The American Journal of Surgery 218 (2019) 411–418

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

Review

Redefining the sensitivity of screening mammography: A review

Alan B. Hollingsworth

Department of Surgery, Mercy Hospital, 4401 W. McAuley Blvd., Suite #1100, Mercy Hospital Coletta Building, Oklahoma City, OK, USA

ARTICLE INFO

Article history:
Received 4 September 2018
Received in revised form
25 January 2019
Accepted 31 January 2019

ABSTRACT

From its inception, screening mammography has enjoyed a perceived level of sensitivity that is inconsistent with available evidence. The original data that imparted erroneous beliefs about sensitivity were based on a variety of misleading definitions and approaches, such as the inclusion of palpable tumors, using the inverse of interval cancer rates (often tied to an arbitrary 12 month interval), and quoting prevalence screen sensitivity wherein tumors are larger than those found on incidence screens. This review addresses the background for the overestimation of mammographic sensitivity, and how a major adjustment in our thinking is overdue now that multi-modality imaging allows us to determine real time mammographic sensitivity. Although a single value for mammographic sensitivity is disingenuous, given the wide range based on background density, it is important to realize that a sensitivity gap between belief and reality still exists in the early detection of breast cancer using mammography alone, in spite of technologic advances. Failure to recognize this gap diminishes the acceptance of adjunct methods of breast imaging that greatly complement detection rates.

Introduction

When screening mammography was being integrated into clinical practice in the late 1970s and early 1980s, a pervasive notion of 90–95% sensitivity accompanied the new test. Confidence in mammography was so high that clinicians began to dismiss lumps they would have previously sent for biopsy. This led to a surge in medical malpractice where Kern's "Triad of Error" was the most common scenario: 1) young patient, 2) self-discovered mass, and 3) negative mammograms.[1] Risk management courses and educational efforts became commonplace for many years before a healthy skepticism about the 90–95% sensitivity took hold. The origins of the 90–95% belief are discussed herein, along with the gradual adjustments downward for mammographic sensitivity that are still in progress due to the impact of real time determinations now possible through multi-modality imaging.

Specificity of various imaging modalities, critical for cost and feasibility analyses, is an inherent component of those analyses intended to justify asymptomatic screening for breast cancer. However, this overview is not intended to address the benefit or harms of screening. The focus is exclusively on redefining mammographic sensitivity levels in order to accommodate the introduction of supplemental imaging methods, which are becoming key components of high-risk and high-density screening guidelines. Much of the resistance to accepting supplemental screening is likely based in outdated – and inflated – sensitivity values ascribed to mammography. This review is intended to create a more realistic approximation of the cancer detection rate when relying on mammograms alone.

Origins of quoted mammographic sensitivity

Little distinction was made in the early days of mammography between studies that included palpable tumors and those limited to asymptomatic screening. In fact, there was not a consensus at that time as to the exact definition of screening, some wanting to include lumps palpated through expert clinical exam that were not apparent to the patient. Yet, Dr. Robert L. Egan, mammography pioneer at what was then called The University of Texas M.D. Anderson Hospital and Tumor Institute at Houston, set the standard in 1962 for today's definition by stating that the occult cancers identified by screening mammography are the ones that are "totally unsuspected following examination by the usual methods used to diagnose breast cancer, including an examination of the breast by an experienced and competent physician. To qualify for this definition, no symptoms or signs should be present."[2]

In spite of this working definition, the first prospective randomized trial to determine if a mortality reduction could be achieved through breast cancer screening included clinical exam along

E-mail address: alan.hollingsworth@mercy.net.

https://doi.org/10.1016/j.amjsurg.2019.01.039
0002-9610/

with mammography. Begun in 1963, the Health Insurance Plan of Greater New York[3] culminated with enough ambiguity and design flaws that the study is often excluded from modern meta-analyses. Subsequent investigators have been challenged in determining how many cancers were discovered by mammography alone in the HIP study, but an accepted figure for mammographic sensitivity is 39%,[4] a sharp contrast to what eventually emerged as the commonly believed 90–95%.

The international screening trials that followed were mixed with regard to inclusion of the clinical exam. The Canadian National Breast Screening Study (CNBSS) emphasized clinical exams (and self-exam instruction) not only in the mammography limbs but also in control groups.[5] While organizers of the trial have long maintained that expert clinical exam and self-exam are as useful as screening mammography, critics point out that the CNBSS cannot be included as a pure study of mammographic screening due to the confounding imparted by clinical exams, and of great concern, the enrollment of patients into the trial subsequent to the detection of suspicious palpable masses.[6]

The definition of mammographic screening as being applicable only to asymptomatic women was gradually accepted, in line with the original description by Dr. Egan. To that end, the remaining international trials employed that definition, generating mortality reduction data based on mammography as the sole screening modality. And while sensitivity in those trials is often referenced as a range, e.g. from 71% to 98%,[7] a closer look at published analyses reveals those numbers are applicable to the first screen only, that is, the prevalence screen where tumors are larger.

Prevalence screens reflect the proportion of breast cancers in a population at a particular point in time. While it takes several screening rounds to reach a steady state, by convention, the term "prevalence screen" describes the first round of screening. Subsequent rounds of screening are referred to as "incidence screens," which reflect the rate of occurrence of new cases (disease incidence). Long-term analysis when analyzing incidence screens only will generate sensitivity levels 15–30% lower than the first round of screening.[7] Yet, these lower sensitivity levels for long-term incidence screening are seldom referenced, even though incidence screens comprise the bulk of screening activity.

