SMOKING AND VARIATION IN BREAST TUMOR BIOMARKER EXPRESSION

Eboneé Nicole Butler

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology in the Gillings School of Global Public Health.

Chapel Hill
2017

Approved by:

Melissa A. Troester

Jeannette T. Bensen

Mengjie Chen

Kathleen Conway

Andrew F. Olshan

ABSTRACT

Eboneé Nicole Butler: Smoking and Variation in Breast Tumor Biomarker Expression
(Under the direction of Melissa A. Troester)

Purpose: Smoking is a suspected risk factor for breast cancer, with hypothesized links to estrogen-mediated, genotoxic, and growth-factor dependent mechanisms. Each mechanism can be modeled by overexpression of ER, p53, and EGFR, respectively. This dissertation examines associations between smoking and biomarkers for each mechanistic pathway. **Methods:** Our population-based study included 1,970 women diagnosed with invasive breast cancer in central and eastern North Carolina. Single and multigene biomarker outcomes were characterized as binary (+/-) or continuous measures for protein or mRNA. Single gene measures included ER/ESR1, p53/MDM2, and EGFR. Multigene mRNA signatures included a luminal score (LS); a p53 signature used to describe wild-type (Wt) or mutant (Mut) activity; and an algorithm-based proliferation score (PS). We used logistic and linear regression models to estimate associations between smoking and biomarker outcomes. **Results:** (Aim 1) When compared with never smokers, the odds of ER+, ESR1+, and LS+ tumors were nearly doubled among current smokers, those who smoked 20 or more years, and those who smoked within 5 years of diagnosis. Quantitative levels of ESR1 mRNA were highest among current smokers compared to never smokers overall and among women with ER+ breast cancer; however, we did not observe associations between smoking and continuous ER protein expression. (Aim 2) ER- cases with a history of ever smoking were at increased odds of having breast tumors with the p53 IHC+ molecular

phenotype. In addition, long smoking duration was also associated with higher quantitative levels of p53 protein among ER- breast tumors but not ER+ breast tumors. The EGFR IHC+ phenotype was inconsistently linked to smoking for both ER+ and ER- tumors. With respect to our multigene mRNA signatures, smoking was not linked to either the p53 Wt or p53 Mut subtype; however, with respect to the proliferation score, smoking metrics were consistently linked to lower odds for the PS+ subtype. **Conclusions:** Both single and multigene measures for each mechanistic pathway captured tumor changes associated with smoking. Findings from our study have implications for understanding potential mechanisms underlying smoking and breast cancer risk.

I dedicate this dissertation to my Granny, Ms. Eddie Mae Johnson.

I love you.

"You are the light of the world—like a city on a hilltop that cannot be hidden. No one lights a lamp and then puts it under a basket. Instead, a lamp is placed on a stand, where it gives light to everyone in the house. In the same way, let your good deeds shine out for all to see, so that everyone will praise your heavenly Father.

Matthew 5:14-16 (NLT)

ACKNOWLEDGEMENTS

The completion of this dissertation would not be possible without the support of many people. I would first like to thank my advisor, Dr. Melissa Troester, for her teaching, support, and wisdom. I am grateful for her guidance as I've grown in scholarship. I would also like to thank the members of my dissertation committee for their invaluable contributions: Dr. Jeannette Bensen has encouraged me to think about the myriad of biological processes that may influence observed associations between risk factors and tumor biomarker expression; Dr. Mengjie Chen has exposed me to new methods in bioinformatics that have been a great complement to my studies in epidemiology; I am thankful to Dr. Kathleen Conway for the many hours she spent discussing my research interests and objectives, which helped me to better articulate my research aims; and I am grateful to Dr. Andrew Olshan who has been a great academic advisor and department chair during my time at UNC.

Navigating the PhD process is tough and I encountered many unfamiliar situations that I could not have processed on my own. Thank you to Dr. Sherry Eaton for her guidance and for helping me to grow in personal and professional confidence. I am encouraged by her and, because of her, I will always remember to let my light shine. Thank you to Ms. Kathy Wood for helping me navigate the academic environment. In the pursuit of higher education, there are many challenges that are unique to black students and other underrepresented minorities. Kathy's work and mission make it easier for student's like me to find a soft place

to land. I am also grateful to Ms. Chandra Caldwell. Thank you for your pep talks and for encouraging me to stay the course.

Last, but not least, thank you to Ms. Mary Elizabeth Bell and the Carolina Breast Cancer Study. It has been a great pleasure to work with you. Thank you for allowing me to be a part of your team. I am forever grateful to the 7,333 women who serve as participants in the study and all staff – past and present – who make CBCS a reality. Thank you especially to Mr. Adam Gardner, Ms. Erin Lutz, Ms. Linda Shaw, Ms. Pamela Mack, Mr. Scott Gee, Ms. Michele Smith, and Ms. Susan Campbell. You all make a work group feel more like family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CBCS	Carolina Breast Cancer Study
CI	Confidence interval
CK 5/6	Cytokeratin 5/6
CPD	Cigarette packs per day
DNA	Deoxyribonucleic acid
DF	Degrees of freedom
EGFR	Epidermal Growth Factor Receptor
ER	Estrogen-receptor (protein)
ESR1	Estrogen-receptor (DNA/RNA)
HER2	Human Epidermal Growth Factor Receptor-2
HR	Hazard ratio
IHC	Immunohistochemistry
IRR	Incidence rate ratio
LS	Luminal score
LRT	Likelihood ratio test
MUT	Mutant
OR	Odds ratio
PAM50	Prediction Analysis of Microarray 50
PR	Progesterone-receptor
PS	Proliferation score
mRNA	Messenger ribonucleic acid
TMA	Tissue microarray
TNBC	Triple-negative breast cancer
TP53	Tumor protein 53 (p53)
WT	Wildtype

CHAPTER 1: SPECIFIC AIMS

Smoking is a suspected risk factor for breast cancer, based on weak-to-moderate measures of association, and the detection of tobacco smoke particulates in breast tissues of smokers. However, the US Surgeon General concludes that there is insufficient evidence to suggest a causal relationship between active smoking and breast cancer risk. Indeed, epidemiologic studies have yielded mixed results. Because many breast cancers are estrogen dependent, studies that report earlier menopause and lower levels of circulating estrogens among smokers support an "anti-estrogenic" effect, and would suggest inverse risk. However, few epidemiologic studies have supported an "anti-estrogenic" hypothesis; studies have more commonly suggested a positive association between smoking and breast cancer risk, consistent with tissue culture and animal experiments, showing that cigarette smoke causes DNA-damage, disrupts cell-cycle regulation, and is linked to malignant transformation.

