB	0.64	0.48	.003	0.50	0.52	.63	0.42	0.51	.27	0.94	0.25	.01
IIIA	0.56	0.50	.05	0.57	0.51	.38	0.42	0.51	.27	0.75	0.46	.36
IIIB	0.33	0.58	.93	0.00	0.00	.27	—	—	—	1.00	0.00	.39
IIIC	0.30	0.47	.63	0.33	0.50	.70	0.10	0.32	.31	0.75	0.50	.48
IV, first line	0.48	0.50	.14	0.31	0.47	.49	0.68	0.48	< .001	0.48	0.51	.63
IV, second line	0.29	0.46	.44	0.36	0.50	.80	0.42	0.50	.17	0.06	0.25	.003
Within subtype: HR negative HER2 positive												
By stage												
IA	0.24	0.43	Ref	0.10	0.31	Ref	0.41	0.51	Ref	0.22	0.44	Ref
IIA	0.84	0.37	< .001	0.90	0.31	< .001	0.72	0.46	.05	0.95	0.23	< .001
IIB	0.97	0.17	< .001	0.91	0.30	< .001	1.00	0.00	.003	1.00	0.00	< .001
IIIA	0.97	0.17	< .001	0.94	0.25	< .001	1.00	0.00	.001	1.00	0.00	.007
IIIB	—	—	—	—	—	—	—	—	—	—	—	—
IIIC	0.80	0.42	< .001	0.80	0.45	.001	0.75	0.50	.23	1.00	0.00	.13
IV, first line	0.36	0.48	.17	0.45	0.51	.01	0.24	0.44	.24	0.36	0.49	.45
IV, second line	0.24	0.43	.98	0.25	0.46	.31	0.41	0.50	.98	0.04	0.21	.12
Within subtype: TNBC												
By stage												
IA	0.57	0.50	Ref	0.65	0.49	Ref	0.43	0.50	Ref	0.68	0.48	Ref
IIA	0.92	0.27	< .001	0.86	0.35	.07	0.93	0.26	< .001	1.00	0.00	.002
IIB	0.94	0.24	< .001	0.82	0.40	.33	1.00	0.00	< .001	1.00	0.00	.05
IIIA	0.96	0.19	< .001	1.00	0.00	.04	1.00	0.00	.008	0.90	0.32	.19

(Continued on following page)

TABLE 2. Concordance Means by Physician Experience Level (Continued)

Variable	All Physician–WfO Concordance			Chief Physician–WfO Concordance			Attending Physician–WfO Concordance			Fellow–WfO Concordance		
	Mean	SD	*P*	Mean	SD	*P*	Mean	SD	*P*	Mean	SD	*P*
IIIB	—	—	—	—	—	—	—	—	—	—	—	—
IIIC	1.00	0.00	.09	1.00	0.00	.48	1.00	0.00	.27	1.00	0.00	.35
IV, first line	0.43	0.50	.10	0.27	0.46	.03	0.64	0.49	.11	0.13	0.35	.001
IV, second line	0.52	0.50	.52	0.38	0.52	.19	0.36	0.48	.50	0.80	0.41	.34

NOTE. Values are proportions. *P* values are from Wilcoxon signed rank test of differences in concordance between indicated and reference groups.
Abbreviations: HR, hormone receptor; HER2/neu, human epidermal growth factor receptor 2; Ref, reference group; TNBC, triple-negative breast cancer (HR and HER2/neu-negative tumors); WfO, Watson for Oncology.

standards or private insurance coverage needs for trastazumab therapy for patients with HER2/neu-positive cancer).[13] Higher rates of concordance can be obtained by modifying AI advisory tools to address local contextual factors and by advances in knowledge that reduce therapeutic uncertainties.

Nonetheless, a moderate level of overall concordance was found between the treatment advisory tool and the three groups of oncologists. Concordance rates varied by level of clinician expertise, with oncologists in training (fellows) exhibiting a slightly higher rate of concordance with the tool's recommendations than senior oncologists. This difference could not be explained by differences between physician groups in patient age, receptor subtype, or TNM stage, as shown by logistic regression modeling results, which were adjusted for these variables.