Prevalence screens introduce the same bias as when palpable cancers are included in reporting performance characteristics – that is, the tumors are larger and therefore more easily detectable by mammography. Sensitivity calculations are therefore higher. In recommending a screening strategy to women, it is misleading to quote a sensitivity rate that applies to the first screen only when the patient is asking, "If I develop breast cancer *at some point in my life*, what are the chances it will appear on my mammogram before I can feel it?" The answer to that question comes from sensitivity calculations based on long-term incidence screens, not the studies that include palpable cancers or quotes applicable to prevalence screens.

Even though the international trials have confirmed a breast cancer mortality reduction in meta-analyses,[7] they did so under remarkable disadvantages. Long intervals between screenings, single view mammography, nascent technology from 40 years ago that preceded quality standards – all conspired to hinder sensitivity, and thus muted the measurable effect on mortality reduction. The fact that lives were saved at all is testimony to the vulnerability of breast cancer biology to early detection. Yet, while these international trials were underway, the United States would take a different approach, one that secured the mistaken belief that mammograms had 90–95% sensitivity in the setting of asymptomatic screening.

Given the confusion and controversy generated by the HIP as the first screening study, the recommendation by many epidemiologists was for the United States to conduct its own prospective, randomized trial with a focus on the controversial age group for screening, 40 to 49. Instead, the National Cancer Institute, with its generous budget provided by the Cancer Control Act of 1971, and with full support by the American Cancer Society, made the decision to launch an observational study to see if it would be feasible to screen the general female population in the U.S. as a routine part of health care, as had been successful with the Pap smear starting in the 1950s.

From 1973 to 1980, over 280,000 women at 29 sites underwent 5 screens using both clinical exam and mammography (and thermography until 1977) in what was called the Breast Cancer Detection Demonstration Project (BCDDP).[8] In the under 50 group, mammography alone detected 44% of the cancers, while both mammography and clinical exam detected 46%, for a 90% combined sensitivity. Only 10% were detected on exam alone. In the over-50 group, 95% sensitivity was recorded. Thereafter, the 90–95% sensitivity for mammography became a presumed fact, manifest most commonly in radiology disclaimers at the bottom of mammography reports, stating the inverse: "5–10% of breast cancers are not visible on mammography."

Lost in the enthusiasm for the new screening tool, few noticed that the cohort was heavily populated with palpable cancers, nearly one-half of all cases, in conflict with the definition of screening as articulated by Dr. Egan years earlier. Indirect evidence that the BCDDP was heavily weighted with palpable cancers is the fact that there were more Stage II patients (n = 1375) diagnosed than Stage 0 and Stage I combined (n = 1306). The shaky origins of the oft-quoted 90–95% sensitivity would go unchallenged for decades, in spite of the much lower sensitivity figures that had been seen in the HIP and the international trials.

However, in an undercurrent of concern, there were questions raised about the true sensitivity of mammography. Studies emerged noting the negative impact of density,[9] the disproportionate incidence of mammographically occult lobular carcinoma,[10] and then the appreciation that diffuse growth patterns were sometimes undetectable on mammography regardless as to whether the histology was lobular or ductal.[11]

Using interval cancer rates to define sensitivity

Interval cancer rates vary widely in a 7-fold range,[12] based on a variety of factors including age, breast density, cohort risk levels, and radiologic expertise. When the concept was introduced to clinicians, it carried a fatalistic connotation – that is, nothing could be done about the underlying biology of a fixed number of breast cancers. If these aggressive tumors began growing after a true negative mammogram, they were going to emerge between screenings no matter what we do. Rather than focusing on the wide range of interval cancer rates, a truism emerged that 20–25% of breast cancers would become clinically evident after a negative mammogram and before the next screening round.

And while this 20–25% might be a pragmatic performance characteristic (stated in the inverse as 75–80% sensitivity), interval cancer rates were not an accurate measure of true sensitivity for screening mammography. A tumor not clinically present at the time of screening cannot be counted as a false-negative, in spite of the fact that it is a screening failure. And while aggressive biology for interval cancers, in general, can be documented from a variety of angles,[13] it became apparent that many of these so-called interval cancers were actually present on the prior film but missed by faulty interpretation,[14] or detectable cancers that were simply buried in a zone of mammographic density,[15] without any real difference in biology from screen-detected cancers.

Without a secondary form of breast imaging, the number of

detectable cancers buried in an area of density on the prior mammogram was completely unknown. The fast-growing tumors would emerge as an interval cancer, but the slower-growing tumors would be discovered on mammography at the next screen, and thus, would *not* be counted as a mammographic miss. Indeed, these latter tumors diagnosed by the next mammogram are considered evidence for successful screening when, in fact, by using multi-modality imaging, many would be considered a miss on the prior mammogram.

Thus, counting interval cancers as an inverse to sensitivity is flawed at many levels unless the next mammographic detection is paradoxically counted as last year's miss. As improbable as that approach might seem, we will see it used by the American College of Radiology in a major trial, discussed below, wherein any cancer found within 15 months after a negative mammogram is considered a prior miss, regardless of the method of detection.

Historically, while interval cancer rates seemed unavoidable, they also prompted skepticism as to the pervasive notion of 90–95% sensitivity of mammography. If mammograms were really so sensitive, then the only way to explain the inverse of interval cancer rates (75–80%) was if interval cancers were nearly all classic, aggressive cancers that begin and grow very quickly after the negative mammogram.

