


Breast Cancer Screening Trials: Endpoints and Overdiagnosis

Ismail Jatoi , MD, PhD,^{1,*} Paul F. Pinsky, PhD²

¹Division of Surgical Oncology and Endocrine Surgery, University of Texas Health Science Center, San Antonio, TX, USA and ²Division of Cancer Prevention, National Cancer Institute, Rockville, MD, USA

*Correspondence to: Ismail Jatoi, MD, PhD, Division of Surgical Oncology and Endocrine Surgery, University of Texas Health Science Center, San Antonio, TX 78229, USA (e-mail: jatoi@uthscsa.edu).

Abstract

Screening mammography was assessed in 9 randomized trials initiated between 1963 and 1990, with breast cancer-specific mortality as the primary endpoint. In contrast, breast cancer detection has been the primary endpoint in most screening trials initiated during the past decade. These trials have evaluated digital breast tomosynthesis, magnetic resonance imaging, and ultrasound, and novel screening strategies have been recommended solely on the basis of improvements in breast cancer detection rates. Yet, the assumption that increases in tumor detection produce reductions in cancer mortality has not been validated, and tumor-detection endpoints may exacerbate the problem of overdiagnosis. Indeed, the detection of greater numbers of early stage breast cancers in the absence of a subsequent decline in rates of metastatic cancers and cancer-related mortality is the hallmark of overdiagnosis. There is now evidence to suggest that both ductal carcinoma in situ and invasive cancers are overdiagnosed as a consequence of screening. For each patient who is overdiagnosed with breast cancer, the adverse consequences include unnecessary anxiety, financial hardships, and a small risk of morbidity and mortality from unnecessary treatments. Moreover, the overtreatment of breast cancer, as a consequence of overdiagnosis, is costly and contributes to waste in health-care spending. In this article, we argue that there is a need to establish better endpoints in breast cancer screening trials, including quality of life and composite endpoints. Tumor-detection endpoints should be abandoned, because they may lead to the implementation of screening strategies that increase the risk of overdiagnosis.

Screening mammography was assessed in 9 randomized trials initiated between 1963 and 1990, with breast cancer-specific mortality as the primary endpoint (1). In contrast, breast cancer detection (ie, cancer detection rates, sensitivity, or area under the receiver operating characteristic curve) has been the primary endpoint in most screening trials (Table 1) and observational studies initiated during the past decade (2–7). These trials have evaluated newer breast screening modalities, including digital breast tomosynthesis (DBT), magnetic resonance imaging (MRI), and ultrasound, and the results provided impetus for implementation of these newer modalities into clinical practice. None of these trials have yet provided mortality data, and novel screening strategies have been recommended solely on the basis of their abilities to detect greater numbers of early stage breast cancers, consistent with the long-held belief that cancers are generally curable if detected early (1).

However, the detection of greater numbers of early stage breast cancers can also lead to overdiagnosis, which refers to the detection of cancers that pose no threat to life and would never have been detected in the absence of screening (8).

Indeed, the detection of greater numbers of early stage cancers in the absence of a subsequent decline in rates of metastatic cancers and cancer deaths is the hallmark of overdiagnosis. Adoption of novel screening strategies solely on the basis of improved tumor-detection endpoints (and the absence of mortality data) may therefore exacerbate the problem of overdiagnosis and do more harm than good.

Evidence for Overdiagnosis

Breast cancer overdiagnosis is a major public health concern, and there is indirect as well as direct evidence to suggest that both ductal carcinoma in situ (DCIS) and invasive cancers are overdiagnosed as a consequence of screening (9, 10). DCIS is rarely palpable, almost always screen detected, and generally regarded as a nonobligate precursor of invasive breast cancer (11). Rates of DCIS detection and extirpation surged nearly 6-fold in the United States between 1975 and 2004, with the widespread implementation of mammography screening (12).