Although this study was not designed to examine the basis for clinician treatment choices, reasoning processes in clinical decision making have been demonstrated to vary by level of clinical experience and to evolve over time, contributing to differences in judgements involving the same complex cases.[14,15] This transition occurs as a result of many factors, including cognitive shifts in the balance between the use of causal biomedical, hypothesis-generating, deductive reasoning in physicians in training and use of broad knowledge bases, pattern recognition, and illness scripts in experts—that is, compilations of reorganized pathophysiologic, clinical, and contextual scenarios.[16,17]

Concordance also differed by HR subtype–TNM stage, with HR-positive and metastatic breast cancers being lower than that of HR-negative and nonmetastatic TNBC cancers. This is a result of a larger number of adjuvant endocrine treatment choices for HR-positive early and HR-positive metastatic breast cancers, which leads to a lower probability of agreement between two experts. Treatment

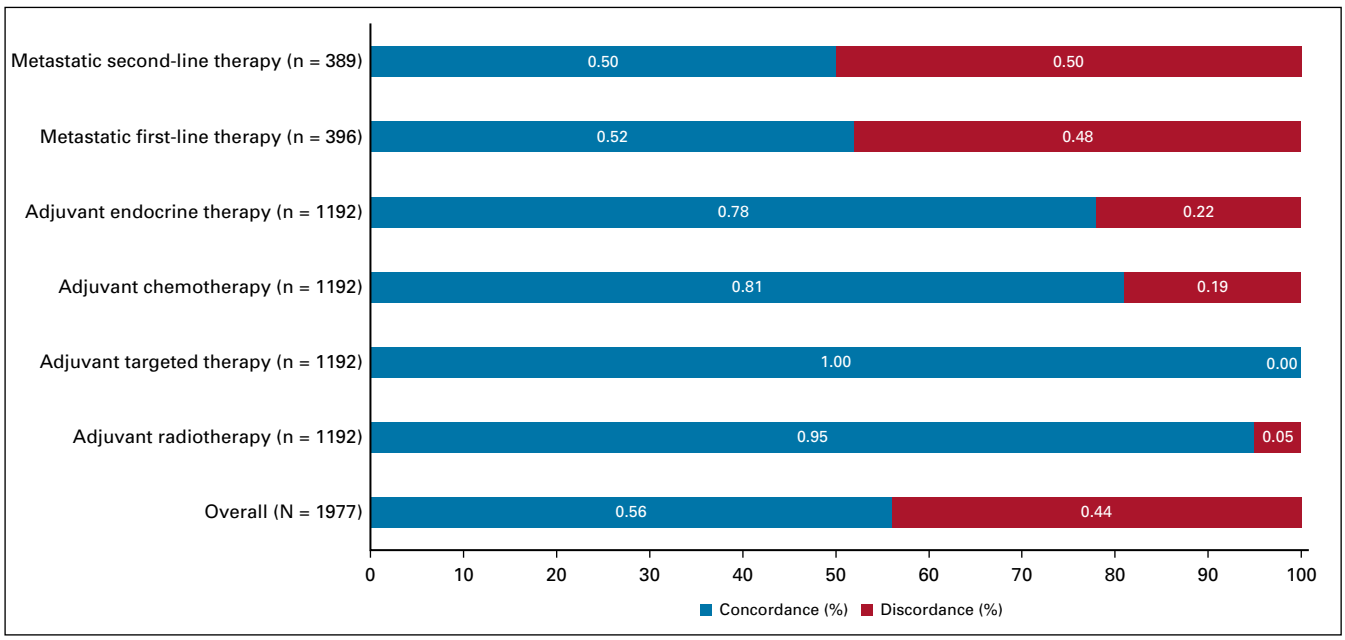

FIG 1. Concordance by adjuvant therapy or metastatic breast cancer therapy type.

TABLE 3. Reasons For Nonconcordance by Physician Experience Level

Reason	All Physicians (N = 881)	Chief Physicians (n = 277)	Attending Physicians (n = 385)	Fellows (n = 219)	*P*
Clinical judgment at variance with WfO	0.48	0.51	0.47	0.45	.43
WfO not consistent with NCCN	0.33	0.32	0.33	0.35	.78
Treatment not available in WfO	0.08	0.08	0.09	0.06	.42
Treatment not available in China	0.11	0.09	0.11	0.14	.21

NOTE. *P* values are from Kruskal-Wallis equality of populations test of difference in the presence of reason for nonconcordance across physician level.
Abbreviations: NCCN, National Comprehensive Cancer Network; WfO, Watson for Oncology.

decisions involving stage IV cancers exhibited the lowest rates of concordance across all HR subtypes. This was expected because of their greater biologic, genomic, and clinical complexity, as well as the limits of current evidence to guide patient-specific treatment.[7] For example, clinical judgments about responses to prior therapies and the degree of risk for adverse systemic and organ effects may differ among physicians, which leads to greater nonconcordance for treatments. Stage-dependent concordance on breast cancer treatment using the WfO advisory tool was also observed in a general population of breast cancer cases in India.[5]

It is important to note that nonconcordance does not imply that one treatment is correct for a given patient and another is not, nor does it necessarily diminish the potential value of a decision support system that provides access to supporting evidence and insight into its reasoning process.[18] A system that provides this information at the point of care may promote the consideration of alternate approaches, allow for comparison of the system's and clinician's logic well, and enable exam room reviews of relevant evidence. This degree of disclosure should be provided by decision support systems for treatment recommendations that are offered as well as for treatments that are expressly Not recommended.

