Breast Cancer Screening in a Multimodality Environment—The Need for a Simple Summary Measure of Marginal Value

David Gur, ScD

In a rapidly changing clinical environment, assessment of imaging-based technologies and practices for periodic screening for the early detection of breast cancer is constrained by cost, complexity, and professional resources, particularly concerning supplementary imaging of subgroups constituting a large fraction of the screened population. Relatively high survival rates after detection make it extremely difficult to adequately assess marginal values of proposed approaches either before the technology in question being widely accepted and used or before it becomes largely obsolete. The author discusses several issues related to the assessment process and proposes the use of a surrogate summary measure of performance for this purpose, namely the number of recalled cases for the diagnostic workup of suspicious findings during repeat examinations, per one additional screen detected cancer that is invasive, node-negative, and classified grade 2 or above.

Key Words: Breast cancer; screening; technology and practice assessment; marginal value.

In recent years, there have been a number of new technologies and practices aimed specifically at improving the early detection of breast cancer, particularly for women with dense breast tissue. Several modalities have been discussed for possible use as primary and/or supplementary procedure(s) for screening the general population and large subgroups of women (1–6). Although magnetic resonance imaging (MRI) is currently used for screening women at high risk (6), screening women with dense breast tissue is clearly the focus of many of these new procedures, because these women are considered to be at intermediate risk. These procedures include, but are not limited to, hand-held and/or automated whole breast ultrasound (WBUS) (1), digital breast tomosynthesis (DBT) (2), molecular breast imaging (MBI) (3), cone beam computed tomography (CBCT) (4), and contrast-enhanced mammography (CEM) (5). In most instances, when a modality supplements a current practice (or simply the addition of more images or views using one modality), cancer detection improves because the information ascertained from the current and added modality is not totally correlated–namely, only when the information generated by two modalities (or more) is totally (100%) correlated, no diagnostic improvements are expected by using both modalities, other than possible improvements related to multiple observers interpreting the same case (namely, gaining from interobserver variability). Findings by independent observers are rarely 100% correlated with each other, even when the abnormality in question is quite obvious.

In breast cancer screening, new modalities are frequently viewed as supplementary to mammography rather than as replacements (1–6). Hence, the cost and often the complexity of these practices are ever increasing. Clearly, this approach is nonsustainable. Clinicians and investigators should focus on assessing the marginal value of these additional procedures, regardless of any other performance measure considerations, such as cost, professional resources, recall rate, and participants' anxiety.

Several fundamental concepts should be examined before considering the topic of assessing the true value in general, and a marginal value in particular. First, it is always possible to successfully compete with (or "beat") poorly performing current (or reference) practices (7). For example, it would clearly be easier to demonstrate an improvement in cancer detection when supplementing screen film mammography with WBUS in women with dense breasts than when supplementing full field digital mammography (FFDM) based screening. However, because FFDM based screening remains a "less than perfect" procedure, particularly in women with dense breasts, supplementing FFDM based screening with any of the proposed approaches (e.g., WBUS, DBT, MBI,

Acad Radiol 2014; 21:325–328

From the Department of Radiology, Imaging Research, University of Pittsburgh, 3362 Fifth Ave, Pittsburgh, PA 15213-3180. Received October 11, 2013; accepted November 12, 2013. **Address correspondence to:** D.G. e-mail: gurd@upmc.edu

http://dx.doi.org/10.1016/j.acra.2013.11.010

CEM or CBCT) would likely result in detecting additional cancers. The question is always at what cost?

Second, when considering adding new modalities/procedures to current practices for subgroups constituting a large fraction of the screened population (e.g., women with breast density 3 and 4 based on Breast Imaging Reporting and Data System [BIRADS], constituting between 40%–50% of the female population between the ages of 40 and 69), practical/operational issues and financial issues must be carefully considered. In addition, supplementary modalities may increase risk because of additional radiation exposure or repeated injections of contrast media; however, for the purpose of this discussion, I assume that all approaches in question can be considered "safe".

