First noted in 1948 by Johannes Clemmesen from the Danish Cancer Registry (1), patterns of female breast cancers in developed countries are consistent with a “mixture model” with at least two main parts (2–5). The first is largely premenopausal with peak incidence near age 50 years; the second is largely postmenopausal with a peak around age 70 years (6–8). Using the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) 17 Registries Database (covering approximately 28% of the US population), Howlader et al. (9), in this issue of the Journal, provide additional molecular insight into breast cancer heterogeneity by studying newly available data—specifically, the recent addition of human epidermal receptor 2 (HER2) to the list of breast tumor characteristics captured by the SEER program.

Prior laboratory work had strongly suggested that breast cancers are molecularly heterogeneous. Indeed, unsupervised hierarchal clustering of gene expression profiles of various tumor collections has demonstrated four to 10 molecular subtypes of breast cancer (10,11). Of these various classifications, the four so-called “intrinsic” subtypes are the best-known and most well-characterized molecular signatures (10,12,13). The four intrinsic types consist of two main classes by hormone receptor (HR) status (estrogen receptor [ER] and/or progesterone receptor [PR]) and epithelial cell of origin (luminal and basal-like). There are two HR-positive subtypes (luminal A and luminal B) and two HR-negative subtypes (HER2-enriched [HER2E] and basal-like).

Fortunately for cancer surveillance researchers, the intrinsic molecular subtypes can be approximated by the joint protein expression of HR± and HER2± (14,15). Howlader et al. (9) describe for the first time the demographic characteristics, tumor features, and area-level poverty data for the four recapitulated subtypes in the entirety of SEER 17, summarized by us in Figure 1. Seventy-three percent of the breast cancers were luminal A (HR+/HER2−), 10% were luminal B (HR+/HER2+), 5% were HER2E (HR−/HER2+), and 12% were triple-negative (HR−/HER2−). SEER began in 1973 and has recorded HR expression since 1990 but did not collect HER2 data until 2010 (the most recent year for which complete SEER data are available). For the past two decades, the HR (especially ER) status has been an epidemiologically useful correlate for population-based breast cancer heterogeneity. So, it might be asked if the new HER2 data add much to HR status. The answer is a definite yes.

Overall, approximately 15% of cases were HER2-positive (Figure 1). Importantly, what Howlader and colleagues (9) have been able to do is split the HR-positive group into two pieces and the HR-negative group into two pieces. We now know that almost one in three HR-negative tumors coexpress HER2 (5 / [5 + 12] or 29% of HR− tumors) (Figure 1). Conversely, approximately one in eight HR-positive tumors also expresses HER2 (10/ [10+73] or 12% of HR+ tumors, Figure 1). Earlier clinical trial data had suggested that 15%–30% of all invasive breast cancers were HER2-positive (16,17). Furthermore, the trial data also suggested that half of the HER2-positive tumors were HR-positive (16), when in fact the true figure in the population is actually two-thirds (10% / [10%+5%]) (Figure 1). The initial estimates were largely obtained from high-risk breast cancer cohorts in randomized clinical trials (18,19), which we now know are substantially different from those in the general population.

**Figure 1.** Frequency distribution of the four intrinsic breast cancer subtypes in the National Cancer Institute’s Surveillance, Epidemiology, and End Results 17 Registries database, covering approximately 28% of the US population. The four mutually exclusive molecular subtypes are defined by the joint positive or negative expression of the hormone receptors (HRs) and human epidermal growth factor 2 receptor (HER2). Seventy-three percent of the breast cancers were luminal A (HR+/HER2−), 10% were luminal B (HR+/HER2+), 5% were HER2-enriched (HER2E; HR−/HER2+), and 12% were triple-negative (HR−/HER2−).
HER2+ expression also shows that most of the HR-positive breast cancers in SEER are luminal A, whereas the majority of the HR-negative breast cancers are triple-negative (Figure 1). HR-positive tumors have been associated with white race, older age at diagnosis, and low tumor grade, which we now see mostly reflects luminal A expression (HR+/HER2−). HR-negative tumors have been associated with black race, younger age at diagnosis, and high tumor grade, which we now know reflects a greater frequency of triple-negative breast cancers. Among other things, this perspective will likely be useful for a better understanding and monitoring of the black-to-white racial disparity in breast cancer mortality (20,21).

As the HER2 data in SEER mature, joint HR and HER2 expression will be used to monitor breast cancer incidence, survival, and mortality rates. For these important studies, it is crucial that the statistical analyses not repeat a serious mistake of the past, whereby nonrandom missing data patterns for the tumor markers were simply overlooked. We have developed simple inverse-probability weighting methods when the rates are of interest (22), and Howlader and colleagues (9) have developed multiple imputation methods for analysis of both rates and individual-level data (23).

Adding HER2 data represents an important step forward for the SEER program. One important question is how SEER should decide whether to invest resources to capture a tumor characteristic such as HER2. In hindsight, one could argue that HER2 expression became relevant nearly 30 years ago with the demonstration that prognosis was worse for HER2-positive than HER2-negative cancers (24,25). However, there are many such specific and general prognostic markers for numerous types of cancers. We would suggest that SEER should prioritize for rapid uptake those markers that prove to be both predictive [a determinant of the response to a targeted treatment (26)] and prognostic, as is the case for HER2.

Going forward, cancer registries are singularly positioned to establish the population-burden of new tumor markers as well as the historic trends in that burden, especially when there is access to historic biospecimens (27–29). For example, data from SEER’s Residual Tissue Repositories established the existence of a hitherto unobserved “epidemic” of human papillomavirus (HPV)-positive oropharyngeal cancers since the 1980s (29). It also demonstrated that those with HPV-positive oropharyngeal cancers had improved 20-year survival vs those with HPV-negative oropharyngeal cancers (29). Although clinical trial databases also contain biospecimens with extensive annotations, registry data necessarily represent the US population (30), long-term survival over decades can be examined, and survival data predate the introduction of treatments, allowing examination of the impact of treatments on the nationwide burden of disease. One important question for the field is whether such collections based on SEER data should be enlarged. Howlader’s data (9) certainly illustrate the enormous potential of molecular data to identify new temporal trends and etiological clues, track the burden of cancer, characterize survival trends and patterns in the population, and shed light on cancer disparities.

References


jnci.oxfordjournals.org


**Funding**

This research was supported by the Intramural Research program of the National Institutes of Health/National Cancer Institute.

**Notes**

The authors have no conflicts of interest to declare.

**Affiliation of authors:** Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD (WFA, FSR, HAK).