How Many Etiological Subtypes of Breast Cancer: Two, Three, Four, Or More?

William F. Anderson, Philip S. Rosenberg, Aleix Prat, Charles M. Perou, Mark E. Sherman

Manuscript received August 15, 2013; revised May 1, 2014; accepted May 13, 2014.

Correspondence to: William F. Anderson, MD, MPH, National Cancer Institute, Division of Cancer Epidemiology and Genetics, Biostatistics Branch, 9609 Medical Center Dr, Rm 7E144, Bethesda, MD 20892–9780 (e-mail: wanderso@mail.nih.gov).

Breast cancer is a heterogeneous disease, divisible into a variable number of clinical subtypes. A fundamental question is how many etiological classes underlie the clinical spectrum of breast cancer? An etiological subtype reflects a grouping with a common set of causes, whereas a clinical subtype represents a grouping with similar prognosis and/or prediction. Herein, we review the evidence for breast cancer etiological heterogeneity. We then evaluate the etiological evidence with mRNA profiling data. A bimodal age distribution at diagnosis with peak frequencies near ages 50 and 70 years is a fundamental characteristic of breast cancer for important tumor features, clinical characteristics, risk factor profiles, and molecular subtypes. The bimodal peak frequencies at diagnosis divide breast cancer overall into a “mixture” of two main components in varying proportions in different cancer populations. The first breast cancer tends to arise early in life with modal age-at-diagnosis near 50 years and generally behaves aggressively. The second breast cancer occurs later in life with modal age near 70 years and usually portends a more indolent clinical course. These epidemiological and molecular data are consistent with a two-component mixture model and compatible with a hierarchal view of breast cancers arising from two main cell types of origin. Notwithstanding the potential added value of more detailed categorizations for personalized breast cancer treatment, we suggest that the development of better criteria to identify the two proposed etiologic classes would advance breast cancer research and prevention.


Clinically, breast cancer is widely recognized as a heterogeneous disease. In this commentary, we focus instead upon the etiological perspective of breast cancer heterogeneity. By etiological heterogeneity, we mean breast cancer subtypes (components, classes, or groupings) that share common sets of causes. This is distinct from a clinical subtype, which refers to tumors with common prognostic characteristics and/or predictive features (response to targeted-treatment) (1).

As efforts proceed to improve the taxonomy of clinical breast cancer, a fundamental question persists; how many etiological subtypes actually exist? Clinical taxonomic systems have defined multiple classes in an effort to optimize therapeutic management. At its extreme, this approach translates into “precision” or “personalized” medicine, with a view that each person’s tumor is unique.

We propose a more parsimonious view for breast cancer etiology, which we believe is consistent with a hierarchical view of breast cancer derived from two main cell types of origin (2–9). We show that any given breast cancer molecular or clinical category demonstrates a mixture of two stereotypical age-specific incidence patterns with bimodal peak frequencies near ages 50 years and 70 years. We hypothesize that the consistency of this pattern supports a two-component mixture model, where different molecular and/or clinical categorizations represent variable combinations of two etiological subtypes. In this model, it is the difference in the relative distributions of the two putative subtypes that endows any given breast cancer categorization with its distinguishable biological features.

Historical Developments in the Understanding of Breast Carcinogenesis and Pathogenesis From an Epidemiological Perspective

Multistage (Log-Linear) Cancer model

More than 50 years ago, Armitage and Doll noted that cancer rates rise exponentially with advancing age for a number of epithelial malignances (10–12), thus providing the theoretical foundation for multistage tumor initiation, promotion, and progression (13–15). An epidemiological prediction of the multistage cancer model is a log-linear (or log-additive) relationship between cancer incidence and chronological age with a linear (or steady) rise in the logarithm of cancer rates as a function of the logarithm of age at diagnosis.