Adding complexity to studies of smoking and breast cancer risk is the observation that breast cancer is a heterogeneous disease defined by distinct and reproducible gene expression profiles. These gene expression profiles reflect breast cancer "intrinsic subtypes" that are prognostic and predictive of response to treatment, and also appear to have distinct etiologic profiles. However, the relationship between smoking and breast cancer intrinsic subtype has not been completely evaluated. Furthermore, proposed mechanisms of the smoking-breast cancer relationship have implicated at least three mechanistic pathways:

estrogen-mediated, genotoxic, and growth factor-mediated. Thus, a critical examination of smoking and breast cancer risk would benefit by examining intrinsic subtype and tumor biomarkers linked to pathogenesis. Moreover, such studies should carefully incorporate information on dose and timing of exposure. The examination of temporal and dose-dependent patterns of smoking in relation to biomarker-defined breast cancer subtypes may identify time-windows of susceptibility that are associated with specific breast cancer subtypes.

The current proposal uses data from the population-based Carolina Breast Cancer Study (CBCS), which combines molecular biology and epidemiology to examine genetic and environmental risk factors for breast cancer. The CBCS has collected protein and RNA expression data on genes involved in breast tumor biology, specifically: estrogen receptor (ER) protein and RNA expression; p53 protein and p53-dependent RNA expression; and epidermal growth factor receptor (EGFR) protein and RNA expression. These hypotheses will be evaluated while simultaneously considering breast cancer intrinsic subtype.

Aim 1: To examine the relationships between temporal and dose-dependent patterns of smoking, breast cancer intrinsic subtype, and ER expression. Rationale: Growing evidence suggests a possible association between smoking and the luminal, estrogen receptor (ER)-positive breast cancer subtypes, ranging from a modest 5% increased risk to more than doubled risk when comparing ever smokers with never-smokers. Hypothesis: Smoking will be quantitatively and qualitatively associated with increased risk of luminal/ER+ breast cancer and inverse risk of basal-like/ER- breast cancer. Approach: To examine the temporal and dose-dependent relationship between smoking and continuous and categorical measures

of i) luminal and basal-like intrinsic subtypes and; ii) ER protein expression and ESR1 RNA levels.

Aim 2a: To examine the relationships between temporal and dose-dependent patterns smoking, breast cancer intrinsic subtype, and p53 expression. Rationale: Studies have reported that breast cancer patients who were self-reported smokers at time of diagnosis had higher prevalence of specific TP53 mutations compared to their non-smoking counterparts. P53 expression regulates genomic stability and the DNA-damage response and may be an important mechanism underlying the relationship between smoking and breast cancer risk. Hypothesis: Temporal and dose-dependent variation in smoking is associated with variations in risk of p53 mutant cancers. Approach: To evaluate the temporal and dose-dependent association between smoking and p53+ breast cancer risk as measured by IHC and p53-dependent RNA signature.

Aim 2b: To examine the relationships between temporal and dose-dependent patterns of smoking, breast cancer intrinsic subtype, and EGFR expression: Rationale: *In vitro* studies of breast epithelial cells treated with nicotine have demonstrated higher expression of EGFR compared to untreated cells, providing evidence for nicotine as a possible environmental agent linked to EGFR+ breast cancer. Hypothesis: Temporal and dose-dependent variation in smoking exposure is associated with variations in EGFR+ breast cancer risk. Approach: To evaluate the temporal and dose-dependent association between smoking and EGFR+ breast cancer risk as measured by IHC and RNA expression.

CHAPTER 2: BACKGROUND

This year nearly 250,000 US women will be diagnosed with breast cancer[1]. Of these cases, only 1 in 4 will be attributed to high-penetrance germline mutations or familial clusters of the disease. The remaining three-fourths will have no known markers of heritable susceptibility, leaving patients, communities, and researchers to grapple with understanding how environments and individual behaviors influence breast cancer risk. The identification of breast cancer intrinsic subtypes adds to the difficult task of understanding breast cancer etiology, as each subtype is hypothesized to have a distinct risk factor profile. It follows that a critical evaluation of any risk factor must consider the inherent heterogeneity across breast tumors, including differential expression of biomarkers linked to pathogenesis. Smoking is a suspected cause of breast cancer and has been linked to breast tumors that arise via estrogen-mediated, genotoxic, and growth-factor dependent mechanisms. In this proposal, we will examine the association between smoking exposure and breast tumor expression of biomarkers that reflect biologic activity of each mechanistic pathway. We will also examine temporal and dose-dependent patterns of smoking to identify etiologically-relevant time windows that may be associated with early or late carcinogenic events in the development of biomarker-defined breast tumors. By examining smoking exposure in relation to biomarkers linked to pathogenesis, we may identify etiologically-relevant subtypes that have been masked in prior studies of smoking and breast cancer risk.

2.1 Two Etiologic Types of Breast Cancer

The heterogeneous nature of breast cancer has been well-established with the identification of distinct and reproducible intrinsic subtypes[2]. Gene expression studies have identified at least four subtypes that occur with predictable frequencies in representative populations of US women[3-5], namely: Luminal A, Luminal B, HER2-enriched (HER2E), and basal-like breast tumors. Luminal A tumors are most common, occurring in approximately seventy percent of cases. In comparison to other subtypes, Luminal A tumors are the most genetically diverse and are characterized by high estrogen signaling, respond to hormone therapy, and may carry the most favorable prognoses[6]. Luminal B and HER2E tumors account for approximately ten and five percent of breast cancer cases, respectively. These two subtypes are characterized by overexpression or amplification of HER2 and may respond to hormone therapy and the monoclonal antibody, trastuzumab. However, Luminal B tumors have higher levels of estrogen signaling when compared with HER2E tumors. Estimates for the prevalence of basal-like tumors range between 10 and 20 percent of breast cancer cases. Basal-like tumors are characterized by low estrogen-signaling, lack targeted therapies, and are associated with clinical markers for aggressive disease. Though Luminal A, Luminal B, and HER2E tumors are defined by distinct gene expression profiles, each expresses proteins that are found predominantly in the luminal epithelial layer of the mammary gland and are thought to share luminal epithelial origins. By contrast, tumors that are classified as basal-like express proteins that are most abundant in the basal/myoepithelial layer of the mammary gland.

Though intrinsic subtypes have both predictive and prognostic value, less is known concerning their utility in studies of etiology. Breast cancer incidence trends have been used

to suggest the existence of two main etiologic types – luminal and basal-like – based on estrogen-receptor expression and average age at onset[7,8]. Basal-like breast cancers are estrogen-receptor negative (ER-) and have a younger average age at onset relative to luminal breast cancers, which are estrogen-receptor positive. This observation supports arguments that breast cancers of luminal epithelial and basal/myoepithelial origins represent two distinct etiologic classes of disease. In addition, a growing number of genomic platforms have identified two distinct breast cancer clusters when representative samples of breast tumors are compared with tissues from other cancer types[9]. These clusters are characterized by tumors that are ER+ and ER-, which map to the luminal and basal-like phenotypes, respectively.