Received: June 16, 2020; Revised: August 19, 2020; Accepted: August 27, 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

Table 1. Breast cancer screening trials initiated from 2010 onward^a

Breast cancer screening trials	No. of trials (%)
All	32 (100)
Screening modality	
2D digital mammography	24 (75.0)
Digital breast tomosynthesis	19 (59.4)
MRI/AB-MRI	7 (21.9)
Ultrasound	6 (18.8)
Other	11 (34.4)
Primary outcome ^b	
Cancer detection rate	10 (31.2)
Sensitivity	6 (18.8)
ROC area/accuracy	7 (21.9)
Recall rate/specificity	6 (18.8)
Advanced cancer rate	4 (12.5)
Interval cancer rate	2 (6.2)

^aBased on ClinicalTrials.gov search. Search criteria: breast cancer; screening AND imaging; interventional studies. Trials with less than 100 subjects, evaluating only a single imaging modality, or in cancer patients excluded. 2D = 2-dimensional; AB-MRI = abbreviated breast magnetic resonance imaging; MRI = magnetic resonance imaging; ROC = receiver operating characteristic.

^bSome trials had 2 primary outcomes.

Substantial declines in the incidence of invasive breast cancer were expected as a result of this surge, and yet, invasive breast cancer incidence gradually increased from about 100.0 to 124.3 cases per 100 000 women during this same time period, providing indirect evidence that most cases of screen-detected DCIS would likely have never progressed to invasive disease (12). Moreover, an autopsy study in Denmark demonstrated that occult DCIS was evident in 15% of a random sampling of women aged 20-44 years with no previous history of cancer who had died of accidents, a prevalence 4 times greater than the number of invasive breast cancers expected to develop over a 20-year period (13). Thus, occult DCIS seems to be highly prevalent in the general population, and most lesions would not be expected to have an adverse effect on mortality.

Observational studies also provide indirect evidence for the overdiagnosis of invasive breast cancer. These population-based studies show that screening may substantially increase the detection rates of early stage invasive breast cancers, but it has only a marginal effect in reducing the incidence of advanced disease (14, 15). Moreover, even though age-specific incidence rates of invasive breast cancer invariably increase among younger women in the age group that undergoes screening, incidence rates of invasive cancer do not ultimately drop below baseline for the elderly population not being screened, thereby providing further indirect evidence for the overdiagnosis of invasive cancers in the screened population (16).

Randomized trials of screening provide direct evidence of overdiagnosis for both in situ and invasive disease. Various methods were used to estimate rates of overdiagnosis in these trials, such as determining the difference in cancer incidence in the presence or absence of screening (ie, observed excess incidence approach), or on the basis of inferences concerning the natural history of breast cancer and the lead time attributable to screening (ie, lead-time approach) (10). Estimates of overdiagnosis rates vary widely because of differences in methodologies and definitions; therefore, it is not possible to give a precise figure for the breast cancer overdiagnosis rate. A recent evidence review for the US Preventive Services Task Force concluded that approximately 11% to 22% of all breast cancer cases (invasive

plus in situ) in the United States may be overdiagnosed (10). However, the evidence cited for that estimate reflected only standard mammography screening; the newer, more sensitive modalities may have higher overdiagnosis rates.

Harms of Overdiagnosis

It is estimated that 25% of total health-care spending in the United States is wasted, and a considerable portion of that waste is attributable to expenditures on unnecessary treatments (17). The overtreatment of breast cancer, as a consequence of overdiagnosis, is costly and undoubtedly contributes to this waste. Moreover, for each patient who is overdiagnosed with breast cancer, there are substantial adverse consequences: unnecessary anxiety, financial hardships, and a small risk of morbidity and mortality from unnecessary treatments. Even for screen-detected DCIS, almost all women receive therapies that are costly and with potential risks; a recent study showed that 99.6% of cases from 2004 to 2015 received some form of therapy, including combinations of different treatment modalities (surgery, systemic treatment, or radiotherapy) (18). Recent advances in breast imaging technology have improved the sensitivity of screening and thereby potentially exacerbated the risk of overdiagnosis and overtreatment.