This study provides insight into NCCN guideline adherence for complex breast cancer cases among Chinese oncologists with different levels of experience. Guidelines are tools that use rigorous evidence evaluation to inform care decisions by practitioners and patients, but not to dictate care for individual patients.[19] Adherence to guidelines will be influenced by factors at many levels: the patient, practitioner, practice, institution, and the broader ecosystem.[20] In the current work, institutional- and practice-level emphasis on guidelines accounts for the high level of adherence among physicians, and the senior oncologists' greater experience applying them likely explains their higher level of adherence compared with fellows.[16,21] The lower level of adherence to NCCN guidelines by the advisory tool was expected as the WfO treatment advisor is not driven solely by these guidelines. The WfO advisory tool also includes MSK treatment guidelines and current medical evidence from an extensive array of sources that are curated by MSK breast cancer experts. Concordance between the tool and any clinician group focused on achieving high rates of NCCN guideline adherence would therefore be diminished. For example, whereas the tool recommended tamoxifen for patients with HR-positive stage III cancers, NCCN guidelines recommended ovarian function suppression along with either tamoxifen or an aromatase inhibitor. In the case of patients with HER2/neu-positive disease, the advisory tool would often recommend a nonguideline treatment—vinorelbine plus pertuzumab plus trastuzumab—reflecting a preference at MSK.

This study possesses some important strengths as well as several notable limitations. Strengths include the large number of cases reviewed, use of physicians with different levels of expertise, and blindness to treatment recommendations provided by the advisory tool and other physicians. By including only patients with breast cancers that were high risk for recurrence and metastatic breast cancers, this study also allowed for a focused examination of the decision support system's recommendations for complex cases in which physicians may desire additional support. Finally, this report provides insight into adherence to the internationally recognized NCCN breast cancer treatment guidelines by oncologists with varying levels of expertise as well as by an advisory tool using NCCN guidelines and expertise from a leading cancer institution to provide treatment recommendations.

The current investigation contains important limitations. First, the study design is cross-sectional in nature. Increases in the experience of treating physicians over time and improvements in decision support technology may produce different results in subsequent time periods. Second, the study used oncologists from top-tier hospitals in China and their numbers at each level of experience were small. The time required by active practitioners with large patient populations to review large numbers of cases limited the ability to obtain more physician prescribers. Third, physician prescribers reviewed different cases rather than an identical set of cases. This prevented direct comparisons of concordance between physician pairs; however, concordance for each physician prescriber was determined by comparing recommendations for an identical set of cases from each clinician with those of the WfO advisory tool.

TABLE 4. NCCN Guideline Adherence Means by Physician Experience Level

	All Physician–NCCN Adherence			Chief Physician–NCCN Adherence			Attending Physician–NCCN Adherence			Fellow–NCCN Adherence			WfO–NCCN Adherence*		
Variable	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***
Overall	0.96	0.20											0.81	0.39	
By prescriber level															
Chief physician	0.96	0.20	Ref										0.82	(0.39)	Ref
Attending physician	0.98	0.15	.04										0.81	(0.39)	.65
Fellow	0.93	0.25	.04										0.81	(0.40)	.57
Within subtype: HR positive HER2 negative															
By stage															
IA	0.92	0.28	Ref	0.93	0.26	Ref	0.90	0.31	Ref	0.93	0.26	Ref	0.89	0.31	Ref
IIA	0.93	0.26	.82	0.97	0.18	.40	0.90	0.30	.34	0.91	0.28	.08	0.88	0.33	.66
IIB	0.92	0.28	.98	0.88	0.33	.58	0.96	0.19	.03	0.90	0.30	.73	0.80	0.40	.08
IIIA	0.96	0.20	.24	0.97	0.18	.50	1.00	0.00	.03	0.88	0.34	.52	0.43	0.50	< .001
IIIB	1.00	0.00	.67	1.00	0.00	.79	—	—	—	1.00	0.00	.79	1.00	0.00	.63
IIIC	0.96	0.19	.25	0.96	0.12	.13	1.00	0.00	.09	0.91	0.03	.84	0.42	0.50	< .001
IV, first line	0.98	0.13	.01	0.99	0.12	.13	1.00	0.00	.01	0.95	0.22	.69	0.99	0.07	< .001
IV, second line	1.00	0.00	< .001	1.00	0.00	.20	1.00	0.00	.001	1.00	0.00	.02	0.92	0.27	.50
Within subtype: HR positive HER2 negative															
By stage															
IA	0.96	0.20	Ref	0.94	0.24	Ref	1.00	0.00	Ref	0.89	0.32	Ref	0.92	0.27	Ref
IIA	0.95	0.23	.67	0.96	0.21	.83	0.97	0.18	.25	0.90	0.31	.91	0.79	0.41	.03
IIB	0.86	0.35	.05	1.00	0.00	.36	1.00	0.00	—	0.63	0.50	.07	0.62	0.49	< .001
IIIA	0.91	0.29	.31	1.00	0.00	.36	0.92	0.29	.07	0.75	0.46	.37	0.56	0.50	< .001
IIIB	0.67	0.58	.02	1.00	0.00	.73	—	—	—	0.00	0.00	.02	0.67	0.58	.13
IIIC	0.83	0.39	.03	0.89	0.33	.64	1.00	0.00	—	0.25	0.50	.01	0.13	0.34	< .001
IV, first line	0.97	0.16	.65	1.00	0.00	.19	1.00	0.00	—	0.90	0.30	.87	0.13	0.37	.13
IV, second line	0.90	0.30	.19	0.73	0.47	.12	0.92	0.28	.07	1.00	0.00	.18	0.67	0.48	< .001
Within subtype: HR positive HER2															
By stage															
IA	1.00	0.00	Ref	1.00	0.00	Ref	1.00	0.00	Ref	1.00	0.00	Ref	0.93	0.25	Ref
IIA	1.00	0.00	—	1.00	0.00		1.00	0.00	—	1.00	0.00	—	0.97	0.18	.40
IIB	1.00	0.00	—	1.00	0.00		1.00	0.00	—	1.00	0.00	—	0.97	0.17	.47