Several epidemiologic studies have examined associations between traditional breast cancer risk factors and breast tumors stratified by ER status (i.e., ER+ vs. ER-) or luminal and basal-like designations. Early age at menarche, African American race, and young age are associated with increased risks of Basal-like or triple-negative breast cancers[10-13]; these non-modifiable risk factors may reflect unmeasured risk factor profiles that increase susceptibility to the basal-like phenotype. Lower body mass index (BMI) and breastfeeding among women with high parity are two modifiable risk factors associated with inverse risk of the basal-like breast cancer type[10,12-16]. Alcohol intake has been consistently linked to increased risk of breast cancer, with many studies showing evidence of a strong association with luminal or estrogen-receptor positive disease[15,17]. Further, risks for luminal and basal-like breast cancers may differ by physical activity engagement, first degree family history of breast cancer[16], menopausal status[18], income, exposure to exogenous hormones[10], and age at first pregnancy[19,20].

2.2 Smoking and Breast Cancer Risk

Although the prevalence of cigarette smoking has steadily decreased since the 1950s, approximately 50% of women in the United States report a history of ever smoking and 14% are self-reported current smokers[21]. The Surgeon General's 2014 report on the "health consequences of smoking" suggests there is sufficient evidence to identify mechanisms by which cigarette smoke could cause breast cancer, based on data from animal studies; the report concludes, however, that current population-based evidence is insufficient to infer causation. This conclusion has been attributed to inconsistent results from epidemiologic investigations, including the lack of an observed dose-response relationship. Indeed, epidemiologic studies of smoking and breast cancer risk have yielded a mix of positive and null findings, suggesting little or no increased risk of disease. And as in most studies of smoking and cancer incidence, duration – but not dose – has been more consistently associated with risk.

Several contemporary studies of smoking and breast cancer risk have reported positive associations in both age- and ethnically-diverse populations of women (Table 2.1). Investigators from the Multiethnic Cohort Study, Cancer Prevention Study II, Black Women's Health Study, and Nurse's Health Study have demonstrated positive, though modest, associations for smoking status (i.e., current, former, never), dose, and duration and breast cancer risk. In a meta-analysis of 15 cohort studies, Gaudet et al. estimated a 10 percent increased risk of breast cancer among women who were current smokers at time of study enrollment (HR 1.1, 95% CI = 1.1 to 1.2) and a 10 percent increased risk among women who were classified as former smokers (HR 1.1, 95% CI = 1.0 to 1.2), when compared with never smokers. These large-scale cohort studies improve upon prior

investigations by addressing biases inherent to observational study designs, including minimizing the potential for recall bias and control for confounding or effect measure modification due to established breast cancer risk factors. However, few studies have considered the heterogeneous nature of breast cancer as a source of bias in etiologic investigations and its potential to mask associations between smoking and distinct molecular subtypes of breast cancer.

2.3 Smoking and Breast Tumor Biomarkers Linked to Pathogenesis

Tobacco smoke includes more than 70 carcinogens that have been evaluated by the International Agency for Research on Cancer (IARC), and which comprise eight chemical classes [20]. Two of the largest classes – the polycyclic aromatic hydrocarbons (PAHs) and the *N*-nitrosamines – are thought to be responsible for cancer initiation in lung tumors. Moreover, PAHs, *N*-nitrosamines – and their predecessor, nicotine – have been examined in human breast tissue and tissue culture for their ability to transform normal breast epithelium to cancer[22-24]. Metabolized forms of PAHs and the tobacco-specific *N*-nitrosamines can form covalent bonds at susceptible nucleotide binding sites to form DNA-adducts[21]. If the affected cell evades an arsenal of DNA repair mechanisms (e.g., nucleotide excision repair), the resultant adduct can yield single-base point mutations that may render a gene's protein product non-functional. Further, smoking exposure may also result in chromosomal breaks and loss of heterozygosity, leading to DNA copy number aberration. Thus, protein and RNA expression levels of key genes may provide clues to understanding the etiology of smoking and breast cancer risk.

Smoking has been linked to breast tumors that arise via estrogen-mediated, genotoxic, and growth-factor dependent mechanisms; these mechanisms can be modeled by

overexpression of the estrogen-receptor, TP53, and epidermal growth factor receptor (ER, p53, and EGFR), respectively. The overexpression of each marker reflects aberrant changes in cell-cycle regulation and homeostatic disruption of the tumor microenvironment, which allow a single cancer cell to gain selective advance and multiply through clonal expansion. By examining smoking exposure in relation to biomarkers linked to pathogenesis, we may identify etiologically-relevant subtypes that have been masked in prior epidemiologic studies of smoking and breast cancer risk.

2.3.1 Intrinsic Subtype

Triple subtypes – defined as the joint expression of the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) – have been used as surrogates for intrinsic subtype in studies of breast cancer etiology. Each marker is measured by IHC or mRNA assay and is designated as positive (+) or negative (-), based on clinicopathologic cut points for overexpression. Luminal types are typically defined as tumors that are ER+ and or PR+, irrespective of HER2 expression. And triple negative tumors, surrogates for Basal-like tumors, are negative for all three markers. Although triple subtypes strongly correlate with gene expression profiles for intrinsic subtype, varying levels of discordance exist[25]. For example, approximately 75% of triple negative tumors are confirmed as Basal-like by gene expression assay, while the remaining 25% can be genetically similar to Luminal breast cancers. Thus, triple subtype designations are a potential source of outcome misclassification and may yield inconsistent results across studies of smoking and subtype-specific breast cancer risk. Smoking exposure has been variably linked to increased risk of the Luminal subtype and has more consistently showed a null association with the Basal-like subtype (Table 2.2). As shown in the Carolina Breast

Cancer Study, triple subtypes can be further refined by IHC assessment of cytokeratin 5/6 (CK 5/6) and epidermal growth factor receptor (EGFR), which are characteristic of the Basal-like phenotype. Future studies of smoking and breast cancer risk can also benefit from gene expression profiling of tumors, as described by the PAM50 gene signature.

2.3.2 Estrogen Receptor

With a large majority of breast cancers driven by exposure to estrogen, it has been suggested that the anti-estrogenic properties of cigarette smoke counteract its carcinogenic effects, leading to null or inconsistent associations of smoking and breast cancer risk among smokers[26]. This hypothesis is supported by the observation that smokers report earlier age at menopause and have lower levels of circulating estrogens. Nevertheless, epidemiologic investigations have commonly demonstrated a positive association between smoking and breast cancer risk, particularly for ER+ tumors (Table 2.3). In a recent population-based case-control study of women in the Seattle Puget-Sound metropolitan area, Kawai et al. reported a 40% increased risk of ER+ breast cancer among ever smokers (OR 1.4, 95% CI = 1.0 to 1.9), but found no association between smoking and ER- tumors. Similarly, in the Cancer Prevention Study II, Gaudet et al. reported a 20% increased risk of ER+ breast cancer among current smokers (OR 1.2, 95% CI = 1.0 to 1.5) and a 10% increased risk of ER+ breast cancer among former smokers (OR 1.1, 95% CI = 1.1 to 1.3); results among women in the AARP cohort show a 40% increased risk of ER+ breast cancer among current smokers (HR 1.4, 95% CI = 1.0 to 1.8). Neither of these studies reported associations between smoking and estrogen-receptor negative (ER-) breast cancer, suggesting that smoking exposure may be related distinct pathophysiologic pathways leading to overexpression of the estrogen receptor. However, in stark contrast to contemporary studies in North American populations,

older studies of smoking and breast cancer risk in Swedish and Swiss populations have demonstrated positive associations between smoking and ER- breast cancer. Thus, a careful investigation should consider era and characteristics of population of interest.