Given the reality of overdiagnosis, cancer detection is an insufficient and potentially harmful endpoint for a screening trial. Instead, we must consider the full range of possible benefits and harms associated with introducing a more sensitive screening test. A framework for conceptualizing the possible benefits and harms of screening with a standard modality 1 (eg, mammography alone) as compared with a more sensitive modality 2 (eg, mammography combined with MRI) is shown in Figure 1. If a cancer is detected with modality 2 but would have been missed at that time with modality 1, one can examine the potential outcomes if the woman had never been screened with modality 2 but had continued screening only with modality 1. The following are possible outcomes: interval cancer diagnosed later, later screen detection of cancer, and cancer would never have been diagnosed (overdiagnosis). For the first 2 outcomes, the cancer could have been diagnosed in the same stage (eg, 1A) as with modality 2 or at a higher stage; presumably, it would be more likely to be at a higher stage with an interval than with a screen-detected cancer. For the same stage, prognosis and treatment can be assumed similar; therefore, earlier diagnosis per se can be treated as a minor harm of modality 2 (relative to 1). In the absence of clinical benefit, the earlier diagnosis of breast cancer may simply increase the time period that a patient lives with the knowledge that she has cancer and thereby possibly increase the period of anxiety. For a higher stage diagnosis, it could make the difference between dying of breast cancer (with modality 1) or not (with modality 2), clearly a major benefit of modality 2. Alternatively, the woman could survive under either modality but have more intense treatment or disease course with modality 1, thus a minor-moderate benefit of modality 2 in terms of improved quality of life. Finally, an overdiagnosed cancer is a major harm of modality 2.

Tumor Detection Endpoints

Unfortunately, current breast cancer screening practices are often predicated on clinical studies with tumor-detection endpoints. As seen in Table 1, of the 32 breast cancer screening trials initiated since 2010, 23 (71.8%) had the cancer detection

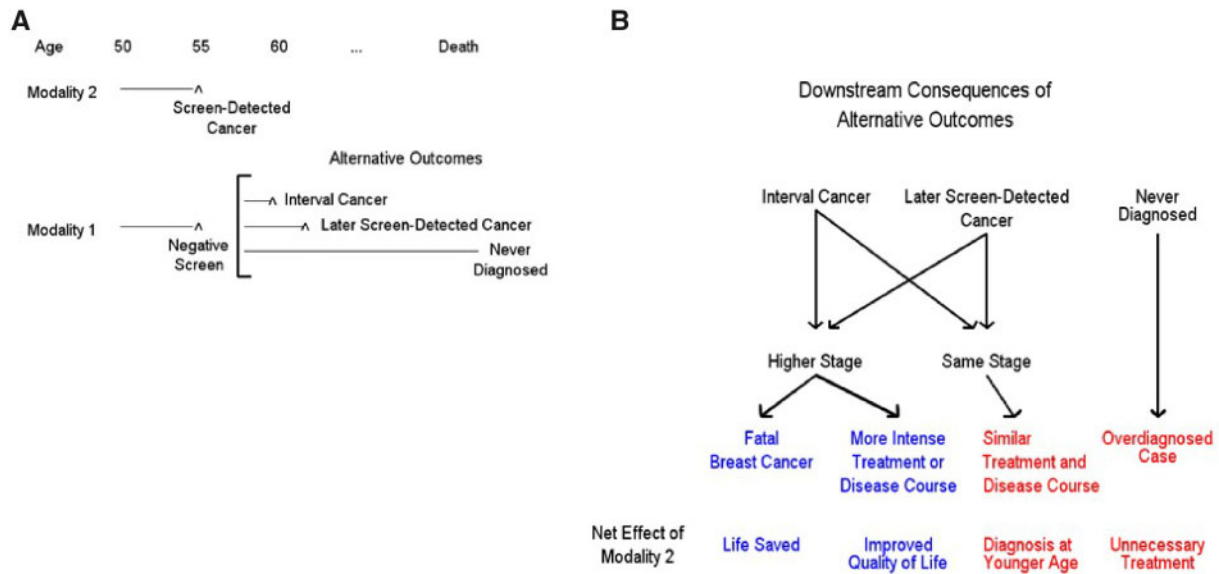


Figure 1. Consequences of earlier detection. **A)** With cancer present at a given age (here, 55 years), the standard modality (1) gives a negative screen, whereas the more sensitive modality (2) finds a screen-detected cancer. Alternative outcomes of what would have transpired if the woman had undergone only modality 1 screening are shown. **B)** Downstream consequences of the alternative outcomes are shown, as well as the net effect of using modality 2. Some scenarios (in blue) are relative benefits of modality 2 relative to 1, whereas others (in red) are potential harms.