However, breast cancer incidence does not demonstrate an exponential increase with advancing age. Incidence rates slow before age 50 years (Figure 1A). The change point in incidence has been termed “Clemmesen’s hook” after its discoverer, Johannes Clemmesen (16,17). Clemmesen’s hook is characteristic of female breast cancers worldwide (18–20), coincides with the female climacteric, and is not found among male breast cancers (21).
To account for the distinctive incidence rate pattern for breast cancer, Pike et al. introduced the concept of “breast tissue age” as a better marker of risk than chronological age (22). In his model, key reproductive events affect the shape of the age-specific incidence rate curve. Specifically, risk factors accelerate and protective factors attenuate breast tissue aging, with Clemmesen’s hook representing the net effect. Others have refined the Pike model (23–26); however, a key feature of all of these models is that all breast cancers share a common pathogenesis, reflected in a single, age-specific incidence rate curve.

Two-Component Cancer Model
Lilienfeld (27) and de Waard (28–30) pioneered the concept that breast cancers develop by two distinct pathways (rather than one), each with a different age-specific incidence rate curve. The first pathway results in mainly premenopausal tumors with peak occurrence early in life, similar to estrogen receptor (ER)-negative cancers in the general US population (Figure 1A). The second pathway results in predominantly postmenopausal cancers with peak incidence later in life, similar to late-onset ER-positive cancers (Figure 1A). In this model, Clemmesen’s hook can be seen as the confluence or superimposition of the age-specific incidence rate curves for the early-onset and late-onset subtypes of breast cancer (31,32).

Descriptive Epidemiology
Age-Specific Incidence Rates
The menopause transition was predicted to affect incidence rates of ER-positive more than ER-negative breast cancers, given the presumed greater role of sex-steroid hormones in the pathogenesis of hormone sensitive cancers (33,34). Nonetheless, and somewhat paradoxically, menopause [or rather its surrogate, age 50 years (35,36)] is associated with greater impact upon ER-negative than ER-positive cancers (31,33,34) (Figure 1A). ER-negative rates rise rapidly early in life then flatten or fall soon after menopause (37), whereas ER-positive rates rise continuously irrespective of menopause, albeit more slowly after age 50 years.

The different incidence rate patterns by ER expression represent an age interaction or effect modification. Under the null hypothesis of no interaction, the age-specific incidence rates for ER-positive and ER-negative cancers would be parallel on the log scale (38–40), yielding a constant incidence rate ratio (IRR) irrespective of age at diagnosis (IRR$_{ERneg \text{ to } ERpos}$ = constant for all ages). We also make a distinction between quantitative (noncrossover) and qualitative (crossover or reversing) age interactions (37–40). A quantitative age interaction varies in magnitude but not direction, whereas a qualitative interaction differs in both magnitude and direction. At the extreme ages in Figure 1A, the incidence rate ratio of ER-negative to ER-positive cancer is 2.3 during ages 20 to 24 years (IRR$_{ERneg \text{ to } ERpos}$ > 1.0) and the incidence rate ratio of ER-negative to ER-positive cancer is 0.10 during the ages 80 to 84 years (IRR$_{ERneg \text{ to } ERpos}$ < 1.0). True qualitative or reversing interactions are considered rare (40–42) but, when found, can be statistical surrogates for age-dependent etiological heterogeneity. In the context of a qualitative age interaction, Clemmesen’s hook can be further seen as the crossover in falling ER-negative rates and rising ER-positive rates.

Bimodal Age Distributions at Diagnosis
If breast cancer followed the log-linear incidence curve described by Armitage and Doll (10–12), breast cancer cases in the general population would be predicted to show a unimodal age distribution at diagnosis. In contrast, breast cancer overall demonstrates a bimodal pattern, with the modal ages near 50 and 70 years representing the central tendencies for the early-onset and late-onset breast cancers (Figure 2A). Density plots are constructed with 1-year increments using a “smoothing” method for the corresponding age distributions at diagnosis (31,43), where the area under the curve includes all of the breast cancer cases in a given population.