2.3.3 Tumor protein p53

TP53 (or p53) is the most frequently mutated gene in breast cancer with estimates ranging between 20 to 40 percent of all cases. Investigators have demonstrated that breast cancer patients who are self-reported smokers at time of diagnosis have a higher prevalence of specific TP53 mutations when compared with their non-smoking counterparts. In the absence of gene sequence technologies, epidemiologists have employed IHC staining as one high throughput method to detect p53 protein expression in breast tumors. Nuclear staining of p53 positively correlates with TP53 missense mutations, which render non-functional forms of the protein. In a case-case analysis, Gammon et al. reported that smokers were twice as likely to be diagnosed with p53+ breast cancer when compared to never smokers, where p53-positivity was defined as moderate to strong staining in 10% or greater of tumor cells (Table 2.4). Furberg et al. suggest that the relationship between smoking and p53+ breast cancer risk may be null; here, the authors defined p53-positivity as dark nuclear staining in 10% or greater of examined tumor cells. The discordant results between the two studies may be explained by differences in protocol and the selection of the p53 IHC cut point. Given the evidence that breast cancer patients who are smokers harbor a greater proportion of TP53 mutations, and that p53 protein IHC expression correlates with TP53 missense mutation, it is plausible that the threshold selected by Furberg et al. did not accurately categorize tumors with and without TP53 missense mutations, thereby yielding a spurious null association between smoking and p53 expression in breast tumors.

2.3.4 Epidermal Growth Factor Receptor

Until recently, nicotine - the most abundant and pharmacologically active component of cigarette smoke – had not been implicated in the development of breast cancer. Mouse xenograft models have shown that normal mammary cells transform to neoplastic cells upon exposure to nicotine and its derivatives[27]. And similar to findings from lung cancer studies[28], breast epithelial cells treated with nicotine show increased rates of cell proliferation through activation of EGFR[29], providing a measurable biological endpoint for a hypothesized association between nicotine and EGFR+ breast cancer risk. EGFR overexpression is present in nearly three-fourths of basal like breast tumors, predicting increased risk of recurrence and poorer overall survival. Although less common, EGFR-positivity is also observed among Luminal A, Luminal B, and human epidermal growth factor receptor-2 (HER2) enriched intrinsic subtypes to varying degrees; however, its prognostic value for these breast tumor subtypes has not been defined. A recent study of basal-like marker expression (i.e., EGFR and/or CK 5/6) in luminal tumors showed that the distribution of traditional breast cancer risk factors (i.e., family history of breast cancer, age at menarche, parity, age at first full-term pregnancy, number of live births, breast feeding, and BMI) did not differ according to EGFR expression[30]. Although there were no apparent etiologic differences, Luminal tumors that expressed EGFR had more favorable clinical features when compared with tumors that did not express EGFR.

At present, no studies have examined the epidemiology of EGFR+ breast cancer. Several studies, however, have examined tumor and clinical characteristics with respect to EGFR expression using IHC cut points of 1% or 10% of stained tumor cells to characterize tumors as positive or negative. Tumors defined as EGFR+ were less likely to express ER or

PR and were positively correlated with chromosome 7 polysomy, HER2 gene amplification, and EGFR protein expression[31-35]. EGFR+ tumors have also been shown to have inverse associations with disease-free survival, are more likely to have positive associations with lymph node status, and may be associated with specialized histological breast cancer types. In addition, EGFR+ tumors may be associated with higher proliferative fractions, increased aneuploidy, and increased tumor size. Although there is little data regarding the epidemiology of EGFR+ disease, investigators for study of early stage breast cancer patients reported positive associations between high EGFR tumor expression, African American race, and young age.

The proposed study will be the first to investigate the association between smoking, concomitant nicotine exposure, and risk of breast cancer characterized by overexpression of EGFR. It is important to emphasize that prior investigations of smoking and breast cancer intrinsic subtype found no link between smoking and basal-like breast cancer, of which nearly three-fourths overexpress EGFR[36]. Recall, however, that EGFR may be overexpressed for all intrinsic subtypes to varying degrees[25,36,37]. Thus, dichotomizing breast cancers by EGFR expression (i.e., EGFR+ and EGFR-) will allow us to specifically assess whether smoking is associated with EGFR+ breast cancer across subtypes.

2.4 Exposure-Time-Windows and Breast Cancer Risk

Prior investigations of smoking and breast cancer risk have typically used crude definitions of induction and latency periods for disease. Recall that induction and latency periods are defined as the time between exposure and disease initiation – and the time between disease initiation and disease manifestation, respectively. Cohort studies have most commonly used baseline measures of smoking, which may not reflect exposure levels at time

of breast cancer diagnosis (Table 2.1). Both cohort and case-control studies have incorporated measures of smoking initiation with respect to categories of age and smoking cessation relative to time-windows preceding disease diagnosis. These exposure-time-windows are defined by the investigator and observed associations may be sensitive to selected cut points. Improper specification of induction and latency periods leads to non-differential exposure misclassification and biases effect estimates toward the null[38]. Thus, methods that allow researchers to examine temporal patterns of smoking without selecting time periods *a priori* may be beneficial in studies of smoking and breast cancer risk.

The current proposal seeks to explore the temporal relationship between smoking and breast cancer risk. Logistic regression models that include parametric latency functions can be used to evaluate variation in disease risk by time since exposure. The inclusion of a parametric latency function in our logistic regression model will allow us to calculate time weighted exposure estimates using maximum likelihood estimation, where the highest weights are assigned during the period where smoking is associated with the greatest risk of EGFR positive breast cancer. In addition, non-parametric functions (e.g., B-splines) – which do not impose a specified probability distribution – can be used to visualize trends that describe differential breast cancer risk along the course of a woman's smoking history. Further, as we explore temporal patterns of smoking in relation to biomarker-defined breast cancer types, we may observe associations that allow us to hypothesize temporal associations for the activation of a given mechanistic pathway with respect to smoking exposure. Specifically, evidence of an association between smoking and breast cancer risk proximal to time of diagnosis infers a late-acting carcinogenic event. By contrast, if smoking exposure is associated with breast cancer risk at a distal point from date of diagnosis, we may infer that

smoking is associated with an early carcinogenic event. By using latency models for protracted exposures, we will identify exposure time frames where smoking is most etiologically relevant to biomarker-defined breast tumors.