(Continued on following page)

TABLE 4. NCCN Guideline Adherence Means by Physician Experience Level (Continued)

	All Physician–NCCN Adherence			Chief Physician–NCCN Adherence			Attending Physician–NCCN Adherence			Fellow–NCCN Adherence			WfO–NCCN Adherence*		
Variable	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***	**Mean**	**SD**	***P***
IIIA	1.00	0.00	—	1.00	0.00		1.00	0.00	—	1.00	0.00	—	1.00	0.00	.14
IIIB	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
IIIC	1.00	0.00	—	1.00	0.00		1.00	0.00	—	1.00	0.00	—	1.00	0.00	.41
IV, first line	0.92	0.27	.05	0.87	0.34	.10	1.00	0.00	—	0.91	0.00	—	1.00	0.00	< .001
IV, second line	0.90	0.31	.03	0.75	0.46	.02	0.93	0.27	.26	0.91	0.27	.26	0.34	0.48	< .001
Within subtype: TNBC															
By stage															
IA	0.99	0.00	Ref	1.00	0.00	Ref	1.00	0.00	Ref	1.00	0.00	Ref	1.00	0.00	Ref
IIA	1.00	0.00	.26	1.00	0.00	—	1.00	0.00		1.00	0.00	.27	1.00	0.00	< .001
IIB	1.00	0.00	.49	1.00	0.00	—	1.00	0.00	—	1.00	0.00	.50	1.00	0.00	.02
IIIA	1.00	0.00	.54	1.00	0.00	—	1.00	0.00	—	1.00	0.00	.50	1.00	0.00	.03
IIIB	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
IIIC	1.00	0.00	.81	1.00			1.00	0.00	—	1.00	0.00	.76	1.00	0.00	.40
IV, first line	0.98	0.13	.92	0.93	0.26	.25	1.00	0.00	—	1.00	0.00	.41	0.97	0.00	.02
IV, second line	0.99	0.11	.92	0.88	0.35	.11	1.00	0.00	—	1.00	0.00	.24	0.81	0.40	.51

NOTE. Values are proportions. *P* values are from Wilcoxon signed rank test of differences in adherence between indicated and reference groups.

Abbreviations: HR, hormone receptor; HER2/neu, human epidermal growth factor receptor 2; NCCN, National Comprehensive Cancer Network; Ref, reference group; TNBC, triple-negative breast cancer (HR and HER2/neu-negative tumors); WfO, Watson for Oncology.

**P* values from Wilcoxon signed rank tests of differences in adherence between all physicians and WfO, both overall and by variable subgroups, are < .003

In conclusion, this study demonstrates that, at the level of individual oncologists, treatment concordance for complex breast cancers was modest between physicians and the AI advisory tool. This degree of concordance is encouraging because therapeutic decisions in these cases are often difficult as a result of the current limits of medical knowledge for treating complex breast cancers and the presence of local contextual factors that affect physician treatment choices. Ongoing advances in breast cancer research and the customization of these tools to account for local contextual factors will improve the performance of this technology. The current study is an early step in a chain of evidence for AI advisory tools that requires replication in other geographies and practice settings to understand its scalability. Of more importance, research must proceed to assess the technology's impact on clinician–patient treatment decisions as well as patient health and health system outcomes in rigorously controlled trials.