But as it gradually became more apparent that many, if not most, interval cancers had an average growth rate but happened to be buried in density, the implication was that the 90–95% sensitivity mantra should be adjusted downward. An example of this subtle modification was an adjustment from the widely held belief of 90–95% sensitivity to "80–90% sensitivity,"[16] as quoted over many editions of *Cancer Facts and Figures*, published by the American Cancer Society. In spite of *Cancer Facts and Figures* being a heavily referenced document, there was never a reference provided for this 80–90% figure. As more data emerged and the complexity of sensitivity became appreciated, when the 2015 edition of *Cancer Facts and Figures* was published, a numerical estimate of mammographic sensitivity was no longer mentioned, while noting that "sensitivity is lower for younger women and women with dense breasts."[17]

The American College of Radiology was attuned to the sensitivity issues when the lexicon was implemented for the Breast Imaging Reporting and Data System (BI-RADS®) in 1994, long before breast density became a cause célèbre.[18] The cautionary disclaimers began, not with Level C density, but with what we would today call Level B, that is, "scattered fibroglandular tissue." Radiologists were instructed to include a statement about "possible diminished sensitivity" at this level of density. As one moved along the density continuum toward "extremely dense" tissue, the lexicon included stronger warnings. But without a numerical reference that quantified the diminished sensitivity, few paid attention to the disclaimers present on standardized mammographic reporting.

Breast density being responsible for missed cancers was recognized early,[19] and while density as an independent risk factor was also introduced early,[20] the acceptance of density as a risk factor would lag more than 20 years behind the sensitivity problem. That said, the fact that disclaimers on sensitivity originally included patients with what we today call Level B density was even more justified if the most likely place for cancer to begin is within a patch of white, as has been proposed.[21]

The overall density pattern, after all, is only a surrogate for the level of density immediately adjacent to the borders of the tumor. A solitary patch of white can conceal a tumor due to the lack of X-ray attenuation, unless tumor margins interface at some point with the darker areas of predominantly adipose tissue. Multi-modality imaging later confirmed the phenomenon of missed cancers in relatively low density breasts.[22]

As the history of sensitivity determinations unfolded, new technology emerged in the form of digital mammography, and it was the hope of many that sensitivity would improve. The immediate goal, however, with digital mammography was not superior sensitivity, but equivalency to film screen technology with respect to general performance characteristics. All radiologic images were being converted to digital technology at the time, and the goal of equivalency was the requirement by the FDA for approval of digital mammography. Still, the comparative approach of two technologies allowed yet another look at mammographic sensitivity.

DMIST and the conventional 12-month follow-up

During the era of single modality breast imaging, studies of sensitivity were limited to a designated follow-up period, usually 12 months. Without another imaging approach to define missed cancers, it became conventional to count any cancer as occurring prior to the next mammogram as a miss, thus introducing the aforementioned linkage of sensitivity to the inverse of the interval cancer rate. An interval cancer rate of 20% translated to a sensitivity of 80%, providing a functional sensitivity perhaps, but not a true sensitivity of the test itself.

When it came time to compare film screen mammography to digital technology, the American College of Radiology study organizers opted to be unconventional. In the Digital Mammographic Imaging Screening Trial (DMIST),[23] the decision was made to monitor patients for 455 days after study entry, this 15-month period of follow-up being every bit as arbitrary as the conventional 365 days. Analysis would be performed at both 12-month and 15-month follow-ups, subsequent to screening mammography that utilized both techniques – film screen and digital – in every participant, which ended up being 42,760 women in the analysis.

In another unconventional vein, the decision was made to designate any cancer arising in the 12- or 15-month period as a miss, regardless of the method of discovery. As mentioned earlier, this created the odd paradox of a mammographically-detected cancer being labeled as a mammographic miss on the prior exam. Stated alternatively, it was assumed that any cancer detected, even on mammography, was present up to 455 days earlier, acknowledging that this was probably not the case in all patients. Regardless, it sets the stage for lower sensitivities than if one uses only the inverse of interval cancers as defined by palpable tumors that arise in between screens.

The trial concluded that accuracy of digital and film mammography was similar, but with digital having the edge on sensitivity for women under age 50, or premenopausal, and in those with dense breasts. While media coverage focused on the benefit of digital technology in younger women with dense breasts, little mention was made of the absolute calculations of sensitivity which, for digital was 70% at the 12-month interval while film technology was 66%.[23] Then, by extending the analysis an additional 3 months, the differences between the two technologies blurred, with both groups demonstrating 41% sensitivity overall, all density levels combined.

While a far cry from what is routinely quoted for mammographic sensitivity, this improbable 41% will closely approximate mammographic sensitivity as defined by multi-modality studies.

Returning to our original "90–95% sensitivity," it is noteworthy that, in a later analysis of DMIST subgroups (10 groups, each with 2 technologies),[24] none of the 20 calculations reached 90%, including women over 65 with non-dense tissue, and only 3 of 20 calculations reached 80% sensitivity. Most concerning was the second largest sub-group ($n = 7315$) – age under 50 and pre- or

perimenopausal with dense breasts – where film screen sensitivity was 27.3%. This was the only subgroup that reached statistical significance when comparing the two approaches ($p = 0.0013$), with the sensitivity of digital mammography calculated at a much higher 59%, albeit still below a reasonable goal in asymptomatic screening.

While digital mammography was FDA-approved and widely accepted, recalling that the stated goal was mere equivalence to film screen, the much-anticipated approach of 3-D tomosynthesis mammography promised more clear-cut improvements in sensitivity.

The relative improvement in sensitivity with tomosynthesis

The introduction of 3-D tomosynthesis mammography was accompanied by comparison studies, both retrospective and prospective, all showing some degree of improved cancer detection rates (CDRs) with the newer technology over 2-D digital mammography.

In considering the prospective studies, the Oslo Breast Cancer Screening Trial[25] demonstrated a CDR for 2-D digital mammography of 6.1 per 1000 exams, while the CDR for 3-D was 8.0 per 1000 exams. This was a 1.9 per 1000 absolute increase in the cancer detection rate, and when expressed in relative terms, constituted a 27% increase in the sensitivity.