2.5 Summary

Identifying the component causes of breast cancer remains one of the greatest public health challenges of the 21st century. Researchers have proposed smoking as a probable risk factor the disease; however, a definitive relationship between smoking and breast cancer risk has not been established. In this proposal, we will consider the heterogeneous nature of breast cancer and its potential to mask associations between smoking and distinct molecular subtypes of breast cancer. We will also examine temporal patterns of smoking exposure using data-driven approaches that incorporate maximum likelihood estimation to identify critical exposure-time-windows. Results from our study may help to elucidate associations masked in prior studies.

2.6 Literature Review Tables

See next page.

Table 2.1. Temporal and dose-dependent measures of smoking and breast cancer risk.

Reference	**Study Design and Overview**	**Temporal** relative to date of study enrollment		**Duration** (years)		**Dose** (cigarettes/day)	
Catsburg (2014)[39]	Case-cohort. The Canadian Study of Diet, Lifestyle and Health (CSDLH). There were 1,096 breast cancer cases and 3,314 women included in the sub-cohort (approximately 4% of those in sub-cohort became cases). Smoking exposure was determined by self-administered questionnaire. The authors observed no apparent association between smoking measures and breast cancer risk.	Current vs. Never Former vs. Never	HR (95% CI) 1.0 (0.8 to 1.4) HR (95% CI) 1.0 (0.9 to 1.2)	Overall <10 10 - < 20 20 - < 30 30 - < 40 40+ Never	HR (95% CI) 0.9 (0.8 to 1.2) 1.2 (0.9 to 1.5) 0.9 (0.7 to 1.2) 1.2 (0.9 to 1.5) 0.9 (0.6 to 1.3) (Ref)	Overall 20+ 15 - < 20 10 - < 15 5 - < 10 < 5 Never	HR (95% CI) 1.0 (0.7 to 1.3) 1.1 (0.8 to 1.4) 1.1 (0.9 to 1.5) 1.0 (0.8 to 1.3) 0.9 (0.7 to 1.1) (Ref)
Cui (2006)[40]	Cohort. Canadian National Breast Screening Study (NBSS). Approximately 90,000 women were recruited for a randomized controlled trial of mammographic screening for breast cancer; 4,445 became cases. Women were followed for an average of 16 years. Smoking exposure was assessed at baseline. Long smoking duration (>40 years), smoking dose (40+ cigarettes/day), and smoking initiation 40 years prior to study enrollment	Smoking Start Years Prior 1-9 10-19 20-29 30-39 40+ Never	 RR (95% CI) 1.0 (0.8 to 1.3) 1.1 (0.9 to 1.2) 1.0 (1.0 to 1.1) 1.1 (1.0 to 1.2) 1.3 (1.1 to 1.6) (Ref)	Overall 1-9 10-19 20-29 30-39 40+ Never	HR (95% CI) 1.0 (0.9 to 1.1) 1.0 (0.9 to 1.1) 1.1 (1.0 to 1.2) 1.1 (1.0 to 1.3) 1.5 (1.2 to 1.9) (Ref)	Overall 1-9 10-19 20-29 30-39 40+ Never	RR (95% CI) 1.0 (0.9 to 1.1) 1.1 (1.0 to 1.2) 1.1 (1.0 to 1.2) 1.1 (0.9 to 1.3) 1.2 (1.0 to 1.4) (Ref)

Reference	Study Design and Overview	Temporal relative to date of study enrollment		Duration (years)		Dose (cigarettes/day)	
(continued)	was associated with increased breast cancer risk.						
Dossus (2014)[41]	Cohort. European Prospective Investigation into Cancer. Of the 322,988 women enrolled in the study, 9,822 developed breast cancer over an average follow-up period of 11 years. Smoking was assessed by baseline questionnaire. The authors observed a slight association between current or former smoking and increased breast cancer risk. There was a trend between increasing smoking duration and increased risk among current smokers. This trend was not evident among former smokers although the highest categories of smoking duration were associated with increased risk for this group. The highest categories of smoking dose were also associated with increased breast cancer risk for current and former smokers.	Current vs. Never Former vs. Never	HR (95% CI) 1.1 (1.0 to 1.1) HR (95% CI) 1.1 (1.0 to 1.1)	Current Smokers 0-10 10-20 20-30 >30 Never Former Smokers 0-10 10-20 20-30 >30 Never	 HR (95% CI) 0.9 (0.7 to 1.3) 1.0 (0.8 to 1.1) 1.0 (1.0 to 1.1) 1.1 (1.0 to 1.2) (Ref) HR (95% CI) 1.1 (1.0 to 1.2) 1.0 (0.9 to 1.1) 1.1 (1.0 to 1.2) 1.0 (0.9 to 1.1) (Ref)	Current Smokers <6 6-10 10-15 ≥15 Never Former Smokers <6 6-10 10-15 ≥15 Never	 HR (95% CI) 1.0 (0.9 to 1.1) 1.0 (0.9 to 1.2) 1.1 (1.0 to 1.2) 1.1 (1.0 to 1.2) (Ref) HR (95% CI) 0.9 (0.9 to 1.0) 1.0 (0.9 to 1.1) 1.1 (1.0 to 1.2) 1.1 (1.0 to 1.3) (Ref)

Reference	Study Design and Overview	Temporal relative to date of study enrollment		Duration (years)		Dose (cigarettes/day)	
Gaudet (2013)[42]	Cohort. The Cancer Prevention Study II (CPS-II) Nutrition Cohort. 97,786 women were enrolled in 1992 to examine cancer incidence and mortality. Median follow-up was 14 years 3,721 invasive breast cancers occurred. Both current and former smoking were associated with slight increased risks of breast cancer. Among former smokers, long duration (31 to 70 years) was associated with breast cancer risk. However, duration was not associated with risk among current smokers. There was no apparent association between smoking dose and breast cancer risk.	Current vs. Never Former vs. Never	HR (95% CI) 1.3 (1.1 to 1.4) HR (95% CI) 1.1 (1.1 to 1.2)	Current Smokers 1-40 40-49 50-73 Never Former Smokers <1-10 11-20 21-30 31-70 Never	HR (95% CI) 1.2 (0.9 to 1.5) 1.0 (0.9 to 1.4) 1.0 (0.9 to 1.6) (Ref) HR (95% CI) 1.2 (1.1 to 1.3) 1.2 (1.1 to 1.3) 1.1 (1.1 to 1.3) 1.0 (1.2 to 1.4) (Ref)	Current Smokers 1-9 10-19 20-29 30-39 40-90 Never	HR (95% CI) 1.2 (0.9 to 1.7) 1.0 (0.8 to 1.3) 1.2 (1.0 to 1.5) 1.1 (0.7 to 1.8) 1.4 (0.8 to 2.4) (Ref)
Gram (2015)[43]	Cohort. The Multiethnic Cohort (MEC) Study. 83,300 women were enrolled and followed between 1993 and 2010. Of these, 4,484 developed invasive breast cancer. Smoking was assessed at baseline via questionnaire. Both current and former smoking were associated with slight increased risk of	Current vs. Never Former vs. Never	HR (95% CI) 1.1 (1.0 to 1.2) HR (95% CI) 1.0 (1.0 to 1.1)	Overall ≤20 21-30 >30 Never	HR (95% CI) 1.0 (0.9 to 1.1) 1.2 (1.1 to 1.3) 1.1 (1.0 to 1.2) (Ref)	Overall ≤10 11-20 >20 Never	HR (95% CI) 1.0 (0.9 to 1.1) 1.1 (1.0 to 1.2) 1.2 (1.0 to 1.3) (Ref)