AFFILIATIONS

[1]Academy of Military Medical Sciences, Beijing, People's Republic of China
[2]IBM Research, Yorktown Heights, NY
[3]The Fifth Medical Center of Chinese People's Liberation Army General Hospital, Beijing, People's Republic of China
[4]The Affiliated Hospital of Qingdao University, Qingdao, People's Republic of China
[5]Medical Molecular Biology, Beijing Institute of Biotechnology, Beijing, People's Republic of China
[6]Jiangsu Province Hospital, Nanjing, People's Republic of China
[7]Henan Cancer Hospital, Zhengzhou, People's Republic of China
[8]RxEconomics, Hunt Valley, MD
[9]Columbia University, New York, NY
[10]Fourth Hospital of Hebei Medical University, Shijiazhuang, People's Republic of China
[11]IBM, Cambridge, MA

CORRESPONDING AUTHOR

Zefei Jiang, MD, No. 8 of East Street, Fengtai District, Beijing, People's Republic of China; 86-13901372170; e-mail: jiangzefei@csco.org.cn.

AUTHOR CONTRIBUTIONS

Conception and design: Fengrui Xu, Zefei Jiang, Yongmei Yin, Edward H. Shortliffe
Financial support: Zefei Jiang
Administrative support: Zefei Jiang, Zhenzhen Liu, Min Yan, Jinhai Tang
Provision of study material or patients: Fengrui Xu, Zefei Jiang, Haibo Wang, Zhenzhen Liu, Min Yan, Yuhua Song, Cuizhi Geng
Collection and assembly of data: Fengrui Xu, Zefei Jiang, Haibo Wang, Zhenzhen Liu, Min Yan, Yuhua Song, Cuizhi Geng, Jinhai Tang
Data analysis and interpretation: Fengrui Xu, Martín-Jose Sepúlveda, Zefei Jiang, Haibo Wang, Jianbin Li, M. Christopher Roebuck, Edward H. Shortliffe, Cuizhi Geng, Kyu Rhee
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Martín-José Sepúlveda
Consulting or Advisory Role: IBM Watson Health

M. Christopher Roebuck
Consulting or Advisory Role: Pharmaceutical Manufacturer Association (Inst), IBM Watson Health (Inst)

Edward H. Shortliffe
Consulting or Advisory Role: IBM Watson Health
Patents, Royalties, Other Intellectual Property: Editor of textbook on biomedical informatics (Springer), now in its fourth edition (fifth in preparation)

Kyu Rhee
Employment: IBM
Leadership: IBM
Stock and Other Ownership Interests: CVS Health (I), Johnson & Johnson (I), Merck (I), Celgene (I), Allergan (I), Eli Lilly (I)

No other potential conflicts of interest were reported.

REFERENCES

1. Buchanan BG, Shortliffe EH, (eds): An expert system for oncology protocol management, in Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, 876-881. Reading, MA, Addison-Wesley, 1984. http://www.shortliffe.net/Buchanan-Shortliffe-1984/Chapter-35.pdf
2. Lieber J, Bresson B: Case-based reasoning for breast cancer treatment decision helping, in Blanzieri E, Portinale L (eds): Advances in Case-Based Reasoning. EWCBR 2000. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1898. Heidelberg, Germany, Springer, 2000, pp 173-185
3. Lisboa PJ, Taktak AFG: The use of artificial neural networks in decision support in cancer: A systematic review. Neural Netw 19:408-415, 2006
4. Suwanrusme H, Issarachai S, Umsawasdi T, et al: Concordance assessment of a clinical decision support software in patients with solid tumors. J Clin Oncol 36, 2018 (suppl; abstr 18584)
5. Somashekhar SP, Sepúlveda M-J, Puglielli S, et al: Watson for Oncology and breast cancer treatment recommendations: Agreement with an expert multidisciplinary tumor board. Ann Oncol 29:418-423, 2018
6. Chang JH, Vines E, Bertsch H, et al: The impact of a multidisciplinary breast cancer center on recommendations for patient management: The University of Pennsylvania experience. Cancer 91:1231-1237, 2001
7. National Comprehensive Cancer Network: Clinical Practice Guidelines in Oncology: Breast cancer, version 1.2018. https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