Slightly higher CDR differences were noted in the Screening with Tomosynthesis or Standard Mammography (STORM) Trial[26] from Italy where there was a 2.8 per 1000 increase in CDR using 3-D tomosynthesis. A comparable increase was seen in the Malmö Breast Tomosynthesis Screening Trial[27] where single-view tomosynthesis generated a 2.6 per 1000 increase in CDR. When these absolute numbers are translated to "relative improvement" in sensitivity, one generates a 53% sensitivity improvement (5.3 CDR to 8.1) for the STORM Trial, and 41% improvement (6.3 CDR to 8.9) with the Malmö Trial. Not unexpectedly, the superior performance of 3-D with regard to CDRs was expressed preferentially in relative terms by vendors and enthusiasts over the small increment in absolute CDR.

A consistent finding in nearly all studies has been the predominance of invasive cancer discoveries over DCIS using 3-D technology, lending support for acceptance of the technology under the premise that the discovery of invasive disease is less likely than DCIS to be categorized as overdiagnosis. The emergence of architectural distortions seen only with tomosynthesis has been the strength of 3-D technology, rather than improvement in the detection and analysis of calcium clusters. It is also important to note that the above studies are only comparative, with a focus on increased CDR. Absolute values for sensitivity have not been a primary endpoint, so CDR serves as a surrogate for absolute sensitivity.

The only current option to calculate absolute sensitivity for 3-D tomosynthesis is to begin with data for 2-D digital mammography, then adjust accordingly dependent on the relative impact of 3-D technology. For example, using the 15-month definition for mammographic sensitivity from DMIST of 41%, a 50% relative increase in sensitivity with 3-D would generate 60% sensitivity for 3-D tomosynthesis mammography. But if one presumes the more conventional values for mammographic sensitivity at 80% baseline, then the same 50% relative increase becomes the mathematical impossibility of 120% when 40% (50% of 80%) is added to 80%. A paradox is thus generated – the higher one calculates the incremental, *relative* benefit of 3-D, the lower one must presume the true sensitivity starting point of 2-D technology.

In truth, this exercise is unfair because the improvements in CDR currently being reported are heavily weighted toward prevalence screens where the benefit of 3-D is going to be higher than long-term incidence screens using the same 3-D technology.

All these issues will be settled when true performance characteristics of 3-D technology are achieved through direct analysis, rather than relative improvements. This information will be forthcoming from prospective, randomized trials, primarily the Tomosynthesis Mammographic Imaging Screening Trial (TMIST), sponsored by the Eastern Cooperative Oncology Group – American College of Radiology Imaging Network (ECOG-ACRIN), with support from the National Cancer Institute.[28]

The TMIST study is a massive undertaking that has a target accrual of 165,000 by the end of year 2020. Women ages 45–74 will be screened for 5 years after being randomized to either 2-D digital mammography or 3-D tomosynthesis. This is in contrast to DMIST where each participant had both film screen mammography and digital mammography performed. And, unlike the aforementioned non-randomized 3-D tomosynthesis studies, the analysis will go well beyond CDRs, using comprehensive criteria that address risks and benefits of screening in general.

In spite of the justified enthusiasm for the most significant technologic improvement in mammography since its inception, all current indicators suggest that multi-modality imaging will increase CDRs to a degree greater than 3-D tomosynthesis.

Mutlimodality imaging (without changing the threshold of detection) – ultrasound

Tumor size remains one of the strongest predictors of outcome in patients with invasive breast cancer. And while "mode of detection" further stratifies prognosis, the value of tumor size continues as a prognostic factor independent of the method of detection.[29] And while tumor biology is emerging as a greater predictor of outcome than size, we have little control over biology as cancer develops and becomes clinically evident. In contrast, we do have control over tumor size, and this depends on the methodology used for screening.

When screening with a tool that is based on anatomic contrasts – mammography or ultrasound – the underlying biology of the tumor is not clearly reflected by the imaging. Sojourn time is the pre-clinical phase wherein a tumor becomes "clinically detectable," yet remains in the breast until the next round of screening when it is actually detected. Sojourn times and tumor biology are interrelated phenomena that impact whether or not a mortality reduction will accompany earlier, or more reliable, detection.

But from the standpoint of measuring sensitivity, if two modalities have the same threshold of detection as measured by mean invasive tumor size, sensitivity of the two modalities can be cross-checked in real time, without changing the definition of "clinically detectable." This approach to measuring sensitivity is far superior to arbitrary follow-up periods of 12, 15 or 24 months after single modality imaging. If a 1.0 cm tumor appears on ultrasound, but not mammography, then it is a clear miss for the latter, and vice versa.

From one of the earliest studies of screening ultrasound,[30] it became apparent that there was no significant difference in mean tumor size whether discovered by mammographic screening or ultrasound screening, even though different tumors were identified. Granted, there are sub-groups where mean tumor size shows variance, such as larger tumors associated with higher density on mammography.[31] In general, however, the mean tumor sizes for invasive disease are comparable using either ultrasound or mammography.

In the American College of Radiology Imaging Network (ACRIN) 6666 Trial,[32] comparing screening mammography to ultrasound in patients with dense breasts and at least one additional risk factor, the average size of tumors detected by mammography alone was

11.5 mm while those detected by ultrasound alone measured 10 mm, the slightly larger size with mammography possibly explained by the density inclusion criteria.