Reference	Study Design and Overview	Temporal relative to date of study enrollment		Duration (years)		Dose (cigarettes/day)	
(continued)	breast cancer. Long smoking duration and high dose were also associated with increased risk of disease.						
Nyante (2014)[44]	Cohort. AARP (formerly American Association of Retired Persons). 186,150 female study participants were enrolled in 1995-96 and followed for an average of 10 years. Smoking exposure was assessed via baseline questionnaire. However, smoking duration was not assessed at baseline. Current and former smoking were associated with slight increased risk of breast cancer. There was no apparent trend between smoking dose and breast cancer risk.	Current vs. Never Former vs. Never	HR (95% CI) 1.2 (1.1 to 1.3) HR (95% CI) 1.1 (1.0 to 1.1)	Not reported.		Current Smokers 1-10 11-20 21-30 31-40 ≥41 Never Former Smokers 1-10 11-20 21-30 31-40 ≥41 Never	 HR (95% CI) 1.2 (1.0 to 1.3) 1.2 (1.1 to 1.4) 1.1 (0.9 to 1.2) 1.1 (0.9 to 1.4) 1.4 (0.6 to 1.5) (Ref) HR (95% CI) 1.1 (1.0 to 1.2) 1.0 (1.0 to 1.1) 1.2 (1.1 to 1.3) 1.2 (1.1 to 1.4) 1.1 (0.9 to 1.2) (Ref)

Reference	Study Design and Overview	Temporal relative to date of study enrollment		Duration (years)		Dose (cigarettes/day)	
Rosenberg (2013)[45]	Cohort. The Black Women's Health Study. 52,425 women were followed for 14 years between 1997 and 2009. 1,377 breast cancer cases occurred. Smoking was assessed at baseline via questionnaire. When compared with never active or passive smokers, current and former smoking were not associated with increased breast cancer risk. Women who smoked 20 pack-years had slight increased risk of breast cancer.	Current vs. Never Active or Passive Former vs. Never Active or Passive	IRR (95% CI) 1.1(0.8 to 1.3) IRR (95% CI) 1.1 (0.9 to 1.4)	Pack-years <10 10-19 20	IRR (95% CI) 1.1 (0.9 to 1.3) 1.0 (0.8 to 1.3) 1.2 (1.0 to 1.5)	Reported composite measure of pack-years.	
Xue (2011)[46]	Cohort. The Nurse's Health Study. 111,140 women were enrolled at baseline (1976) and were followed through 2006. 8,772 incident cases of breast cancer occurred. Active smoking exposure was assessed via baseline questionnaire and updated biennially. Current and former smoking were associated with increased risk of breast cancer.	Current vs. Never Former vs. Never	HR (95% CI) 1.1 (1.0 to 1.2) HR (95% CI) 1.1 (1.0 to 1.1)	Overall <20 20-39 ≥40 Never	HR (95% CI) 1.0 (1.0 to 1.1) 1.1 (1.0 to 1.1) 1.2 (1.0 to 1.3) (Ref)	Current Smokers 1-14 15-24 ≥25 Never Former Smokers 1-14 15-24 ≥25 Never	HR (95% CI) 1.0 (0.9 to 1.2) 1.1 (1.0 to 1.2) 1.1 (1.0 to 1.3) (Ref) HR (95% CI) 1.0 (1.0 to 1.1) 1.1 (1.0 to 1.2) 1.1 (1.0 to 1.2) (Ref)

Table 2.2. Associations between smoking and risk of breast cancer intrinsic subtype.

Reference	**Study Design and Overview**	**Molecular Platform** and definition of positive subtype		**Smoking and Biomarker Contrast**		**Measure of Effect**
Butler (2016)[47]	Case-control. The Carolina Breast Cancer Study. Study enrollment included 1,803 cases and 1,564 controls. Data on smoking exposure was obtained during a nurse-administered interview. Current smoking was associated with increased risk of Luminal breast cancer and slight decreased risk of Basal-like breast cancer.	Medical Records and IHC Staining	Luminal: ER+ and/or PR+, HER2- Basal-like: ER-, PR-, HER2-EGFR+ and/or CK 5/6+	Current vs. Never Current vs. Never	Luminal vs. Controls Basal-like vs. Controls	OR and 95% CI 1.5 (1.1 - 2.0) OR and 95% CI 0.8 (0.5 - 1.5)
Kabat (2011)[48]	Cohort. The Women's Health Initiative (WHI). 148,030 women aged 50-79 were enrolled between 1993 and 1998 and followed over an 8-year period. Smoking exposure was assessed at baseline via questionnaire. There was no apparent association between current or former smoking and TNBC risk.	Medical Records and Pathology Reports	TNBC: ER-/PR-/HER2-	Current vs. Never Former vs. Never	TNBC vs. Non-cases TNBC vs. Non-cases	HR and 95% CI 1.1 (0.7 - 1.7) HR and 95% CI 0.9 (0.7 - 1.2)

Reference	Study Design and Overview	Molecular Platform and definition of positive subtype		Smoking and Biomarker Contrast		Measure of Effect
Kawai (2014)[49]	Case-control. Seattle-Puget Sound metropolitan area. There were 960 cases and 938 controls. Smoking exposure was obtained via questionnaire and restricted to those that occurred prior to reference date. Current and former smoking were not associated with TNBC breast cancer risk.	ER+ or PR+: staining in ≥ 1% of tumor cells. HER2+: FISH 3+	TNBC: ER-/PR-/HER2-	Current vs. Never Former vs. Never	TNBC vs. Controls TNBC vs. Controls	OR and 95% CI 1.2 (0.7 - 2.1) OR and 95% CI 0.9 (0.6 - 1.5)
Millikan (2008)[13]	Case-control. The Carolina Breast Cancer Study. Study enrollment included 1,803 cases and 1,564 controls. Data on smoking exposure was obtained during a nurse-administered interview. Smoking duration was not differentially associated with Luminal A or Basal-like breast cancer.	Medical Records and IHC Staining	Luminal A: ER+ and/or PR+, HER2- Basal-like: ER-, PR-, HER2-EGFR+ and/or CK 5/6+	Years <10 11-19 20+ Never	Basal-like vs. Luminal A	OR (95% CI) 0.9 (0.6 to 1.5) 1.1 (0.7 to 1.7) 0.7 (0.5 to 1.1) (Ref)
Tariq (2014)[50]	Prospective cohort. Tumor registry at the University of Florida at Jacksonville (2000-2005). Smoking status (ever = current or past) was recorded in the tumor registry and was not associated with TNBC tumors.	Tumor registry	TNBC: ER-/PR-/HER2-	Ever vs. Never	TNBC vs. Non-TNBC	Proportion 20% vs. 28% p = 0.4