Crossing ER-positive and ER-negative age-specific rates (Figure 1A) also shows bimodal age distributions at diagnosis.
Theoretically, divergent ER trends in the United States could have resulted from statistical anomalies (53), and/or the implementation of lower thresholds for classifying ER tests as positive (46–48). However, secular trends in other countries with better control of these potentially confounding factors (eg, Denmark) showed similar trends as in the United States (53), consistent with the hypothesis that divergent ER-positive and ER-negative trends might be due to changes in different risk factor profiles by ER subtype over time (47,48,54).

Age-Adjusted Secular Trends
Recent studies in the United States show an unexpected divergence of ER-positive and ER-negative breast cancer trends (46–48); ER-positive cancers have risen over the long term, whereas ER-negative breast cancers have declined (Figure 1B). Theoretically, divergent ER trends in the United States could have resulted from statistical anomalies (49), changes in assay methodology or application of lower thresholds for classifying ER tests as positive (50), and/or the implementation of organized screening mammography (51,52). However, secular trends in other countries with better control of these potentially confounding factors (eg, Denmark) showed similar trends as in the United States (53), consistent with the hypothesis that divergent ER-positive and ER-negative trends might be due to changes in different risk factor profiles by ER subtype over time (47,48,54).

Biostatistical Models
Our group has leveraged the application of biostatistical models to complement descriptive epidemiology. First, we used two-component mixture models to determine whether bimodal age distributions at diagnosis fitted the data better than a single density (55) (Figure 2, A and B). Then, to confirm qualitative age interactions (eg, Figure 1A), we used age–period–cohort (APC) models to evaluate age-specific effects independent of calendar-period effects (that relate to screening, changing diagnostic and/or practice patterns) and/or birth-cohort effects (generational and/or exposure factors) (56,57).

Two-Component Mixture Models
We initially identified three distinct age-specific incidence rate patterns that were closely associated with seven histopathological breast cancer subtypes (58). Incidence rates of infiltrating duct, tubular, and lobular carcinomas rose rapidly until age 50 years, then increased more slowly, similar to rates for breast cancer overall. Rates for medullary and inflammatory breast carcinomas increased rapidly until age 50 years, then flattened or fell, similar to ER-negative rates (Figure 1A). Finally, rates for papillary and mucinous carcinomas increased steadily with age, similar to ER-positive rates (Figure 1A) and much like cancers at many other organ sites such as colorectal cancer (12).

Notwithstanding the three distinct incidence rate patterns, two-component mixture models demonstrated that six of the histopathological subtypes had bimodal age distributions at diagnosis with a dominant late-onset mode near age 70 years and a minor mode around age 50 years. Density plots for ER-positive tumors show bimodal age distributions at diagnosis with a dominant late-onset mode near age 70 years and a minor mode around age 50 years. The risk for breast cancer–specific death can be expressed as an annual hazard rate, which describes the instantaneous rate of dying from breast cancer in a specified time interval (ie, percentage dying per year) after diagnosis among women who are alive at the beginning of that time interval. Nonparametric hazard function estimators were applied that modeled the hazard profile of ER-positive and ER-negative cancers, allowing both the shape and magnitude to be estimated free of ad hoc mathematical assumptions. Specifically, the hazard rate curves were generated using cubic splines with joinpoints selected by Akaike's information criteria and 95% confidence intervals applied with bootstrap resampling (187–189). Bimodal age distributions at diagnosis among women (B) are associated with two very different cancer-specific outcomes. ER-negative hazards for breast cancer death peak near 7.5% per year approximately 2 years after initial diagnosis and then decline rapidly. ER-positive hazards lack a sharp peak but are relatively constant at 1% to 2% per year. Falling ER-negative and constant ER-positive hazards cross over approximately 8 years after breast cancer diagnosis.