Sensitivity calculations were separated into the prevalence screen and two subsequent incidence screens. For mammography alone, incidence screening sensitivity was 52%, and ultrasound sensitivity was 45.3%. Even using both modalities, combined sensitivity was only 76%. Notably, only 55% of the mammography-detected tumors were invasive, while 94% of the ultrasound-detected tumors were invasive. Thus, given one modality or the other, the argument can be made that ultrasound is preferred, given more biologically significant tumors.

Consistent with most studies, the increase in CDR with screening ultrasound in ACRIN 6666 was 3.7 cancers per 1000 incidence screens (2nd and 3rd screens).[32] The slightly higher CDR for ultrasound screening over that seen with 3-D tomosynthesis has raised the question as to which of the two technologies affords a greater improvement in detection. While 3-D tomosynthesis has great appeal in the form of single modality imaging, only one comparative study has reported initial results so far in which the same patients who had 3-D added to 2-D, also underwent ultrasound screening.

An interim report from this prospective, comparative study, the Adjunct Screening with Tomosynthesis or Ultrasound in Women With Mammography-Negative Dense Breasts (ASTOUND) Trial,[33] revealed a CDR for tomosynthesis of 4 per 1,000, while ultrasound CDR was 7.1 per 1,000, both numbers higher than what has been observed in other studies. These higher CDRs for both modalities are likely due to the fact that the interim report was released after a single screening session in each patient, making the results 100% prevalence data at this point. Furthermore, enrollment was limited to women with Level C or Level D density, thus expanding the impact of both modalities.

A breakdown of the 24 additional cancers identified beyond 2-D digital mammography in the ASTOUND Trial reveals a remarkable shift toward invasive disease (23/24) and a rather high rate of node positivity (34.8%). Importantly, 12 cancers were detected on both 3-D and ultrasound, then 11 additional cancers using ultrasound, while only one additional tumor was detected on 3-D tomosynthesis alone. Thus, using ultrasound only as a second modality to 2-D digital mammography identified 23 of 24 additional cancers (96%), while 3-D technology detected 13 of the 24 additional cancers (54%). Rarely do we have to pick one or the other, as 3-D tomosynthesis is gradually replacing 2-D digital, providing improved detection even before ultrasound is considered. Still, the benefit of adding ultrasound appears, at this point, to surpass the added benefit of 3-D tomosynthesis.

Subsequent to the comparative studies that added ultrasound to mammography, largely in higher density patients, it became clear that mammographic sensitivity could not be 80–90% through all levels of density. What emerged was a split approach to describing sensitivity, applying the 80–90% estimate once ascribed to all women by the American Cancer Society,[16] applicable today only to density levels A & B, while admitting that sensitivity in dense breast tissue (levels C & D) slips to 50% or below, borrowing from studies like ACRIN 6666.[32]

While this is a reasonable approach, it reinforces the notion that breast density is a sharp dichotomy when, in fact, it is a continuum that has qualitative aspects in addition to quantitative. In truth, each individual has her own level of mammographic sensitivity. In this era of personalized medicine, it should be considered a worthy goal to calculate this level of sensitivity for individuals, just as we routinely calculate individual levels for breast cancer risk using mathematical models.

The road from "90–95%" sensitivity to the wide range quoted today has taken decades for the adjustment. But we are about to see the quantification of mammographic sensitivity drop even further when the threshold of detection is altered through those imaging methods that employ a functional component – gadolinium for breast MRI or a radionuclide for molecular imaging or contrast-enhanced 3-D tomosynthesis. For purposes of general discussion here, these contrast-enhanced approaches are considered nearly equivalent.

Multimodality imaging that redefines the threshold of detection – breast MRI (or equivalent)

Although ACRIN 6666 focused on adjunct screening ultrasound, there was a secondary modality studied as well – breast MRI. Participants in ACRIN 6666 were offered a single breast MRI to be performed at the end of the study, with only one-fourth of participants accepting this option up front. This generated a sub-group of 612 women in whom 7 cancers were identified over the course of three screens at 0, 12 and 24 months, using two modalities – mammography and ultrasound. Yet, the single MRI at the study's conclusion identified 9 additional cancers, with 8 of the 9 invasive, all node-negative. Mean tumor size for the invasive discoveries was 0.85 cm, which in terms of volume ($V = 4/3\pi^3$), is less than half the size of the usual mammographic discovery (1.15 cm diameter in this particular study). This smaller size with MRI detection not only extends lead time, but redefines sojourn time to an earlier point. With this new definition of "clinical detectability," both mammographic and ultrasound sensitivity calculations drop substantially.

And while there is general agreement that MRI is more sensitive than mammography or ultrasound, the quantification of that difference is not straightforward as most studies are performed in high risk and/or high density patients. In ACRIN 6666, where both density and risk were modestly increased for enrollment, the combined sensitivity of mammography and ultrasound without MRI was 76%. But within the sub-group that opted for MRI, combined sensitivity of mammography and ultrasound was only 44%, while mammographic sensitivity alone was 31.3%.[32]

If this sounds improbably low, similar results are achieved if one uses any of the contrast-enhanced modalities, be it molecular imaging,[34] positron emission mammography,[35] or contrast-enhanced mammography.[36] In brief, we never realized how many cancers were truly missed with mammography until a physiologic agent (contrast) was added to anatomic-based imaging. While these modalities often detect occult tumors that should be large enough to be seen on mammography or ultrasound, the additional reason for re-calculating to lower sensitivity levels with conventional imaging is that the threshold of detection has been lowered.

Thus, there are two ways that contrast-enhanced imaging appears to lower mammographic sensitivity – 1) simply detecting the tumors large enough to be seen on mammography, but missed due to density (these are the ones where we expect mortality reductions, the same as if the tumor had been seen on mammography), and 2) lowering the threshold of detection. The latter may prove to be unnecessary as there is likely a diminishing return with the discovery of smaller and smaller tumors. If an imaging method is developed in the future with a 5 mm mean tumor size with screening, then the calculated sensitivity of breast MRI et al. will drop accordingly. Still, it is important to make the distinction between the two effects in that improved detection rates of MRI and comparable imaging methods can be sometimes dismissed as based entirely on lowering the size threshold.

