ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

Annals of Oncology 29: 418–423, 2018
doi:10.1093/annonc/mdx781
Published online 9 January 2018

ORIGINAL ARTICLE

Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board

S. P. Somashekhar[1]*, M.-J. Sepúlveda[2], S. Puglielli[3], A. D. Norden[3], E. H. Shortliffe[4], C. Rohit Kumar[1], A. Rauthan[1], N. Arun Kumar[1], P. Patil[1], K. Rhee[3] & Y. Ramya[1]

[1]Manipal Comprehensive Cancer Centre, Manipal Hospital, Bangalore, India; [2]IBM Research (Retired), Yorktown Heights; [3]Watson Health, IBM Corporation, Cambridge; [4]Department of Surgical Oncology, College of Health Solutions, Arizona State University, Phoenix, USA

*Correspondence to: Prof. Sampige Prasannakumar Somashekhar, Manipal Comprehensive Cancer Centre, Manipal Hospital, Old Airport Road, Bangalore 560017, Karnataka, India. Tel: +91-9845712012; Fax: +91-80-2502-3759; E-mail: somashekhar.sp@manipalhospitals.com

Background: Breast cancer oncologists are challenged to personalize care with rapidly changing scientific evidence, drug approvals, and treatment guidelines. Artificial intelligence (AI) clinical decision-support systems (CDSSs) have the potential to help address this challenge. We report here the results of examining the level of agreement (concordance) between treatment recommendations made by the AI CDSS Watson for Oncology (WFO) and a multidisciplinary tumor board for breast cancer.

Patients and methods: Treatment recommendations were provided for 638 breast cancers between 2014 and 2016 at the Manipal Comprehensive Cancer Center, Bengaluru, India. WFO provided treatment recommendations for the identical cases in 2016. A blinded second review was carried out by the center's tumor board in 2016 for all cases in which there was not agreement, to account for treatments and guidelines not available before 2016. Treatment recommendations were considered concordant if the tumor board recommendations were designated 'recommended' or 'for consideration' by WFO.

Results: Treatment concordance between WFO and the multidisciplinary tumor board occurred in 93% of breast cancer cases. Subgroup analysis found that patients with stage I or IV disease were less likely to be concordant than patients with stage II or III disease. Increasing age was found to have a major impact on concordance. Concordance declined significantly ($P \leq 0.02$; $P < 0.001$) in all age groups compared with patients <45 years of age, except for the age group 55–64 years. Receptor status was not found to affect concordance.

Conclusion: Treatment recommendations made by WFO and the tumor board were highly concordant for breast cancer cases examined. Breast cancer stage and patient age had significant influence on concordance, while receptor status alone did not. This study demonstrates that the AI clinical decision-support system WFO may be a helpful tool for breast cancer treatment decision making, especially at centers where expert breast cancer resources are limited.

Key words: Watson for Oncology, artificial intelligence, cognitive clinical decision-support systems, breast cancer, concordance, multidisciplinary tumor board

Introduction

Oncologists who treat breast cancer are challenged by a large and rapidly expanding knowledge base [1, 2]. As of October 2017, for example, there were 69 FDA-approved drugs for the treatment of breast cancer, not including combination treatment regimens [3]. The growth of massive genetic and clinical databases, along with computing systems to exploit them, will accelerate the speed of breast cancer treatment advances and shorten the cycle time for changes to breast cancer treatment guidelines [4, 5]. In addition, these information management challenges in cancer care are occurring in a practice environment where there is little time available for tracking and accessing relevant information at the point of care [6]. For example, a study that surveyed 1117 oncologists reported that on average 4.6 h per week were spent keeping

current in the field; while 53 h per week were spent on patient care and administrative tasks [7].