8. Martei YM, Matro JM: Identifying patients at high risk of breast cancer recurrence: Strategies to improve patient outcomes. Breast Cancer (Dove Med Press) 7:337-343, 2015
9. Medrio Corporation: EDC: Electronic data capture system. https://medrio.com/products/edc/
10. Somashekhar SP, Sepúlveda M-J, Puglielli S, et al: Watson for Oncology and breast cancer treatment recommendations: Agreement with an expert multidisciplinary tumor board. Annals Oncol 29:418-423, 2018. https://academic.oup.com/annonc/article/29/2/418/4781689#supplementary-data
11. Kruskal WH, Wallis WA: Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47:583-621, 1952
12. Wilcoxon F: Individual comparisons by ranking methods. Biometrics 1:80-83, 1945
13. Li J, Shao Z, Xu B, Jiang Z, et al: Use of trastuzumab as an adjuvant/neoadjuvant therapy in patients with HER2-positive breast cancer in China: The Nvwa study. Medicine (Baltimore) 97:e10350,2018
14. Dowie J, Elstein AS: Professional Judgement: A Reader in Clinical Decision Making. New York, NY, Cambridge University Press, 1988
15. Boshuizen HPA, Schmidt HG: On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. Cogn Sci 16:153-184, 1992
16. Schmidt HG, Norman GR, Boshuizen HPA: A cognitive perspective on medical expertise: Theory and implication. Acad Med 65:611-621, 1990
17. Patel VL, Frederiksen CH: Cognitive processes in comprehension and knowledge acquisition by medical students and physicians, in Schmidt HG, de Volder MC (eds): Tutorials in Problem-Based Learning. Assen, the Netherlands, van Gorcum, 1984, pp 143-157
18. Shortliffe EH, Sepúlveda M-J: Clinical decision-support in the era of artificial intelligence. JAMA 320:2199-2200, 2018. https://jamanetwork.com/journals/jama/fullarticle/2713901
19. National Institutes of Health: Clinical practice guidelines. https://nccih.nih.gov/health/providers/clinicalpractice.htm
20. Cabana MD, Rand CS, Powe NR, et al: Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 282:1458-1465, 1999
21. Patel VL, Arocha JF, Diermeier M, et al: Cognitive psychological studies of representation and use of clinical practice guidelines. Int J Med Inform 63:147-167, 2001

TABLE A1. Odds Ratios (95% CI) From Logistic Regressions of Physician–WfO Concordance

Variable	All Physician Model (N = 1,977)	Chief Physician Model (n = 595)	Attending Physician (n = 790)	Fellow Model (n = 592)
Chief physician (Ref)	1 (1.00 to 1.00)			
Attending physician	0.88 (0.69 to 1.13)			
Fellow	2.25* (1.72 to 2.93)			
Age	1.04* (1.03 to 1.05)	1.05* (1.02 to 1.07)	1.05* (1.03 to 1.07)	1.02† (1.00 to 1.05)
Stage IA (Ref)	1 (1.00 to 1.00)	1 (1.00 to 1.00)	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IIB	0.76 (0.45 to 1.29)	0.56 (0.22 to 1.40)	0.33† (0.13 to 0.85)	2.6 (0.98 to 6.89)
Stage IIB	0.74 (0.39 to 1.37)	0.32 (0.10 to 1.00)	0.5 (0.17 to 1.52)	2.22 (0.71 to 6.92)
Stage IIIA	0.57 (0.31 to 1.05)	0.69 (0.24 to 1.97)	0.4 (0.15 to 1.13)	0.7 (0.23 to 2.12)
Stage IIIB	0.97 (0.08 to 12.29)	1 (1.00 to 1.00)		1 (1.00 to 1.00)
Stage IIIC	0.67 (0.33 to 1.34)	0.9 (0.27 to 2.93)	0.59 (0.20 to 1.81)	0.38 (0.09 to 1.62)
Stage IV, first line	2.09* (1.21 to 3.60)	0.99 (0.40 to 2.43)	7.15* (2.65 to 19.35)	1.21 (0.45 to 3.29)
Stage IV, second line	1.18 (0.70 to 2.01)	2.77 (0.82 to 9.36)	0.74 (0.31 to 1.73)	2.38 (0.93 to 6.11)
HR positive HER2 negative (Ref)	1 (1.00 to 1.00)	1 (1.00 to 1.00)	1 (1.00 to 1.00)	1 (1.00 to 1.00)
HR positive HER2 positive	0.49† (0.25 to 0.94)	0.48 (0.14 to 1.65)	0.38 (0.13 to 1.07)	0.8 (0.24 to 2.67)
HR negative HER2 positive	0.25* (0.11 to 0.57)	0.08* (0.01 to 0.40)	0.7 (0.20 to 2.42)	0.16† (0.03 to 0.95)
TNBC	1.03 (0.54 to 1.99)	1.41 (0.42 to 4.69)	0.76 (0.27 to 2.18)	1.22 (0.37 to 4.00)
Interaction effects				
Stage IIA × HR positive HER2 positive	3.28* (1.39 to 7.75)	2.78 (0.57 to 13.46)	6.27† (1.54 to 25.46)	5.78 (0.51 to 65.31)
Stage IIA × HR negative HER2 positive	24.14* (7.97 to 73.10)	148.18* (15.29 to 1,436.49)	12.53* (2.45 to 64.13)	25.54† (1.65 to 395.90)
Stage IIA × TNBC	13.67* (4.75 to 39.33)	6.50† (1.29 to 32.73)	58.80† (8.96 to 385.60)	
Stage IIB × HR positive HER2 positive	4.47* (1.61 to 12.45)	5.81 (0.91 to 36.99)	4.57 (0.78 to 26.86)	5.8 (0.47 to 71.17)
Stage IIB × HR negative HER2 positive	132.92* (14.70 to 1202.07)	276.37* (17.20 to 4,441.92)		
Stage IIB × TNBC	18.54* (3.60 to 95.60)	8.11 (0.96 to 68.32)		
Stage IIIA × HR positive HER2 positive	4.21* (1.48 to 11.96)	2.82 (0.47 to 16.97)	6.85† (1.22 to 38.28)	3.41 (0.39 to 29.63)
Stage IIIA × HR negative HER2 positive	221.32* (24.52 to 1,997.93)	238.93* (15.59 to 3,661.15)		
Stage IIIA × TNBC	37.14* (4.32 to 319.12)			7.16 (0.57 to 90.06)