rate, sensitivity, or area under the receiver operating characteristic curve as a primary endpoint. A large observational study from 13 academic and nonacademic breast centers showed that the addition of DBT to screening mammography decreased false-positive rates and statistically significantly increased breast cancer detection rates (2). The results of this and similar single institution studies fueled the rapid implementation of screening DBT throughout the United States (19). For women with a familial breast cancer risk, a randomized trial from the Netherlands showed that screening mammography combined with MRI detects greater numbers of cancers and those at an earlier stage when compared with screening mammography alone (3). As a consequence of this trial and observational studies with similar findings, the addition of breast MRI screening to mammography is generally recommended for women with a familial risk (20). Similarly, for women with dense breasts and a negative screening mammogram, numerous studies have shown that supplemental screening with either DBT, MRI, or ultrasound statistically significantly increases breast cancer detection rates, and supplemental screening is now widely recommended for women with dense breasts (4, 5). To further elucidate the optimal screening method for women with dense breasts, a phase II trial (ECOG-ACRIN 1141) compared abbreviated breast MRI and DBT, with rates of invasive breast cancer detection as the primary endpoint (6). Abbreviated breast MRI detected statistically significantly greater numbers of invasive breast cancers, and this was interpreted as a favorable screening outcome, supporting its use as the preferred screening method for women with dense breasts.

The Dense Tissue and Early Breast Neoplasm Screening trial in the Netherlands compared screening with mammography alone vs mammography combined with supplemental MRI for women with extremely dense breasts, and the primary outcome of interest was the interval cancer detection rate (ie, cancers detected during the intervals between screening sessions) in the 2 arms of the study (7). Supplemental MRI screening substantially improved sensitivity and thereby decreased the

interval cancer detection rate, but the relation between rates of interval cancer detection and breast cancer mortality is not clear. Although interval cancers generally have a more aggressive tumor biology and worse prognosis than screen-detected cancers, the interval cancer detection rate has not been validated as a proper surrogate outcome measure for mortality in screening trials (21–23). Validation would require conducting a randomized screening trial assessing the screening modalities of interest that analyzed both breast cancer mortality (ie, the true endpoint) and interval cancer detection rate (ie, the surrogate endpoint) (24). Of note, in the Dense Tissue and Early Breast Neoplasm Screening trial, 42% of the interval cancers in the mammography-alone arm were early stage (0 or 1) and 55% were node-negative, and these cancers would be unlikely to have a substantial adverse effect on mortality (7).

Today, nearly half of all women in the United States are categorized as having dense breasts, and most states have enacted legislation requiring that women be notified of their breast density status following screening mammography (25). Furthermore, some states require insurers to cover the cost of supplemental screening for women with dense breasts (26). Although supplemental screening has been shown to increase breast cancer detection rates, its effect on mortality is not known, and the potential for overdiagnosis is a concern.

Clinically Relevant Endpoints

The ideal primary outcome measure for any breast cancer screening trial is all-cause mortality. It is an unambiguous endpoint not prone to assessor bias, but it generates a huge sample size requirement. Therefore, breast cancer-specific mortality was the surrogate endpoint for each of the 9 mammography screening randomized trials initiated during 1963–1990, but even so, tens of thousands of women were required for each trial (27). In recent years, breast cancer treatments have improved and mortality rates have declined substantially, so

larger sample sizes would be required for clinical trials to discern any incremental breast cancer mortality benefits of novel screening methods. It is generally understood that if such large sample sizes were required to demonstrate further mortality benefits of screening, then these gains would likely be very small in absolute terms. Moreover, newer breast cancer screening technologies are developing at a very rapid pace, and the lengthy follow-up required to assess these emerging technologies with mortality endpoints is no longer feasible. Therefore, to circumvent the problem of larger (and unattainable) sample sizes and to reduce the time required to conduct screening trials, investigators have, as noted above, resorted to other surrogate outcomes (ie, tumor-detection endpoints) to assess the efficacy of novel screening strategies.

However, the assumption that increases in breast cancer detection rates invariably result in reductions in mortality has not been validated. For example, after 15 years of follow-up in the Canadian National Breast Cancer Screening Study, there was a residual excess of 106 breast cancers among women randomly assigned to mammography screening when compared with the unscreened control group, and yet there was no difference in mortality between the 2 arms of the trial (28).