Clinical decision-support systems (CDSSs) are computer programs that have the potential to help clinicians manage the cognitive demands of these information developments. Many have emerged from earlier work in artificial intelligence (AI) and expert systems [8]. They gather and represent knowledge in ways that allow the computer to simulate human reasoning in order to generate advice. AI systems in health care have generally focused on acquiring knowledge from unstructured data such as text (using natural language processing) or large structured datasets (using machine-learning methods). Having stored, indexed, and curated such knowledge, they then use computational reasoning approaches to apply it to a specific case, generating and evaluating hypotheses and, in turn, offering assessments for clinicians to consider. Several CDSSs for oncology exist such as the American Society for Clinical Oncology's CancerLinq [9] and a system known as OncoDoc [10]. However, a cognitive-support approach for oncology therapy selection has, to our knowledge, not been offered until the introduction of IBM's Watson for Oncology (WFO). It is a unique system, with an ability to acquire much of its knowledge by 'reading' the literature, protocols, and patient charts, and learning from test cases and experts from Memorial Sloan Kettering Cancer Center.

We conducted a study assessing the level of agreement regarding cancer treatment between WFO and the multidisciplinary tumor board from a major comprehensive cancer center in India. The objective of the study was to determine the level of recommended treatment concordance (degree of agreement) in a large population of breast cancer cases. We report herein the results of this assessment and discuss the potential value of the technology as a learning system both for cases where concordance is found and where it is absent.

Methods

Study design and patient population

We studied treatment recommendations for 638 breast cancer cases, comparing the degree of agreement between the AI system WFO (IBM Watson Health, Cambridge, MA) and a 15-member multidisciplinary tumor board at the Manipal Comprehensive Cancer Center (MCCC) in Bangalore, India. Patients were female with ductal, lobular, metaplastic, or mixed histology breast cancer and were either naïve to systemic therapy or experienced disease recurrence after systemic and/or surgical treatment. All patients who presented with breast cancer at MCCC within the 3 years preceding acquisition of the WFO advisor (2014–2016) were included except for patients with colloid, adenocystic, tubular, or secretory breast cancer since WFO was not trained to offer treatment recommendations for these tumor types at the time of study. Patients were also excluded if they were male or had metastatic cancer with disease progression following treatment with systemic therapy (second line and beyond). The study protocol was reviewed and approved by the Manipal Hospital institutional review board.

Watson for Oncology

WFO supports oncologists in making breast cancer treatment decisions using a curated body of knowledge including text from more than 300 medical journals and textbooks, Memorial Sloan Kettering (MSK) Cancer Center treatment guidelines (developed based on national treatment guidelines and the center's clinical experience), and literature hand-selected by MSK experts. WFO has also indexed data from over 550 breast cancer cases, including patient characteristics and comorbidities, functional status, tumor characteristics and stage, imaging, and other laboratory findings. WFO was trained using an iterative process. First, it was taught to analyze breast cancer training cases to provide evidence-based treatment recommendations. After WFO's recommendations were reviewed and scored by MSK experts, the feedback was incorporated into the system and used by WFO to refine its analytical process. Additional details on WFO technology, how it reaches decision, and how it has been evaluated during development, are contained in the supplementary material, available at *Annals of Oncology* online (Development and Training of Watson for Oncology). WFO version 16.4 was used in this study.

WFO's treatment recommendations (which generally include several options for a single case) are categorized into three groups with a corresponding label: green represents 'recommended treatments' with a strong base of evidence, amber represents treatments 'for consideration' that oncologists may consider as suitable alternatives based on their clinical judgment, and red represents treatments that are 'not recommended' due to specific contraindications or strong evidence against their use. Therapies recommended by the tumor board were classified as 'not available' if they were not known to WFO at the time of analysis. Evidence supporting the recommended treatments is provided, as are case-specific clinical trials that may be available, prescribing information, potential adverse reactions related to treatment, and a comparison of treatment options.