When five international MRI screening trials were subjected to a combined analysis, mammographic sensitivity was calculated at 40%, and when screening ultrasound was also performed in triple modality studies, sensitivity was 43%.[37] This created a significant

disconnect between the "80–90%" dogma and the new reality. Rather than redefining a more reasonable sensitivity level for mammography, the results of the MRI trials were considered applicable only to high-risk women. Risk levels, however, are not part of the sensitivity equation, so the explanation for 40% sensitivity of mammography lay in the presumption that the high-risk trials were skewed toward younger women and thus, higher levels of breast density must account for poor mammographic sensitivity.

The combined analysis of the international MRI screening trials also revealed sensitivity for MRI to be only 81%,[37] with several issues accounting for this relatively low sensitivity for MRI. Dynamic MRI was still used at the time in some European studies where the technology focused on improved specificity at the cost of sensitivity. A typical scenario at that time would be a calcium cluster identified on mammography, with biopsy showing DCIS, yet low resolution, dynamic MRI being unable to detect any abnormality.[38] With high spatial resolution MRI in use today, sensitivity is routinely determined at 90% or higher.[39] In fact, the true sensitivity of the technology alone is higher still, in that interpretive errors have been attributed to 31% of cancers missed by high-risk screening MRI.[40]

Although we do not have sensitivity data for MRI when screening the general population at baseline risk, we do have CDRs in this group that would indicate that the low sensitivity figures for mammography are not limited to high-risk patients.

In a unique study of 2120 average-risk women who underwent 3861 screening MRIs after negative mammograms (the majority with negative ultrasound as well),[41] the supplemental CDR was 15.5 per 1,000, with invasive cancers in a 2:1 ratio. After the prevalence screen with MRI (CDR = 22.6/1000), 13 cancers were then identified on subsequent MRI screening, with 12 of the 13 visible only on MRI. Median invasive tumor size was 8 mm, node-negative in 93.4%. Cancers were high-grade in 41.7% of cases at prevalence screening and 46.0% of cases at incidence screening.

Remarkably, there were *no interval cancers* encountered during the study, lending credence to the position that most so-called interval cancers are present (and detectable) on the prior screen. And by comparison to CDRs accomplished through general mammographic screening, e.g., 5 per 1,000, one faces the troubling conclusion that mammograms miss more cancers than they detect in general population screening (understanding that this effect is largely due to the lowered threshold of detection).

In spite of the striking differences in sensitivity between breast MRI and mammography, attempts to minimize this disparity are common. In the 2009 recommendation statement from the U.S. Preventive Services Task Force (U.S.P.S.T.F.),[42] mammography was stated to have a sensitivity of 77%–95%, drawing from their own 2002 policy statement,[7] while selectively offering the higher numbers found with prevalence screening only, well above the sensitivity levels found on long-term incidence screens. The recommendations go on to claim that screening MRI has a sensitivity of 71%–100%, this range having no appreciable difference when compared to the range provided for mammography. This 71%–100% was, in fact, the range encountered in the international MRI screening trials, but includes the widespread use of aforementioned dynamic (low sensitivity) technology no longer in use and ignores the 90%-plus sensitivity that was well-established by the time of this 2009 report.

Then, in 2015, the U.S.P.S.T.F. addressed the emerging awareness of breast density as a predictor of lower mammographic sensitivity.[43] However, since a driving principle of this public health organization is to insist on proven mortality reductions before endorsing a breast cancer screening modality, the U.S.P.S.T.F. issued a blanket "I" (insufficient evidence) recommendation for all imaging modalities beyond standard 2-D mammography. As an aside, they had previously balked on digital technology as well, but by the time of the next update, there was essentially no film screen technology left in the United States to endorse. The "I" designation went on to mention specifically 3-D tomosynthesis, ultrasonography, MRI, "or other methods in women identified to have dense breasts on an otherwise negative screening mammogram." When it comes to sensitivity, at least, the Task Force ended up endorsing arguably the worst imaging method available to detect breast cancer in high-density patients, especially those in the Level D group.

It has been previously recognized that while a mortality reduction is the ideal parameter upon which to base effective screening, it is largely impractical to carry out costly prospective, randomized trials to answer this question for many reasons, not the least of which is that the technology utilized will be obsolete by the time efficacy has been determined. In one of the most comprehensive studies of mammographic sensitivity,[4] the authors begin their discussion with this comment: "It is important to evaluate early detection trials as soon as possible without waiting for long-term mortality results. For this purpose, screening sensitivity can be used as an early indicator to assess the screening efficacy."

Conclusion

After many decades of heavy promotional efforts for screening mammography, a more accurate sensitivity level has become evident through multi-modality imaging. This redefinition of mammographic sensitivity at lower levels than once thought now poses a number of challenges with regard to patients, their physicians, screening guidelines, researchers and third-party payors. The recent adoption of 3-D tomosynthesis mammography certainly affords an improvement in sensitivity, but still falls well short of the cancer detection rates that are accomplished with contrast-enhanced technologies.