Clearly, we need to be wary of tumor-detection endpoints for screening studies, because increased and earlier detection does not necessarily result in increased net benefit (benefits minus harms). The detection of large numbers of nonlethal cancers may result in an unnecessary excess in treatment-related morbidity and mortality. Rates of distant metastases capture the mortality effects of screening and are therefore better surrogate outcome measures. In the ongoing Tomosynthesis Mammographic Imaging Screening Trial, women are randomly assigned to standard digital breast mammography vs DBT, and the primary outcome is the rate of advanced cancers, an aggregate endpoint that includes distant metastases (29). The trial was designed this way, instead of with a cancer detection endpoint, precisely because the investigators understood that finding more cancers does not necessarily lead to preventing more breast cancer deaths. However, this trial still required the very large sample size of 165 000 women, indicating that the anticipated effect size is very small and raising concerns that a statistically significant benefit of DBT might not be clinically meaningful. Therefore, there is a need to identify endpoints that are more clinically relevant to the average woman who is screened. Specifically, composite endpoints that incorporate both the benefits and harms of screening should be considered for screening trials.

Quality-of-life outcome measures should be considered as part of an overall composite endpoint (30). After all, screening programs target large numbers of healthy, asymptomatic women, and screening would be expected to have a far greater impact on quality of life than on mortality for most women. Patient-reported quality-of-life outcomes, such as pain and discomfort from the screening procedure, as well as recall rates (ie, false-positive results), could potentially be included as part of a composite endpoint. A general issue with composite endpoints is how to apportion the weights, especially when very disparate outcomes are included in the composite (eg, anxiety due to a false-positive vs diagnosis of a metastatic cancer). Clearly, the latter would carry a much greater weight, but determining the exact choice of weights needs careful thought. Thus, the patient-reported quality-of-life outcomes and mortality surrogate outcomes would need to be appropriately weighted with input from potential participants of screening programs, and sample sizes determined accordingly.

Observational Studies and Modeling

Given that newer screening modalities have already been introduced into clinical practice on the basis of studies demonstrating improvements in breast cancer detection rates, it is unlikely that they will be further assessed in randomized trials with mortality or advanced cancer endpoints. Nonetheless, it is important to learn as much as possible about the impact of these newer screening modalities through well-conducted observational studies. Assessing the effects of screening in large integrated health-care systems using data from electronic health records can be performed relatively quickly and efficiently, as, for example, was done in a recent study on colonoscopy (31). Although proving causality is always problematic in observational studies because of unmeasured confounders and other issues, such studies can still provide insight into harms and costs and give some idea of potential magnitude of benefit. Additionally, simulation modeling may help quantify the short- and long-term benefits and harms of novel screening methods (32). Modeling can assess a wide range of screening outcomes, including death, quality of life, and costs, without necessitating long-term patient follow-up. However, the major disadvantage of any modeling study is that it is based on important assumptions, such as the natural history of breast cancer and the effect of earlier diagnosis on mortality. The general framework for examining intermediate and longer-term outcomes with alternate methods of cancer detection is depicted in Figure 1 and may be useful in assessing relative net benefit.

Conclusion

In summary, we believe that more attention should be paid to establishing better endpoints, including composite endpoints, for breast cancer screening trials. Cancer detection as primary endpoints in randomized trials and observational studies should be abandoned, because they may lead to the implementation of breast cancer screening strategies that increase the risk of overdiagnosis.

Notes

Disclosures: The authors have no conflicts of interest to disclose.

Data Availability

The data in this article are available in ClinicalTrials.gov, at <http://www.Clinicaltrials.gov>

References

1. Jatoi I, Anderson WF, Miller AB, Brawley OW. The history of cancer screening. *Curr Probl Surg*. 2019;56(4):138-163.
2. Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA*. 2014; 311(24):2499-2507.
3. Saadatmand S, Geuzing HA, Rutgers EJT, et al. MRI versus mammography for breast cancer screening in women with familial risk (FaMRIsc): multicentre, randomised, controlled trial. *Lancet Oncol*. 2019;20(8):1136-1147.
4. Melnikow J, Fenton JJ, Whitlock EP, et al. Supplemental screening for breast cancer in women with dense breasts: a systematic review for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2016;164(4):268-278.
5. Ohuchi N, Suzuki A, Sobue T, et al. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. *Lancet*. 2016;387(10016):341-348.