Reference	Study Design and Overview	Molecular Platform and definition of positive subtype		Smoking and Biomarker Contrast		Measure of Effect
Turkoz (2013)[51]	Cross-sectional. Department of Medical Oncology at Hacettepe University, Institute of Oncology. The study identified 1,884 invasive cases that were eligible for analysis. Smoking exposure was obtained during physician-led interview.	ER+ or PR+: staining in ≥ 5% of tumor cells. HER2+: FISH 3+	Luminal: ER+ or PR+ TNBC: ER-/PR-/HER2-	Ever vs. Never Ever vs. Never	Luminal vs. Non-Luminal TNBC vs. Non-TNBC	OR and 95% CI 1.0 (0.7 - 1.4) OR and 95% CI 1.0 (0.7 - 1.4)

Table 2.3. Associations between smoking and risk of ER-defined breast cancer.

Reference	Study Design and Overview	Molecular Platform and definition of positive subtype		Smoking and Biomarker Contrast		Measure of Effect
Cooper (1989)[52]	Case-control. Adelaide, South Australia. The study included 451 cases identified through the South Australian Central Cancer Registry between 1982 and 1984. There were 451 age-matched controls. Smoking exposure data was obtained in-person interview. Former smoking exposure was associated with increased risk of ER- breast cancer.	Saturation analysis assay	ER+: ≥ 10 fmol/mg	Current vs. Never Former vs. Never Current vs. Never Former vs. Never	ER+ vs. Controls ER+ vs. Controls ER- vs. Controls ER- vs. Controls	OR and 95% CI 1.3 (0.8 - 2.0) OR and 95% CI 0.9 (0.6 - 1.4) OR and 95% CI 1.3 (0.7 - 2.5) OR and 95% CI 1.9 (1.0 - 3.6)
Gaudet (2013)[53]	Cohort. The Cancer Prevention Study II (CPS-II) Nutrition Cohort. 97,786 women were enrolled in 1992 to examine cancer incidence and mortality. Median follow-up was 14 years 3,721 invasive breast cancers occurred. Both current and former smoking were associated with slight increased risks of breast cancer. Current and former smoking were associated with increased risk of ER+ breast cancer. However, neither were associated with risk of ER- breast cancer.	Medical Records and Pathology Reports	SEER (NOS)	Current vs. Never Former vs. Never Current vs. Never Former vs. Never	ER+ vs. Controls ER+ vs. Controls ER- vs. Controls ER- vs. Controls	OR and 95% CI 1.2 (1.0 - 1.5) OR and 95% CI 1.1 (1.0 - 1.3) OR and 95% CI 0.9 (0.5 - 1.4) OR and 95% CI 1.0 (0.8 - 1.2)

Reference	Study Design and Overview	Molecular Platform and definition of positive subtype		Smoking and Biomarker Contrast		Measure of Effect
Kabat (2011)[48]	Cohort. The Women's Health Initiative (WHI). 148,030 women aged 50-79 were enrolled between 1993 and 1998 and followed over an 8-year period. Smoking exposure was assessed at baseline via questionnaire. Former smokers had a slight increased risk of ER+ breast cancer.	Medical Records and Pathology Reports	ER+: NOS	Current vs. Never Former vs. Never	ER+ vs. Non-cases ER+ vs. Non-cases	HR and 95% CI 1.1 (0.9 - 1.3) HR and 95% CI 1.1 (1.1 - 1.2)
Kawai (2014)[49]	Case-control. Seattle-Puget Sound metropolitan area. There were 960 cases and 938 controls. Smoking exposure was obtained via questionnaire and restricted to those that occurred prior to reference date. Current and former smoking were associated with increased risk of ER+ breast cancer.	Medical Records and Pathology Reports	ER+: nuclear staining in ≥ 1% of tumor cells.	Current vs. Never Former vs. Never	ER+ vs. Controls ER+ vs. Controls	OR and 95% CI 1.4 (1.0 - 1.9) OR and 95% CI 1.3 (1.0 - 1.7)
Manjer (2001)[54]	Cohort. Malmo, Sweden. 10,902 women born between 1926 and 1949 with average age at enrollment of 50 years. Women were enrolled between 1974 and 1992 for average follow-up of 12 years; 268 incident cases with available tumor tissue	IHC	ER+: Immuno-reactivity	Current vs. Never Former vs. Never Current vs. Never	ER+ vs. Controls ER+ vs. Controls ER- vs. Controls	RR and 95% CI 0.9 (0.6 - 1.2) RR and 95% CI 1.0 (0.7- 1.5) RR and 95% CI 2.2 (1.2 - 4.0)

Reference	Study Design and Overview	Molecular Platform and definition of positive subtype		Smoking and Biomarker Contrast		Measure of Effect
(continued)	occurred. A self-administered questionnaire was used to obtain data on smoking exposure. Current and former smoking exposure was associated with increased risk of ER-, but not ER+, breast cancer.			Former vs. Never	ER- vs. Controls	RR and 95% CI 2.7 (1.4 - 5.0)
Morabia (1998)[55]	Case-control. Geneva, Switzerland. 372 cases diagnosed between 1992 and 1993; 1,059 controls were included. Data on smoking exposure was obtained during in-person interview. Ever smokers who smoked 20 or more cigarettes per day were at increased risk of developing ER+ and ER- breast cancer.	IHC	ER+: nuclear staining in ≥ 20% of tumor cells.	Ever 20+ cpd vs. Never Ever 20+ cpd vs. Never	ER+ vs. Controls ER- vs. Controls	OR and 95% CI 2.4 (1.4 - 4.5) OR and 95% CI 4.3 (1.4 - 13.0)
Nishino (2014)[56]	Case-control. Miyagi Cancer Center Hospital. 1,309 breast cancer cases and 3,878 controls. Smoking exposure was assessed via questionnaire. Current or former smoking was not associated with either of the 4 ER/PR subtypes.	Medical Records and Pathology Reports IHC and EIA	ER+/PR+: NOS	Current vs. Never	ER+/PR+ vs. Controls ER+/PR-vs. Controls ER+/PR-vs. Controls ER-/PR- vs. Controls	Null Associations for current and former smoking for all combinations of ER+/- and PR+/-.

Reference	Study Design and Overview	Molecular Platform and definition of positive subtype		Smoking and Biomarker Contrast		Measure of Effect
Nyante (2014)[44]	Cohort. AARP (formerly American Association of Retired Persons). 186,150 female study participants were enrolled in 1995-96 and followed for an average of 10 years. Smoking exposure was assessed via baseline questionnaire.	SEER Cancer Registry	ER+/PR+: NOS	Current vs. Never Current vs. Never Current vs. Never	ER+/PR+ vs. Controls ER-/PR+ vs. Controls ER-/PR- vs. Controls	OR and 95% CI 1.0 (0.9 - 1.2) OR and 95% CI 1.4 (1.0 - 1.8) OR and 95% CI 1.1 (0.8 - 1.4)

Table 2.4. Associations between smoking and risk of p53-defined breast cancer.