For patients and their physicians, the promotional efforts for mammography have imparted a false sense of security with a negative study, such that entertaining a second modality for periodic screening is believed unnecessary. Even in those states that have passed legislation that mandate insurers cover the cost for adjunct ultrasound screening in high density patients, the experience to date reveals that only one-fourth to one-third of eligible patients accept this option, with no evidence so far of improved compliance over time.[44]

When it comes to establishing guidelines, organizations that endorse an expanded role for multi-modality imaging, such as the Society of Breast Imaging, are largely discounted by insurers who seek so-called neutral sources for their policies. But guidelines from influential organizations can cast the issue in artificial light by comparing apples to oranges as seen above with the U.S. Preventive Services Task Force and their implication that sensitivity ranges for mammography and MRI are basically equivalent. As always, it begs the question: If mammograms are 80% sensitive, then how could it be possible that in head-to-head studies, MRI will detect double, or even triple, the number of cancers found by mammography? The math simply doesn't work.

Screening researchers are impacted as well, where funding can be a challenge already, given increasing skepticism about the value of breast cancer screening in general, and where harms have become the focal point. To propose a more sensitive tool for screening generates concerns that the associated problems, e.g., overdiagnosis, will be made worse Granted, we probably do not need new technologies to take us below the 8 mm mean size of invasive tumors discovered by MRI. At the same time, for larger tumors in dense breasts, a sensitivity of less than 50% should not be

acceptable either. The research challenge is to match these contrast-enhanced, physiology-based studies to the correct patients, efficiently finding those occult cancers that, today, can be easily detected with technology already available.

While high density and high risk are our only guideposts for multi-modality imaging at this point in time, this approach excludes many, if not most, of the patients headed toward breast cancer. Using current guidelines for adjunct imaging can appear to be cost-effective initially, based on the original reports where prevalence screens have high yields. But eventually, the steady state is reached with any modality, in which the lower-yield incidence screens match disease incidence. Thus, when used at fixed intervals, high-cost imaging loses its cost-effectiveness over time. It is possible that more efficient utilization of MRI and similar technologies can come through blood testing (high specificity ruling in adjunct imaging; high sensitivity tests ruling out),[45] or various approaches of artificial intelligence applied to normal screening mammograms,[46] turning fixed interval adjunct screening into more of a diagnostic study, as needed.

Rather than criticizing high-sensitivity imaging as an approach that will primarily exacerbate the harms of screening, an alternative view is that the modest mortality reductions ascribed to screening mammography have been accomplished by identifying *only half* of the detectable cancers, based on 40-year-old technology no longer in use. An obvious conclusion, after acceptance of this lower sensitivity value, is that breast cancer has a more vulnerable biology to early detection than we previously believed, and as such, by identifying those cancers missed by screening mammography, a major reduction in mortality could occur well above what is seen today with routine screening. But if one is grounded in the false belief that mammograms are "80–90 sensitive" across the board, there would be little gained through multi-modality approaches to close a small detection gap.

A single number to define mammographic sensitivity is meaningless, given all the caveats above. Even quoting a range can be flawed. The commonly stated, "80% sensitivity overall, with 50% or less for young women, or women with dense breasts" is certainly fair, but it treats the density issue as if it were an outlier. Roughly one-half of the population undergoing mammographic screening has a Level C or D density. Perhaps, we should be stating a specific sensitivity for each of the four density levels. Better yet, we should support those research efforts that attempt to individualize sensitivity wherein each patient is given her personal sensitivity level based on the quantitative and qualitative features on her mammogram.

Once systemic therapies are capable of eradicating nearly all breast cancers, we will no longer need to screen at all. But in the meantime, an era that might be measured in decades, we need to take advantage of the available adjunct imaging technologies, continuing to study how to use them in a cost-effective manner. There is no pressing need to find breast cancer *earlier* than what is already afforded by contrast-enhanced technologies – the task at hand is to find those cancers large enough to be detected, but are currently missed, on routine screening mammography.

Conflict of interest/Support statement

This study was supported in part by a grant from the National Cancer Institute,United States, NCI Grant R01CA197150 – "Increasing Cancer Detection Yield Using Breast MRI Screening Modality".

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amjsurg.2019.01.039.