(Continued on following page)

TABLE A1. Odds Ratios (95% CI) From Logistic Regressions of Physician–WfO Concordance (Continued)

Variable	All Physician Model (N = 1,977)	Chief Physician Model (n = 595)	Attending Physician (n = 790)	Fellow Model (n = 592)
Stage IIIC × HR positive HER2 positive	1.34 (0.39 to 4.64)	0.99 (0.12 to 7.90)	0.63 (0.05 to 7.44)	6.48 (0.37 to 112.11)
Stage IIIC × HR negative HER2 positive	25.40* (4.02 to 160.45)	45.15† (2.45 to 831.32)	8.49 (0.56 to 128.69)	
Stage IIIC × TNBC	1 (1.00 to 1.00)	1 (1.00 to 1.00)	1 (1.00 to 1.00)	1 1.00 to 1.00)
Stage IV, first line × HR positive HER2 positive	0.83 (0.35 to 1.97)	0.74 (0.16 to 3.15)	1.05 (0.23 to 4.70)	0.64 (0.13 to 3.21)
Stage IV, first line × HR negative HER2 positive	0.94 (0.35 to 2.57)	9.12† (1.41 to 58.89)	0.07* (0.01 to 0.38)	1.93 (0.25 to 15.16)
Stage IV, first line × TNBC	0.34† (0.14 to 0.83)	0.2 (0.04 to 1.17)	0.45 (0.11 to 1.89)	0.07† (0.01 to 0.51)
Stage IV, second line × HR positive HER2 positive	0.63 (0.24 to 1.60)	0.37 (0.05 to 2.70)	3.38 (0.84 to 13.63)	0.02* (0.00 to 0.25)
Stage IV, second line × HR negative HER2 positive	0.7 (0.24 to 2.05)	1.29 (0.11 to 15.75)	1.4 (0.31 to 6.43)	0.06† (0.00 to 0.96)
Stage IV, second line × TNBC	0.74 (0.32 to 1.72)	0.14 (0.02 to 1.19)	1.05 (0.29 to 3.83)	0.92 (0.19 to 4.52)
Constant	0.17* (0.09 to 0.35)	0.16* (0.04 to 0.56)	0.09* (0.03 to 0.27)	0.52 (0.14 to 1.96)

NOTE. Reported as odds ratio (95% CI).

Abbreviations: HR, hormone receptor; HER2, human epidermal growth factor receptor 2; Ref, reference group; TNBC, triple-negative breast cancer; WfO, Watson for Oncology.

*$P < .01$.

†$P < .05$.