Figure 2. Invasive female breast cancer case data were obtained from the National Cancer Institute's Surveillance, Epidemiology, and End Results 9 Registries Database from 1990 through 2010 database overall and for estrogen receptor (ER)–positive and ER-negative cancers. Bimodal breast cancer populations have fluctuated over time likely because of complex interactions between age-related biologic, risk factor, and screening phenomena, as previously described (186). For illustration, this figure has been restricted to the 1995 to 1998 period, during which a bimodal female breast cancer population was evenly distributed between early-onset and late-onset subtypes. Age distributions at diagnosis (or density plots) with 95% confidence intervals were constructed in 1-year age increments using a kernel density estimator applied to the corresponding age-at-diagnosis frequency histogram. The area under the curve represents 100% of the cancer records. The vertical axis shows the smoothed distribution (or proportion) with the frequency value x 100 = percentage distribution. A) Density plot for breast cancer overall demonstrates a bimodal age distribution at diagnosis with the modal ages near 50 and 70 years representing the central tendencies for early-onset and late-onset breast cancers. B) Density plot for ER-negative tumors also shows a bimodal age distribution at diagnosis with a dominant early-onset mode near age 50 years and a minor mode around age 70 years. Density plot for ER-positive tumors shows bimodal age distributions at diagnosis with a dominant late-onset mode near age 70 years and a minor mode around age 50 years. C) The risk for breast cancer–specific death can be expressed as an annual hazard rate, which describes the instantaneous rate of dying from breast cancer in a specified time interval (ie, percentage dying per year) after diagnosis among women who are alive at the beginning of that time interval. Nonparametric hazard function estimators were applied that modeled the hazard profile of ER-positive and ER-negative cancers, allowing both the shape and magnitude to be estimated free of ad hoc mathematical assumptions. Specifically, the hazard rate curves were generated using cubic splines with joinpoints selected by Akaike's information criteria and 95% confidence intervals applied with bootstrap resampling (187–189). Bimodal age distributions at diagnosis among women (B) are associated with two very different cancer-specific outcomes. ER-negative hazards for breast cancer death peak near 7.5% per year approximately 2 years after initial diagnosis and then decline rapidly. ER-positive hazards lack a sharp peak but are relatively constant at 1% to 2% per year. Falling ER-negative and constant ER-positive hazards cross over approximately 8 years after breast cancer diagnosis.
stereotypical ages of 50 and 70 years (45). The one notable exception was medullary carcinoma, which showed a unimodal age distribution with mode close to age 50 years. Medullary carcinomas are linked to the loss of BRCA1 function (59–61), which we propose may represent the closest known approximation to an etiologically pure early-onset subtype of breast cancer.

Similar to ER-positive and ER-negative breast cancers (Figure 2B), the bimodal peaks for different histopathological categories do not sharply divide cancers into pure groups, but rather reflect central tendencies of what we propose are two fundamental etiological classes. Breast cancers that develop at extreme ages are likely to be highly enriched for one etiological class, but both of these classes span the entire lifespan, with substantial mixing during the middle years, when many cancers occur. The mixing fraction or proportion of each etiological group varies within a class of breast cancers depending on its definition; however, the peak ages remain near ages 50 and 70 years.

**APC Models**

A useful APC function is the fitted (or longitudinal) age-specific incidence rate curve (37,62) (Figure 1A). The fitted curve stitches together the age-specific incidence rates from a collection of birth cohorts, each one observed over a limited and variable age span (ie, younger cohorts are observed at younger ages and older cohorts at older ages). The resulting curve estimates the age-specific rates of the middle or reference cohort over the entire age range. In contrast with the typical cross-sectional, age-specific incidence rate curve that may be confounded by period and cohort effects (63,64), the fitted curve is conditioned upon cohort and adjusted for period changes.

Clemmesen’s menopausal hook for breast cancer overall was once dismissed as a birth-cohort artifact (65), where the progressive increase in breast cancer risk from one generation to the next gave the appearance of falling incidence rates among older persons (64). This view is refuted by APC models (37), which demonstrate that the Clemmesen’s phenomenon is a true age-related event that persists in the fitted age-specific incidence rate curve (Figure 1A). Similar modeling approaches confirm that the qualitative age interaction remains for the fitted curves for ER-positive and ER-negative breast cancers, respectively (37).