6. Comstock CE, Gatsonis C, Newstead GM, et al. Comparison of abbreviated breast MRI vs digital breast tomosynthesis for breast cancer detection among women with dense breasts undergoing screening. *JAMA*. 2020;323(8):746–756.
7. Bakker MF, de Lange SV, Pijnappel RM, et al. Supplemental MRI screening for women with extremely dense breast tissue. *N Engl J Med*. 2019;381(22):2091–2102.
8. Welch HG, Prorok PC, O'Malley AJ, Kramer BS. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. *N Engl J Med*. 2016;375(15):1438–1447.
9. Srivastava S, Koay EJ, Borowsky AD, et al. Cancer overdiagnosis: a biological challenge and clinical dilemma. *Nat Rev Cancer*. 2019;19(6):349–358.
10. Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the 2009 U.S. Preventive Services Task Force recommendation. *Ann Intern Med*. 2016;164(4):256–267.
11. Jatoi I, Baum M. Mammographically detected ductal carcinoma in situ: are we overdiagnosing breast cancer? *Surgery*. 1995;118(1):118–120.
12. Virnig BA, Tuttle TM, Shamlilian T, Kane RL. Ductal carcinoma in situ of the breast: a systematic review of incidence, treatment, and outcomes. *J Natl Cancer Inst*. 2010;102(3):170–178.
13. Nielsen M, Thomsen JL, Primdahl S, Dyreborg U, Andersen JA. Breast cancer and atypia among young and middle-aged women: a study of 110 medicolegal autopsies. *Br J Cancer*. 1987;56(6):814–819.
14. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *N Engl J Med*. 2012;367(21):1998–2005.
15. Jorgensen KJ, Gotzsche PC, Kalager M, Zahl PH. Breast cancer screening in Denmark: a cohort study of tumor size and overdiagnosis. *Ann Intern Med*. 2017;166(5):313–323.
16. Jatoi I, Anderson WF. Breast cancer overdiagnosis with screening mammography. *Arch Intern Med*. 2009;169(10):999–1000; author reply 1000–1001.
17. Shrank WH, Rogstad TL, Parekh N. Waste in the US health care system: estimated costs and potential for savings. *JAMA*. 2019;322(15):1501.
18. Fan B, Pardo JA, Alapati A, Hopewood P, Mohammad Virk Z, James TA. Analysis of active surveillance as a treatment modality in ductal carcinoma in situ. *Breast J*. 2020;26(6):1221–1226.
19. Richman IB, Hoag JR, Xu X, et al. Adoption of digital breast tomosynthesis in clinical practice. *JAMA Intern Med*. 2019;179(9):1292.
20. Kuhl CK. Underdiagnosis is the main challenge in breast cancer screening. *Lancet Oncol*. 2019;20(8):1044–1046.
21. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer*. 2017;3(1):12.
22. Irvin VL, Zhang Z, Simon MS, et al. Comparison of mortality among participants of women's health initiative trials with screening-detected breast cancers vs interval breast cancers. *JAMA Netw Open*. 2020;3(6):e207227.
23. Baker SG, Kramer BS. A perfect correlate does not a surrogate make. *BMC Med Res Methodol*. 2003;3(1):16.
24. Grimes DA, Schulz KF. Surrogate end points in clinical research: hazardous to your health. *Obstet Gynecol*. 2005;105(5 pt 1):1114–1118.
25. Saulsberry L, Pace LE, Keating NL. The impact of breast density notification laws on supplemental breast imaging and breast biopsy. *J Gen Intern Med*. 2019;34(8):1441–1451.
26. Horny M, Shwartz M, Duszak R Jr, Christiansen CL, Cohen AB, Burgess JF. Jr. Characteristics of state policies impact health care delivery: an analysis of mammographic dense breast notification and insurance legislation. *Med Care*. 2018;56(9):798–804.
27. Nelson HD, Fu R, Cantor A, Pappas M, Daeges M, Humphrey L. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. Preventive Services Task Force recommendation. *Ann Intern Med*. 2016;164(4):244–255.
28. Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. 2014;348:g366.
29. TMIST PI Pisano: "are we further reducing breast cancer mortality the more intensively we screen?" *The Cancer Letter*. 2019.
30. Jatoi I, Gail MH. The need for combined assessment of multiple outcomes in noninferiority trials in oncology. *JAMA Oncol*. 2020;6(3):420.
31. Levin TR, Corley DA, Jensen CD, et al. Effects of organized colorectal cancer screening on cancer incidence and mortality in a large community-based population. *Gastroenterology*. 2018;155(5):1383–1391.e5.
32. Lowry KP, Trentham-Dietz A, Schechter CB, et al. Long-term outcomes and cost-effectiveness of breast cancer screening with digital breast tomosynthesis in the United States. *J Natl Cancer Inst*. 2020;112(6):582–589.