References

1. Kern KA. The delayed diagnosis of breast cancer: medicolegal implications and risk prevention for surgeons. *Breast Dis*. 2001;12:145–158.
2. Egan R. Fifty-three cases of carcinoma of the breast, occult until mammography. *AJR*. 1962;88:1102–1108.
3. Shapiro S, Venet W, Strax P, Venet L. Current results of the breast cancer screening randomized trial: the health insurance plan (HIP) of greater New York study. In: Day NE, Miller AB, eds. *Screening for Breast Cancer*. Toronto: Hans Huber; 1988:3–15.
4. Shen Y, Zelen M. Screening sensitivity and sojourn time from breast cancer early detection clinical trials: mammograms and physical examinations. *J Clin Oncol*. 2001;19:3490–3499.
5. Baines CJ, Miller AB, Bassett AA. Physical examination. Its role as a single screening modality in the Canadian National Breast Screening Study. *Cancer*. 1989;63:1816–1822.
6. Bevers TB. Flaws in CNBSS are vast, impact on screening recommendations is nil. *The ASCO Post*. 2014;5(6).
7. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2002;137:347–360.
8. Cunningham MP. The breast cancer detection demonstration Project 25 years later. *Ca - Cancer J Clin*. 1997;47:131–133.
9. Feig SA, Shaber GS, Patchefsky A. Analysis of clinically occult and mammographically occult breast tumors. *AJR Am J Roentgenol*. 1977;128:403–408.
10. Sickles EA. The subtle and atypical mammographic features of invasive lobular carcinoma. *Radiology*. 1991;178:25–26.
11. Hollingsworth AB, Taylor LDH, Rhodes DC. Establishing a histologic basis for false-negative mammograms. *Am J Surg*. 1993;166:643–647.
12. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Canc*. 2017 Apr 13;3:12. https://doi.org/10.1038/s41523-017-0014-x. eCollection 2017.
13. Collett K, Stefansson IM, Eide J, et al. A basal epithelial phenotype is more frequent in interval breast cancers compared with screen detected tumors. *Cancer Epidemiol Biomark Prev*. 2005;14:1108–1112.
14. Lekanidi K, Dilks P, Suaris T, et al. Breast screening: what can the interval cancer review teach us? Are we perhaps being a bit too hard on ourselves? *Eur J Radiol*. 2017 Sep;94:13–15. https://doi.org/10.1016/j.ejrad.2017.07.005. Epub 2017 Jul 12.
15. Sala M, Domingo L, Louro J, et al. Survival and disease-free survival by breast density and phenotype in interval breast cancers. *Cancer Epidemiol Biomark Prev*. 2018 May 31. https://doi.org/10.1158/1055- 9965.EPI-17-0995. pii: cebp.0995.2017.
16. American Cancer Society. *Cancer Facts & Figures 2007*. Atlanta, GA: American Cancer Society; 2007:9.
17. American Cancer Society. *Cancer Facts & Figures 2015*. Atlanta, GA: American Cancer Society; 2015:10.
18. Burnside ES, Sickles EA, Bassett LW, et al. The ACR BI-RADS® experience: learning from history. *J Am Coll Radiol*. 2009;6:851–860.
19. Mann BD, Giuliano AE, Bassett LW, et al. Delayed diagnosis of breast cancer as a result of normal mammograms. *Arch Surg*. 1983;118:23–24.
20. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. *AJR*. 1976;126:1130–1139.
21. Ursin G, Hovanessian-Larsen L, Parisky YR, et al. Greatly increased occurrence of breast cancers in areas of mammographically dense tissue. *Breast Cancer Res*. 2005;7:605–608.
22. Bigenwald RZ, Warner E, Gunasekara A, et al. Is mammography adequate for screening women with inherited BRCA mutations and low breast density? *Cancer Epidemiol Biomark Prev*. 2008;17:706–711.
23. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*. 2005;353:1773–1783.
24. Pisano ED, Hendrick RE, Yaffe MJ, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. *Radiology*. 2008;246:376–383.
25. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*. 2013;267:47–56.
26. Ciatto S, Houssami N, Bernardi D, et al. Integration of 3-D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. *Lancet Oncol*. 2013;14:583–589.
27. Lang K, Andersson I, Rosso A, et al. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study. *Eur Radiol*. 2016;26:184–190.
28. Access: https://clinicaltrials.gov/ct2/show/NCT03233191.
29. Michaelson JS, Silverstein M, Wyatt J, et al. Predicting the survival of patients with breast carcinoma using tumor size. *Cancer*. 2002;95:713–723.
30. Kolb TM, Lichy J, Newhouse JH. Occult cancer in women with dense breasts: detection with screening US – diagnostic yield and tumor characteristics. *Radiology*. 1998;207:191–199.

31. Nickson C, Kavanagh A. Tumor size at detection according to different measures of mammographic breast density. *J Med Screen*. 2009;16:140–146.
32. Berg WA, Zhang Z, Lehrer D, et al. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *J Am Med Assoc*. 2012;307:1394–1404.
33. Tagliafico AS, Calabrese M, Mariscotti G, et al. Adjunct screening with tomosynthesis or ultrasound in women with mammography-negative dense breasts: interim report of a prospective comparative trial. *J Clin Oncol*. 2016;34:1882–1888.
34. O'Connor M, Rhodes D, Hruska C. Molecular breast imaging. *Expert Rev Anticancer Ther*. 2009;9:1073–1080.
35. Berg WA, Weinberg IN, Narayanan D, et al. High-resolution fluorodeoxyglucose positron emission tomography with compression ("positron emission mammography") is highly accurate in depicting primary breast cancer. *Breast J*. 2006;12:309–323.
36. Fallenberg EM, Schmitzberger FF, Amer H, et al. Contrast-enhanced spectral mammography vs. mammography and MRI – clinical performance in a multi-reader evaluation. *Eur Radiol*. 2017;27:2752–2764.
37. Sardanelli F, Podo F. Breast MR imaging in women at high-risk of breast cancer. Is something changing in early breast cancer detection? *Eur Radiol*. 2007;17:873–887.
38. Harms SE. The use of breast magnetic resonance imaging in ductal carcinoma in situ. *Breast J*. 2005;11:379–381.
39. Hillman BJ, Harms SE, Stevens G, et al. Diagnostic performance of a dedicated 1.5-T breast MR imaging system. *Radiology*. 2012;265:51–58.
40. Vreemann S, Gubern-Merida A, Lardenoije S, et al. The frequency of missed breast cancers in women participating in a high-risk MRI screening program. *Breast Canc Res Treat*. 2018;169:323–331.
41. Kuhl CK, Strobel K, Bieling H, et al. Supplemental breast MR imaging screening of women with average risk of breast cancer. *Radiology*. 2017;283:361–370.
42. US Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2009;151:716–726.
43. Nelson HD, Fu R, Cantor A, et al. 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;164:244–255.
44. Weigert JM. The Connecticut experiment; the third installment: 4 years of screening women with dense breasts with bilateral ultrasound. *Breast J*. 2016;23:34–39.
45. Hollingsworth AB, Pearce MR, Stough RG. Modeling the impact of a screening blood test on the use of adjunct breast imaging. *Breast J – accepted for publication June*. 2018;25.
46. Heidari M, Khuzani AZ, Hollingsworth AB, et al. Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. *Phys Med Biol*. 2018 Jan 30;63(3):035020. https://doi.org/10.1088/1361-6560/aaa1ca.