**Analytic Epidemiology**

Consistent with descriptive epidemiology and biostatistical models, analytic epidemiology also supports a two-component breast cancer mixture model based upon the identification of major risk factors and genetic susceptibility.

**Risk Factor Epidemiology**

Despite the well-established protective effect of full-term pregnancy for breast cancer overall (22,66,67), it has been suggested, albeit without total agreement (68,69), that parity is associated with an early increase in risk, followed by long-term protection (70–86). Alternatively, the early risk and late protection for parity could reflect another qualitative (crossover or reversing) age interaction for breast cancer (40), where parity increases risk of early-onset subtypes and reduces risk of late-onset ones (37). Indeed, parity and multiple live births are associated with reduced risk for ER-positive and luminal A intrinsic breast cancer subtypes (Table 1) and increased risk of ER-negative tumors (specifically, ER-negative basal-like or triple-negative intrinsic breast cancer subtypes) (87–91). The combined effect of early age at first birth and lack of breast feeding appears to impart an especially high risk for basal-like (triple-negative and BRCA1) cancers (88,89,92–98), particularly among some specific ethnic groups.

Obesity is another risk factor with dual effects by age at diagnosis and ER status (90,99,100), which are possibly mediated through the cholesterol metabolite 27-hydroxycholesterol (101–103). Body weight has a direct association with postmenopausal breast cancer and an inverse relationship with premenopausal cancer (104–109). Obesity also has a stronger positive association with hormone receptor–positive than hormone receptor–negative cancers (101,102,110), although obesity may also increase the risk of basal-like, triple-negative, and inflammatory breast cancers (88,90,111,112). Rising ER–positive cancers among older women and falling ER–negative tumors among younger women (Figure 1B) are consistent with rising obesity and declining parity in the United States (113) and Denmark (114), as well as many other parts of the world (115). Other studies have also found dual risk factor associations by ER status [reviewed in (54)].

**Genetic Susceptibility**

Many genetic loci are known to contribute to the risk of familial breast cancer, including highly penetrant deleterious germline mutations in the BRCA1 and BRCA2 tumor suppressor genes. Of note, BRCA1 and BRCA2 breast cancers are distinct in their expression of hormone receptors (116–118). Roughly 75% of BRCA1 breast cancers are ER negative and 25% are ER positive. On the other hand, 75% of BRCA2 breast cancers are ER positive and 25% are ER negative, similar to breast cancer in the general population. Recent genome-wide association studies also have identified more than 75 low-penetrant susceptibility loci with evidence for specificity by tumor subtype (119–129). For 13 loci (119), there is greater relative risk for ER-positive than ER-negative cancers. For two notable exceptions (single nucleotide polymorphisms rs6828523 and rs7072776), the relative risk is opposite for ER-positive and ER-negative tumors (119), consistent with risk factor differences by ER expression. Lastly, there are a number of susceptibility loci only for ER-negative but not ER-positive tumors (eg, the hTERT loci) (130,131).

**Molecular Class Discovery And Prognosis**

Analyses of gene-expression profiling data reveal that ER-positive and ER-negative tumors are fundamentally distinct diseases in molecular terms (132). There are two predominantly ER-positive intrinsic molecular subtypes (ie, luminal A and luminal B) and two predominantly ER-negative intrinsic subtypes (ie, HER2-enriched and basal-like) (Table 1). The intrinsic molecular subtypes are largely distinguished by the expression of genes involved in luminal epithelial differentiation (eg, ER and PR genes), proliferation (eg, Ki67 gene), human epidermal growth factor receptor 2 pathway (eg, HER2 gene), and basal differentiation (133–136).

The intrinsic molecular signatures are robust across multiple genomic platforms (137,138), apply to both carcinoma in situ and...
Invasive breast cancers (139–141), and are identifiable within different racial groups (142–144). Although the intrinsic gene set was originally developed through agnostic profiling, the power of this approach reflects its ability to define molecular subtypes that vary with respect to prognosis and treatment (134,136,137,145–156). The intrinsic subtypes have been variously approximated with immunohistochemical staining algorithms (Table 1); however, there can be considerable discordance between the gene-based and immunohistochemical-based expression profiles for the intrinsic subtypes (157–159).