Data collection and concordance determination

Patient's data were abstracted from Manipal clinical records and entered manually into WFO by two trained senior oncology fellows. The Manipal multidisciplinary tumor board (MMDT) had previously reviewed and recommended treatment regimens for all cases between 2014 and 2016 (referred to as $T1_{MMDT}$). WFO analyzed the same cases, with their clinical information, in 2016 (referred to as $T2_{WFO}$). WFO and the physicians who ran the cases were blinded to the treatment recommendations that had been made by the MMDT.

Concordance was measured based on how the tumor board's treatment recommendation at $T1_{MMDT}$ was categorized in WFO's list of treatment recommendations at $T2_{WFO}$ ('recommended'/green, 'for consideration'/amber, or 'not recommended'/red; or 'not available' to WFO). If the tumor board's recommendation corresponded to the 'recommended' or 'for consideration' categories, it was defined as concordant. If the tumor board's recommendation was not available in WFO it was designated 'not available', and together with the 'not recommended' category comprised the 'non-concordant' cases.

Since in some cases WFO's analysis was as many as 3 years following the tumor board's initial treatment recommendation, a second analysis was carried out in which non-concordant cases were re-presented to the tumor board to determine whether its recommendation would change in light of new medical advances and guidelines. This second analysis was blinded. The tumor board was unaware of the concordance data, of WFO's treatment recommendation, that they had previously reviewed these cases, or that they were cases being considered in the evaluation of concordance with WFO. Cases that were concordant in the initial analysis were not re-presented for the second analysis due to the amount of time and work it would have required from an already fully engaged tumor board.

Data analysis and statistics

Descriptive statistics of breast cancer case characteristics were calculated using Microsoft Excel and presented as means ± standard deviation or median (min, max). Row percentages are presented where indicated. Concordance was expressed as percent agreement. Cancer characteristics included patient age, cancer stage, and receptor status. To control for

Table 1. Characteristics of breast cancer cases

N	638
Female, %	100%
Age±SD, years	52±11
Stage, *n* (%)	
I	61 (10)
II	262 (41)
III	191 (30)
IV	124 (19)
Receptor status, *n* (%)	
HR (+)	261 (41)
HER2/neu (+)	184 (29)
Triple (−)	193 (30)

HER2/neu, human epidermal growth factor receptor 2; HR, hormone receptor; SD, standard deviation.

these three determinants of concordance simultaneously, a logistic regression model was estimated with odds ratios and 95% confidence intervals reported.

Results

Patients had a mean age of 52 years and presented with non-metastatic disease in 81% of cases (Table 1). Receptor status varied slightly among the cases, with more patients having HR-positive tumors (41%) and fewer patients having HER2/neu-positive (29%) or triple-negative (30%) tumors.

Overall treatment concordance was 73% when no account was taken of the difference in time at which the MMDT originally reviewed the cases (2014–2016, $T1_{MMDT}$) versus when WFO reviewed the identical case information (2016, $T2_{WFO}$; Table 2). Since the MMDT recommendations were made earlier than those by WFO, a blinded, second review of the non-concordant cases (27%, $n = 175$) was conducted in 2016 by the MMDT. This re-review led to a change in treatment recommendations for 128 of 175 (73%) originally non-concordant cases, and improved overall treatment concordance from 73% to 93% (Table 2). Additional results of the second review are presented in supplementary Table S1, available at *Annals of Oncology* online.

Subgroup analyses of treatment concordance by stage and receptor status were also carried out. Concordance varied by stage, with stage I and IV cancers exhibiting lower concordance (80% and 86%, respectively) than stage II and III cancers (97% and 95%, respectively; Figure 1). In contrast, receptor status did not affect concordance, which narrowly ranged from 92% for HR-positive or triple-negative tumors to 95% for HER2/neu-positive tumors (supplementary Figure S1, available at *Annals of Oncology* online). When concordance was analyzed by both stage and receptor status, cases of HR-positive or triple-negative metastatic breast cancer were found to have lower concordance (75% and 85%, respectively) compared with non-metastatic or HER2/neu-positive cases (94%–98%; Figure 2).