TABLE A2. Odds Ratios (95% CI) From Logistic Regressions of Adherence

Variable	All Physicians–NCCN Model (N = 1,977)	WfO–NCCN Model (N = 1,977)
Chief physician (Ref)	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Attending physician	1.77 (0.92 to 3.38)	0.97 (0.70 to 1.36)
Fellow	0.55* (0.32 to 0.95)	0.88 (0.62 to 1.25)
Age	1 (0.97 to 1.02)	1.05† (1.03 to 1.06)
Stage IA (Ref)	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IIA	1.13 (0.44 to 2.90)	0.8 (0.35 to 1.81)
Stage IIB	1.03 (0.34 to 3.09)	0.43 (0.18 to 1.05)
Stage IIIA	1.93 (0.54 to 6.90)	0.08† (0.03 to 0.17)
Stage IIIB	0.11 (0.01 to 1.59)	0.22 (0.02 to 2.92)
Stage IIIC	2.13 (0.42 to 10.77)	0.08† (0.03 to 0.19)
Stage IV, first line	4.88* (1.22 to 19.51)	23.95† (2.97 to 192.96)
Stage IV, second line	1.13 (0.07 to 18.60)	1.41 (0.59 to 3.34)
HR positive HER2 negative (Ref)	1 (1.00 to 1.00)	1 (1.00 to 1.00)
HR positive HER2 positive	1.87 (0.46 to 7.59)	0.3 (0.44 to 3.87)
HR negative HER2 positive	0.69 (0.03 to 14.23)	1.46 (0.37 to 5.74)
TNBC	6.22 (0.74 to 52.19)	0.56 (0.22 to 1.45)
Interaction effects		
Stage IIA × HR positive HER2 positive	0.69 (0.11 to 4.20)	0.42 (0.12 to 1.56)
Stage IIA × HR negative HER2 positive	1 (1.00 to 1.00)	2.79 (0.37 to 20.82)
Stage IIA × TNBC	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IIB × HR positive HER2 positive	0.29 (0.05 to 1.82)	0.36 (0.09 to 1.42)
Stage IIB × HR negative HER2 positive	1 (1.00 to 1.00)	4.95 (0.42 to 58.83)
Stage IIB × TNBC	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IIIA × HR positive HER2 positive	0.24 (0.03 to 1.95)	1.45 (0.37 to 5.63)
Stage IIIA × HR negative HER2 positive	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IIIA × TNBC	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IIIC × HR positive HER2 positive	0.09* (0.01 to 0.86)	0.17* (0.03 to 0.97)
Stage IIIC × HR negative HER2 positive	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IIIC × TNBC	1 (1.00 to 1.00)	1 (1.00 to 1.00)
Stage IV, first line × HR positive HER2 positive	0.35 (0.04 to 3.47)	0.02† (0.00 to 0.22)
Stage IV, first line × HR negative HER2 positive	0.31 (0.01 to 8.89)	0.00† (0.00 to 0.03)
Stage IV, first line × TNBC	0.16 (0.01 to 3.63)	0.3 (0.02 to 4.06)
Stage IV, second line × HR positive HER2 positive	0.36 (0.02 to 8.60)	0.13† (0.04 to 0.52)
Stage IV, second line × HR negative HER2 positive	1 (1.00 to 1.00)	0.02† (0.01 to 0.12)
Stage IV, second line × TNBC	1 (1.00 to 1.00)	0.64 (0.19 to 2.14)
Constant	13.72† (3.13 to 60.22)	1.19 (0.44 to 3.20)

NOTE. Reported are odds ratios with 95% CIs in brackets.

Abbreviations: HR, hormone receptor; HER2, human epidermal growth factor receptor 2; NCCN, National Comprehensive Cancer Network; Ref, reference group; TNBC, triple-negative breast cancer; WfO, Watson for Oncology.

*$P < .05$.

†$P < .01$.

TABLE A3. Concordance by Adjuvant Therapy or Metastatic Breast Cancer Therapy Type: Examples

Adjuvant Therapy Type	Physician/NCCN *v* WfO	WfO/NCCN *v* Physician	Reasoning For WfO/NCCN *v* Physician Discordance
Endocrine therapy	Premenopausal HR positive, more than four positive nodes or age < 40 years, physician/NCCN OFS + endocrine therapy *v* WfO TAM	Premenopausal HR positive N0 and age ≥ 40 years, WfO/NCCN TAM *v* physician OFS + AI	Physician felt OFS + AI suitable for premenopausal patients not at high risk of recurrence *v* WfO/NCCN where OFS + AI recommend OFS + AI for premenopausal patients with high risk of recurrence
Chemotherapy	TNBC T1N0M0, physician/NCCN TC or AC-T *v* WfO CMF	HR positive HER2/neu negative T1N1M0, WfO/NCCN recommend adjuvant chemotherapy *v* physician did not	Physician felt that with one nodal metastasis and Ki-67 < 15%, patient was exempt from chemotherapy
Targeted therapy	All adjuvant targeted therapy recommended by physician/NCCN was in line with WfO	HER2/neu positive T1cN0M0, WfO/NCCN chemotherapy + targeted therapy (trastazumab) *v* physician chemotherapy alone	Trastazumab not covered by Chinese Medicare at the time, so physician did not recommend T1c
Radiotherapy	A T3N0M0 patient, physician/NCCN recommend adjuvant radiotherapy *v* WfO did not	A 60-year-old patient with N2, WfO/NCCN recommend adjuvant radiotherapy *v* physician did not	Physician considered patient comorbidities and elected not to include radiotherapy
Metastatic first- and second-line therapies	Adjuvant therapy failure with taxanes and trastazumab, physician/NCCN LX *v* WfO NPH	TNBC AC-T treatment failure, WfO/NCCN capecitabine or gemcitabine *v* physician TX	Patient experienced immediate recurrence after completion of AC-T, physician felt T had not yet taken full effect and so chose TX (docetaxel + capecitabine)

Abbreviations: AC-T, doxorubicin + cyclophosphamide + taxane; AI, artificial intelligence; HR, hormone receptor; HER2, human epidermal growth factor receptor 2; NCCN, National Comprehensive Cancer Network; NPH, vinorelbine + pertuzumab + trastuzumab; OFS, ovarian function suppression; T, taxane; TAM, tamoxifen; TC, taxane + cyclophosphamide; TNBC, triple-negative breast cancer; TX, taxane + xeloda; WfO, Watson for Oncology.