Luminal breast cancers are the most heterogeneous intrinsic subtypes, with the luminal A tumors distinguished by the high expression of luminal epithelial genes, low expression of the Ki67, and the best prognosis (160,161). The difference between the luminal A and luminal B gene patterns is less distinct than the difference between the luminal A and basal-like subtypes, which appear to be anticorrelated (88,146,162–165). Among the intrinsic subtypes, the basal-like tumors have the most unique and distinctive genomic profile (132,166–169); they are in many ways more similar to squamous cell carcinomas of the lung and high-grade serous ovarian carcinomas than to all other subtypes of breast cancer (9,170). Basal-like tumors are enriched with BRCA1-mutated and triple-negative breast cancers but also include some special histopathological subtypes such as medullary and adenoid cystic tumors (138,157,158,171). Finally, the HER2-enriched (HER2E) subtype shows a global gene signature that lies more closely to the luminal than basal-like cancers (172), with HER2 cell surface expression possibly playing an important role in regulating the luminal cancer stem cell population (5–7).

If every molecular subtype was a unique biological entity, one might anticipate that each would demonstrate a distinct age-specific incidence rate curve and unimodal age frequency distribution at diagnosis (45,58). Unfortunately, incidence rate data are not readily available for the intrinsic molecular signatures because gene-expression analyses have mostly been limited to case series and/or small observational studies from convenience and/or hospital-based samples. Nonetheless, in an early study that used the National Cancer Institute’s Surveillance, Epidemiology, and End Results (SEER) Residual Tissue Repository (164,173), we estimated the age-specific incidence rates for the molecular subtypes with immunohistochemical staining of breast cancer tissue microarrays and imputed population data. Luminal A (defined as ER positive and HER2 negative) incidence rates rose continuously, although more slowly after age 50 years (similar to ER-positive cancers) (Figure 1A). Basal-like rates (defined as ER negative and HER2 negative) increased rapidly early in life then flattened or fell (similar to ER negative cancer) (Figure 1A). These patterns are consistent with subsequent studies from the California Cancer Registry (174–177) and emerging data from SEER’s large-scale population-based database (178).

Given the limited availability of population-based incidence rate data, we applied the intrinsic gene set classification algorithm (147) to approximately 2000 breast cancer cases reported by METABRIC Group (179,180). Although METABRIC described 10 distinct groups based primarily upon copy number data, the age distribution patterns by gene expression–defined intrinsic molecular subtypes are reminiscent of the bimodal patterns defined by ER protein expression (Figure 2B). Luminal A and luminal B cases had bimodal age distributions with predominant late modes near age 70 years and minor modes near age 50 years (Figure 3A), similar to ER-positive cancers in SEER (Figure 2B). Basal-like cancers had an early-onset mode around age 50 years (Figure 3, A and B), similar to ER-negative cancers in SEER (Figure 2B). The HER2-enriched age distribution lay midway between the luminal and basal-like cancers (Figure 3A). Combining the HER2-enriched, luminal A, and luminal B cases into a single non-basal-like group did not appreciably alter the bimodal shape of the molecular subtypes (Figure 3B).

In METABRIC, hazard rates for breast cancer–specific death for basal-like, HER2-enriched, and luminal B tumors peaked near 7.5% per year approximately 2 years after initial breast cancer diagnosis then declined (Figure 3C), similar to ER-negative cases in SEER (Figure 2C). Luminal A hazard rates lacked a sharp peak and were relatively constant at 2% to 2.5% per year (Figure 3C), similar to ER-positive cases in SEER (Figure 2C). Overlaying the hazard rate curves for non-basal-like and basal-like cancers demonstrated crossing hazard rates approximately 8 years after diagnosis (Figure 3D), similar to ER-positive and ER-negative cancers in SEER (Figure 2C).