Table 3 presents results from the logistic regression of concordance as a function of patient age (grouped in five categories), cancer stage, receptor status, and stage by receptor status interactions. Reported are odds ratios, which denote the odds that concordance will occur in a given group relative to the odds of concordance in the reference group. An odds ratio greater than 1 indicates greater odds of concordance, equal to 1 suggests equal odds, and less than one indicates lesser odds.

Compared with patients <45 years of age, concordance declined significantly with increasing age ($P \leq 0.02$ or $P < 0.001$), except for the age group 55–64 years. Concordance was particularly low for patients 75 years of age and older. Odds ratios of concordance by stage, showed that compared with stage I disease, breast cancer stages II–IV were significantly ($P < 0.05$) more likely to be concordant. Finally, cancer stage by receptor status was also included in the logistic regression analysis. It revealed that only stage IV-triple negative breast cancers were significantly less likely to be concordant ($P < 0.05$).

Discussion

This retrospective, observational study shows that WFO can make breast cancer treatment recommendations that are consistent in a large proportion of cases with those of an expert panel of cancer specialists in Bangalore, India. This is an important observation because it demonstrates that an AI CDSS trained by cancer experts at Memorial Sloan Kettering Cancer Center in the United States carried out well in India with different cases and breast cancer experts. The study also demonstrated that this level of agreement between the expert panel and WFO was dependent on the contemporaneous assessment of cases by both agents. The time sensitivity in concordance levels observed in this study were likely the result of rapid advances in breast cancer treatment over the 3-year period.

Non-concordance between WFO and the multidisciplinary tumor board occurred in 7% of cases. In 23% of these cases, non-concordance was due to the availability of therapies in India that were not included in the US-trained oncology advisor. This can result from differences in regulatory approval processes between countries, and can be remedied by including locally approved therapies in the expert system's knowledge base, which had not been done for this study. Non-concordance can also result from variations in the aggressiveness of treatment approaches in patient subpopulations based on demographic characteristics such as comorbidity burden, patient preferences, and level of social support systems. In fact, when we examined concordance by age, we found that concordance decreased as age increased, especially in elderly patients.

Treatment non-concordance may mean that WFO has suggested a treatment regimen not considered by the treating oncologist rather than a regimen that the oncologist had considered but rejected. Since WFO provides evidence for its decision, the oncologist can examine the data and consider the basis for WFO's recommendation. Although we did not assess this situation in the current study, in routine use, an unexpected recommendation by WFO may spur the clinician to examine his or her own weighted evidence and to reconsider it based on the oncology treatment advisor's assessment. Treatment decisions might not change because of this transparency, but learning would be fostered. Use of a CDSS may also help assure more standardization of treatment when appropriate, decreasing the

Table 2. MMDT and WFO recommendations after the initial and blinded second reviews

Review of breast cancer cases ($N = 638$)	Concordant cases, *n* (%)			Non-concordant cases, *n* (%)		
	Recommended	For consideration	Total	Not recommended	Not available	Total
Initial review ($T1_{MMDT}$ versus $T2_{WFO}$)	296 (46)	167 (26)	463 (73)	137 (21)	38 (6)	175 (27)
Second review ($T2_{MMDT}$ versus $T2_{WFO}$)	397 (62)	194 (30)	591 (93)	36 (5)	11 (2)	47 (7)

$T1_{MMDT}$, original MMDT recommendation from 2014 to 2016; $T2_{WFO}$, WFO advisor treatment recommendation in 2016; $T2_{MMDT}$, MMDT treatment recommendation in 2016; MMDT, Manipal multidisciplinary tumor board; WFO, Watson for Oncology.