Third, the slope of incremental performance improvements per unit cost (or additional effort) typically has a negative second derivative resulting in incrementally diminishing returns–namely, adding new procedures to an already accepted group of prior procedures tends to result in a decreasing rate of incremental benefits. At the extreme, when the practice is "perfect" in terms of diagnostic performance, adding supplementary procedures can only add "noise" (and cost); thereby, actually decreasing overall performance without any chance of improving performance.

Fourth, frequently the addition of supplementary procedures to current practices increases not only cancer detection rates, but also false-positive rates in actually "negative" women or those with benign findings. Women with false-positive findings constitute the vast majority of all women with "positive findings" in part because breast cancer has relatively low prevalence in the screening population.

Fifth, during subsequent screening ("incidence years") with a new approach, marginal values may increase or decrease, but typically because of the lower false-positive rates of the current practice (without the supplemental imaging) and the lower detection rates due to prior detection of a large fraction of prevalence cancers, the marginal value is more likely to decrease, unless one assumes that after the introduction one can persistently have higher detection rates for many years. This type of a finding eventually violates/defies some fundamental biological concepts not discussed here and should always be questioned unless convincingly demonstrated experimentally–namely, one expects that after an "introductory" or a "transition" period (typically 2–4 years) cancer detection rates with the new and improved approach will largely return to the preintroduction rates. Hence, performance data on the impact of the new practice beyond the initial implementation period are needed. In screening for the early detection of breast cancer, performance data for a minimum of three to five subsequent screens are needed for this purpose. Otherwise, we may overestimate substantially the actual marginal value.

Sixth, when observers become an integral part of the diagnostic system, one cannot always separate the interpreter's performance from the system's performance. Unfortunately, interobserver variability typically constitutes the largest component of all factors contributing to total variance in performance measures. Hence, whatever validation approach is used for comparing different practices, a large number of participants (radiologists) from different types of practices should be included in the experiment, and assessments of marginal values should account for individualized performance levels. Actually, one interesting measure of performance similar to analysis of variance could be the change, if any, in inter-reader variability (or disagreement rates) when comparing two approaches, but this topic is beyond the scope of this article.

And last, when assessing the value of supplementary modalities to a screening program, it is important to assess marginal value, if any, not only in terms of absolute numbers but also (and perhaps more important) in terms of the types of additionally detected cancers. However, in this discussion I will focus primarily on the former, namely detection rates, as the topic of the types of cancers that may be more important to find earlier in terms of overall societal benefit is quite complex.

Optimally, one would hope to focus on assessing the value of screening in terms of reducing mortality or morbidity, or in terms of total lifetime management costs as a result of more effective treatments/management of cancers that if left alone and/or detected later would have resulted in death, additional morbidity, or a sizable cost in managing the disease. Unfortunately, these cancers are frequently not known a-priori. Defining and finding the so called "killer cancers", as opposed to cancers that could result in "over diagnosis", is a complex topic that is extremely difficult to study rigorously, particularly in a society that is reluctant to not treat all potentially harmful findings, thereby perform long term studies on the natural history of different abnormal breast findings.

Unlike diagnostic procedures that address a specific need to assess a specific medical question, periodic screening is a repeated procedure (individualized or not) that is designed primarily for detecting abnormalities at an early stage under the reasonable, and in many instances validated, assumption that the screening program in question is indeed beneficial to the individual and to the society. However, these benefits may come with some potential for harm as well (8,9)

Because of the nature of lung cancer, namely the relatively rapid growth in many instances and the associated high mortality when detected at a later stage, it is relatively easy to demonstrate mortality reduction by computed tomography (CT) screening of high-risk individuals (e.g., heavy smokers) for the early detection of lung cancer. Data in support of such screening are quite compelling and validated in several large independent studies (10–12). Hence, the recent task force recommendation is both appropriate and timely (13).

This is not the case in screening for early detection of prostate or breast cancer. The relatively long survival of most patients postdetection of breast cancer combined with a rapidly changing environment with many therapeutic advances for different types of breast cancers at all stages, makes it largely impossible to assess without bias attributable mortality related marginal value of supplemental imaging approaches regardless