**Directions for Future Research**

We present the hypothesis that breast cancer comprises two fundamental etiological components, classes or subtypes; which, as of yet, are not specifically defined but which induce bimodal age distributions at diagnosis irrespective of the classification applied. The two putative main etiological subtypes are characterized by sharply contrasting tendencies related to age-specific incidence

---

**Table 1. Immunohistochemical staining for the intrinsic breast cancer molecular subtypes**

<table>
<thead>
<tr>
<th>IHC subtype (%)†</th>
<th>ER and/or PR</th>
<th>HER2</th>
<th>Ki67</th>
<th>Intrinsic subtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal A (73%)</td>
<td>Positive</td>
<td>Negative</td>
<td>Low</td>
<td>Luminal A</td>
</tr>
<tr>
<td>Luminal B (10%)</td>
<td>Positive</td>
<td>Negative or positive</td>
<td>Low or high</td>
<td>Luminal B</td>
</tr>
<tr>
<td>HER2 positive, nonluminal (5%)</td>
<td>Negative</td>
<td>Positive</td>
<td>Not needed</td>
<td>HER2-enriched</td>
</tr>
<tr>
<td>Triple-negative (12%)</td>
<td>Negative</td>
<td>Negative</td>
<td>Not needed</td>
<td>Basal-like</td>
</tr>
</tbody>
</table>

* There are two predominantly hormone-positive (estrogen receptor [ER] and/or progesterone receptor [PR]) intrinsic molecular subtypes (luminal A and luminal B) and two predominantly hormone-negative intrinsic subtypes (human epidermal growth factor receptor 2 [HER2] enriched and basal-like) (133–138). Additionally, there are two predominantly HER2-positive intrinsic molecular subtypes (luminal B and HER2 enriched) and two predominantly HER2-negative intrinsic subtypes (luminal A and Basal-like). Adapted from Goldhirsh et al. (187). IHC = immunohistochemical.

† Estimated percentage distribution for IHC-derived subtypes among women with breast cancer and known ER, PR, and HER2 expression in the general population of the United States in 2010, provided by the National Cancer Institute’s SEER database (178).
rates, clinical prognosis, risk factor profiles, and somatic gene and protein expression. ER expression by age at diagnosis provides an epidemiologically useful correlate for our two-component breast cancer mixture model, but it is not a perfect surrogate.

From the molecular perspective, multianalyte genomic profiling has revealed many novel breast cancer subtypes with distinguishable, if not distinctive, clinical tendencies. We speculate that these categorizations also may reflect varying mixtures of two main etiological components, as demonstrated by a ubiquitous bimodal age distribution at diagnosis lies in between the density plots for the luminal and basal-like cancers. Combining HER2 and luminal cases into a single non-basal-like group did not appreciably alter the shape of the bimodal age distribution plots shown for molecular subtypes. Hazard rates of breast cancer–specific death for basal-like, HER2E, and luminal B cancers peaked near 75% per year approximately 2 years after initial breast cancer diagnosis then declined, similar to estrogen receptor (ER)–negative cases in the National Cancer Institute’s Surveillance, Epidemiology, and End Results (SEER) database (Figure 2C). Luminal A hazard rates lacked a sharp peak and were relatively constant at 2% to 2.5% per year, similar to ER-positive cases in SEER (Figure 2C). Combining the HER2E and luminal cases into a single non-basal-like group resulted in a hazard plot that was intermediate between the basal-like and luminal A hazard rates. Hazard rates crossed over approximately 8 years after breast cancer diagnosis, similar to ER-positive and ER-negative cancers in SEER (Figure 2C).

Furthermore, emerging molecular evidence shows that across different types of cancer, breast cancer is one of the few cancers with two major divisions. One division consists of basal-like tumors and the other is composed of luminal A, luminal B, and HER2-enriched cancers (9,170), which we have conceptualized with a cartoon in Figure 4. These data, together with the epidemiological findings reported herein, suggest that breast cancer overall may be viewed as a hierarchal disease derived from two main cell types of origin (ie, basal/myoepithelial vs luminal cellular compartment) (2–9). Basal-like breast cancers arise from the basal/myoepithelial cell compartment, whereas non-basal-like cancers (ie, HER2-enriched, luminal B, and luminal A) emerge from a more luminal-like cell compartment.