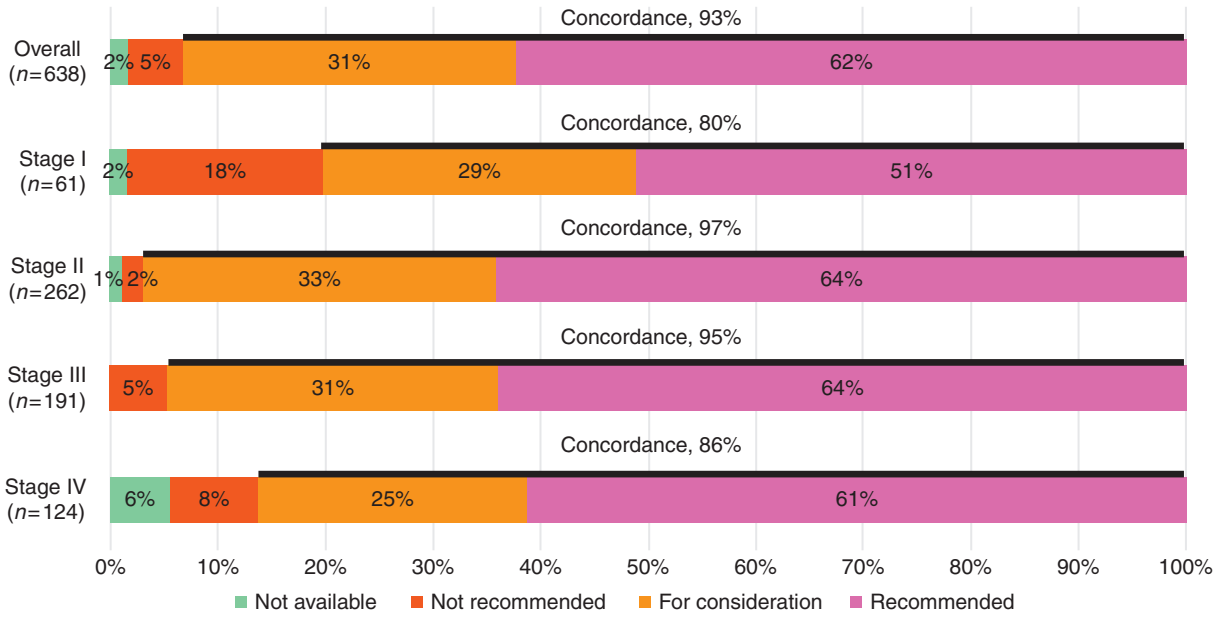

Figure 1. Treatment concordance between WFO and the MMDT overall and by stage. MMDT, Manipal multidisciplinary tumor board; WFO, Watson for Oncology.

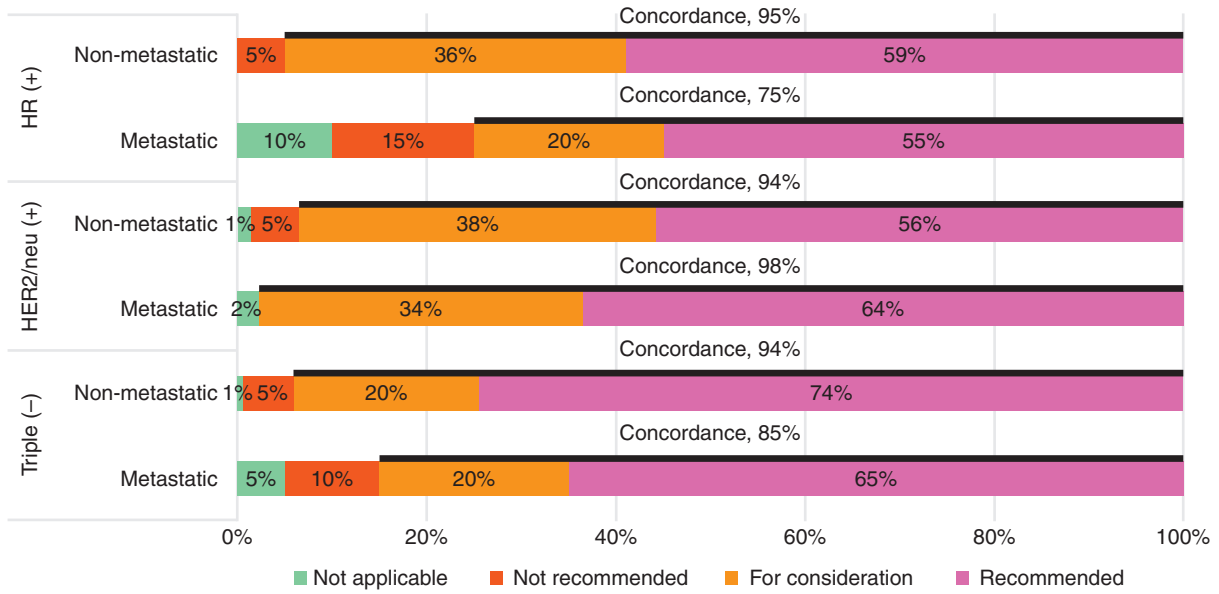

Figure 2. Treatment concordance between WFO and the MMDT by stage and receptor status. HER2/neu, human epidermal growth factor receptor 2; HR, hormone receptor; MMDT, Manipal multidisciplinary tumor board; WFO, Watson for Oncology.

Table 3. Logistic regression model of concordance between WFO and the MMDT

Variable	Odds ratio (95% CI)	*P* value
Age, years		
<45 (reference)	1.00	
45–54	0.22 (0.06–0.82)	0.02
55–64	0.34 (0.09–1.31)	0.12
65–74	0.07 (0.02–0.30)	<0.001
75+	0.007 (0.001–0.04)	<0.001
Stage		
I (reference)	1.00	
II	16.07 (1.74–148.34)	0.01
III	8.53 (1.78–40.77)	0.01
IV	23.24 (2.32–233.09)	0.01
Receptor status		
HER2/neu (+) (reference)	1.00	
HR (+)	0.53 (0.11–2.48)	0.42
Triple (−)	3.28 (0.27–39.59)	0.35
Stage by receptor status		
II×HR (+)	2.03 (0.15–27.55)	0.60
II×Triple (−)	0.13 (0.005–3.48)	0.23
III×HR (+)	1.35 (0.17–10.49)	0.77
III×Triple (−)	0.14 (0.008–2.42)	0.18
IV×HR (+)	0.15 (0.01–2.32)	0.17
IV×Triple (−)	0.02 (0.001–0.76)	0.04

CI, confidence interval; HER2/neu, human epidermal growth factor receptor 2; HR, hormone receptor; MMDT, Manipal multidisciplinary tumor board; WFO, Watson for Oncology.

vagaries of therapy that can occur when personal preference or well-known cognitive biases dominate individuals' decisions [11, 12]. Finally, lack of concordance does not necessarily provide evidence regarding whether the MMDT or WFO was 'correct' in its recommendation. As previously discussed, disagreements can have many valid explanations such as differences in how patients with comorbidities and increasing age are treated. There is no 'gold standard' beyond expert opinion, which studies have shown can vary profoundly during the evaluation of CDSSs [13].

This study has several important strengths. First, it includes a large number of cases and compares the recommendation of WFO to that of a multidisciplinary tumor board (MMDT). Second, the MMDT and WFO were blinded to each other's treatment recommendations, eliminating potential influence by the other's previous assessment. Third, the time discrepancy between the original assessment by the tumor board and WFO are partially addressed by a blinded, second review of non-concordant cases by the tumor board.