Data in this commentary also suggest themes for future etiological and clinical research. Large-scale and population-based epidemiological studies could stratify etiological analyses by molecular subtypes such as basal-like vs non-basal-like or luminal. Our current thinking is that there are two main etiological subtypes for sporadic breast cancers, but there could be a few more, potentially including rare subtypes such as inflammatory breast cancer (112). Given the challenges in molecular subtyping in large-scale population-based studies, it would be helpful to further develop and validate parsimonious robust marker panels for studies that merge individual-level molecular and clinical data in populations within well-defined catchment areas (173,182,183).
Understanding the etiological heterogeneity of breast cancer, whether it is fundamentally bimodal as proposed herein or more complex, would have critical implications for breast cancer prevention. For example, selective ER modulators are effective in reducing breast cancer incidence among women at higher risk overall (184), but the means for identifying women specifically at elevated risk for hormonally driven cancers are lacking. Furthermore, ER-positive status is only a surrogate of hormone dependence; selective ER modulators do not prevent all ER-positive cancers nor do all ER-positive tumors respond to adjuvant endocrine therapy (4,185). If the hormone-responsive phenotype is linkable to a risk factor profile (eg, the etiological class of breast cancers enriched, but not exclusively within the late onset peak), then identifying this subtype more specifically would represent an advance with translational potential for both prevention and treatment. Similarly, understanding the basic biology of tumors enriched within the early-onset age distribution at diagnosis might offer analogous opportunities for BRCA1-related, basal-like, and/or triple-negative breast cancers.

As a final thought, although a two-component mixture model may seem too simplistic for breast cancer clinical heterogeneity, it is not too simple for etiology. Much of the clinical heterogeneity for breast cancer may result from tumor promotion and progression, which is very likely far downstream of tumor initiation (31). Therefore, breast cancer etiology (tumor initiation) may be less complex than subsequent tumor promotion and progression. Additionally, although parsimonious, the complexity of mixture models should not be underappreciated. The potential for an infinite number of mixtures of just two main components is complicated enough to account for much of the observed breast cancer heterogeneity.

**Figure 4.** Results and data herein suggest that breast cancer overall may be viewed as a hierarchical disease, consisting of a two-component mixture of two main cell types of origin (ie, luminal vs basal/myoepithelial). As hypothesized in this cartoon, human epidermal growth factor receptor 2–enriched (HER2) and luminal intrinsic molecular subtypes are initiated (lightning bolts) from the luminal cell compartment. On the other hand, in this model, basal-like breast cancers (HER2+/HR−) are initiated within the basal/myoepithelial cell compartment.

**References**

miological studies in 30 countries, including 50,302 women with breast cancer. Do they differ according to age at diagnosis?


types of ductal carcinoma in situ and invasive breast cancer. Like subtype of invasive breast carcinoma.


Funding

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute (NCI), and by extramural funds to CMP (NCI Breast SPORE program; P50-CA58223-09A1).

Notes

The authors had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of reported findings. CMP is an equity stock holder and board of director member of BioClassifier LLC and University Genomics. CMP is also listed an inventor on a patent application on the PAM50 molecular assay.

We wish to thank Dr Hyuna Sung and Mr David Check, Department of Health and Human Services/NIH/NCI/Division of Cancer Epidemiology and Genetics for advice and enhancements for Figure 4.

Affiliations of authors: Division of Cancer Epidemiology and Genetics Biostatistics Branch (WFA, PSR), and Division of Cancer Prevention (MES), National Cancer Institute, National Institutes of Health, Bethesda, MD; Translational Genomics Group, Vall d’Hebron Institute of Oncology, Barcelona, Spain (AP); Department of Genetics and Pathology & Laboratory Medicine, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC (CMP).