This study contains several notable limitations. First, the study design was observational and lacked controls, making the results potentially vulnerable to the influence of unmeasured factors. Second, although the physicians who entered cases into WFO were familiar with the system and the key elements that needed to be abstracted during the chart review process, the quality and reproducibility of these tasks were not formally tested and could have affected WFO's recommendations. Poor quality in the performance of these tasks would be expected to cause lower quality recommendations by WFO and to reduce concordance. However, since overall concordance with the multidisciplinary board was high, we conclude that the quality of data abstraction and entry was good. Third, the time discrepancy difference between the original tumor board assessments and the later WFO review was addressed only partially by the blinded tumor board re-review of the 175 cases that were originally non-concordant. It is assumed, but not proven, that the remaining 463 originally concordant cases would have remained concordant if they also underwent a second blinded review. The concordance results reported in this study should not be extrapolated to other implementation sites where levels of multidisciplinary expertise and previously discussed factors affecting concordance may differ.

In conclusion, treatment decisions made by WFO exhibited a high degree of agreement with those of the multidisciplinary tumor board. Rates of concordance varied by stage and patient's age. While drug therapy availability, professional treatment guidelines, and the judgment of experts who train the computing system will influence concordance, WFO was robust in this application. It suggests that an AI-based advisory system may have broad value in offering breast cancer treatment advice, particularly for environments where expert resources are not readily available. Furthermore, these results suggest that WFO offers an AI computing methodology that may be an effective decision-support tool in cancer therapy.

Acknowledgements

The authors wish to thank Mark Megerian and Dr Irene Dankwa-Mullan of IBM Watson Health for their support on the technical description of the Watson for Oncology technology.

Funding

IBM Watson Health provided funds to assist with biomedical informatics and professional medical writing assistance (no grant number applies). The Manipal Hospital Comprehensive Cancer Center's statistician assigned to the research study was funded by an IBM grant of US$20,137, #SSPMS0912. There were no additional external funding sources.

Disclosure

Co-authors M-JS and EHS were paid technical consultants of IBM Watson Health providing biomedical informatics and professional medical writing assistance. KR, SP, and AN were IBM employees. SPS and all other co-authors have declared no conflicts of interest.

References

1. Seidman AD. Computer-Assisted Decision Support in Medical Oncology: We Need It Now. In ASCO Post 2016. http://www.ascopost.com/issues/april-10-2016/computer-assisted-decision-support-in-medical-oncology-we-need-it-now/ (16 December 2017, date last accessed).
2. Pusic M, Ansermino JM. Clinical decision support systems. BCMJ 2004; 46: 236–239.

3. National Cancer Institute. Drugs Approved for Breast Cancer. National Institutes of Health. 2017. https://www.cancer.gov/about-cancer/treatment/drugs/breast (16 December 2017, date last accessed).
4. Board on Health Sciences Policy. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk (2015). Institute of Medicine. Washington DC: National Academy Press. 2015.
5. Taichman DB, Backus J, Baethge C et al. Sharing clinical trial data—a proposal from the International Committee of Medical Journal Editors. N Engl J Med 2016; 374(4): 384–386.
6. Woolhandler S, Himmelstein DU. Administrative work consumes one-sixth of U.S. physicians' working hours and lowers their career satisfaction. Int J Health Serv 2014; 44(4): 635–642.
7. Shanafelt TD, Gradishar WJ, Kosty M et al. Burnout and career satisfaction among US oncologists. J Clin Oncol 2014; 32(7): 678–686.
8. Duda RO, Shortliffe EH. Expert systems research. Science 1983; 220(4594): 261–268.
9. Mayo RM, Summey JF, Williams JE et al. Qualitative study of oncologists' views on the CancerLinQ rapid learning system. J Oncol Pract 2017; 13(3): e176–e184.
10. Seroussi B, Bouaud J. Using OncoDoc as a computer-based eligibility screening system to improve accrual onto breast cancer clinical trials. Artif Intell Med 2003; 29(1–2): 153–167.
11. Groopman J. How Doctors Think. New York: Houghton Mifflin Company 2007.
12. Kassirer JP, Kopelman RI. Cognitive errors in diagnosis: instantiation, classification, and consequences. Am J Med 1989; 86(4): 433–441.
13. Yu VL, Fagan LM, Wraith SM et al. Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. JAMA 1979; 242(12): 1279–1282.