Reference	**Study Design and Overview**	**Molecular Platform** and definition of positive subtype		**Smoking and Biomarker Contrast**		**Measure of Effect**
Conway (2002)[57]	The Carolina Breast Cancer Study. 456 invasive breast cancer cases were evaluated for specific TP53 mutations. Of these, 108 breast cancers (or 24%) harbored specific TP53 mutations. 71% of mutations were missense; the remaining were deletions or insertions. Relative to non-smokers, current smokers were more likely to harbor p53 mutations.	Gene Sequencing	p53+: presence of somatic mutation in exons 4-8	Current vs. Never Former vs. Never	p53+ vs. p53- p53+ vs. p53-	OR and 95% CI 2.1 (1.2 - 3.8) OR and 95% CI 0.6 (0.4 - 1.2)
Furberg (2002)[58]	The Carolina Breast Cancer Study. The authors examined smoking exposure in relation to overexpression of p53 protein among 683 cases. There was no apparent association between smoking status and p53 tumor expression. The authors also investigated the associations between smoking dose, duration, and p53 expression and observed no apparent associations.	IHC	p53+: dark nuclear protein staining in ≥ 10% of tumor cells.	Current vs. Never Former vs. Never	p53+ vs. p53- p53+ vs. p53-	OR and 95% CI 0.8 (0.6 - 1.3) OR and 95% CI 1.0 (0.7 - 1.4)

Reference	Study Design and Overview	Molecular Platform and definition of positive subtype		Smoking and Biomarker Contrast		Measure of Effect
Gammon (1999)[59]	The Long Island Breast Cancer Study. Investigators examined the prevalence of p53 overexpression in breast tissues of young women (age < 45 years). Current smoking - but not former smoking - was associated with increased risk of p53 overexpression in breast tumors.	IHC	p53+: moderate to strong nuclear protein staining in ≥ 10% of tumor cells.	Current vs. Never Former vs. Never	p53+ vs. p53- p53+ vs. p53-	OR and 95% CI 2.0 (1.1 - 3.5) OR and 95% CI 1.4 (0.8 - 2.4)
Mordukhovich (2010)[60]	The Long Island Breast Cancer Study Project. 859 invasive breast tumors were evaluated for TP53 mutations (exons 5 - 8). Of these, 151 harbored a p53 mutation. Current and former smokers were slightly less likely to have a p53 mutation, when compared with never smokers.	Gene Sequencing	p53+: presence of somatic mutation in exons 5-8	Current vs. Never Former vs. Never	p53+ vs. p53- p53+ vs. p53-	OR and 95% CI 0.7 (0.4 - 1.2) OR and 95% CI 0.9 (0.6 - 1.4)

CHAPTER 3: METHODS

3.1 Overview

Breast cancer is not a single disease, but exists as a collection of genetically distinct subtypes. These subtypes have prognostic and predictive value in clinical settings and may also provide clues to breast cancer etiology. Homogenous classifications of breast cancer may have masked associations in prior studies of smoking and breast cancer risk; thus, a critical examination would benefit by evaluating smoking in relation to etiologically-relevant subtypes. In addition, it is important to consider dose and timing of exposure. By examining temporal and dose-dependent patterns of smoking, we may identify etiologically-relevant time periods that are associated with risk of specific breast cancer subtypes. Using data from phase III of the Carolina Breast Cancer Study (CBCS III), we seek to examine temporal and dose-dependent relationships between smoking exposure, breast cancer intrinsic subtype, and breast tumor expression of biomarkers linked to pathogenesis. Phase III is a case only study of approximately 3,000 women diagnosed with invasive breast cancer between 2008 and 2013 in central and eastern North Carolina. The CBCS includes self-reported data on smoking exposure and biomarker data obtained from tumor specimens, including protein and RNA expression data on genes with suspected links between smoking and breast carcinogenesis.

3.2 Study Design

Phase III of the Carolina Breast Cancer Study (CBCS) is a population-based case-only study that combines epidemiology and molecular biology to examine environmental and

genetic risk factors for breast cancer. Breast cancer cases were identified by a rapid case ascertainment system, implemented through collaboration between Lineberger Comprehensive Cancer Center (LCCC) and the North Carolina Central Cancer Registry (NCCCR). To be eligible for inclusion, patients must have been female and received a first and primary diagnosis of breast cancer between May 1, 2008 and October 31, 2013. The patient also must have resided in the 44-county study region and been between the ages of 20 and 74 at the time of diagnosis.

To examine potential risk factor differences by age and race, the CBCS employed a randomized recruitment strategy that was designed to oversample young and African American women. The patient's primary physician was contacted to obtain permission to invite the patient into the study and the overall expected response rate is 70%. Patients who declined participation were not substantially different from those who chose to participate in the study. In total, 2,998 women were enrolled in CBCS III. Study participants were asked to consent to a nurse-administered in-person interview that took place in the study participant's home or another pre-arranged location. During the in-person interview the nurse administered a questionnaire that included items on family and personal medical history, reproductive history, smoking, alcohol, diet, medication use and occupational history. Upon consent, the nurse also collected a blood sample and objective anthropometric measurements of height (m), weight (kg, waist (m), and hip (m) circumference. The average time between study enrollment and interview was 6 months.

At the time of interview, study participants were asked permission to obtain formalin-fixed, paraffin-embedded (FPPE) tumor blocks or tissue slides from the hospital where the diagnostic surgery was to be performed. Participants were also asked for permission to obtain

pathology reports and medical records from the treating facilities. Clinical and pathological data abstracted from medical records and pathology reports included tumor size, stage, and node status. For all cases, a single pathologist (Dr. Joseph Geradts) determined tumor grade.

3.3 Outcome Assessment

The CBCS includes protein and RNA expression data on genes used to define intrinsic subtype and genes involved in mechanistic pathways that may link smoking to breast cancer risk (Table 3.1). We will examine binary and continuous outcomes of biomarker expression for intrinsic, estrogen-mediated, genotoxic, and growth-factor dependent mechanistic pathways. Specifically, we will use gene expression signatures and protein-specific cut points to characterize tumors as positive (+) or negative (-) for the given biomarker or biomarker pathways. We will use continuous measures of protein and counts of RNA transcripts to examine whether smoking exposure modulates biomarker expression. Tissue microarrays (TMAs) were constructed for immunohistochemical (IHC) staining of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor-2 (HER2), cytokeratin 5/6 (CK 5/6), epidermal growth factor receptor (EGFR), and p53. Automated quantification of staining was performed using Genie classifier and protein-specific algorithms[61]. RNA was extracted from the same tumor specimens used to construct the TMAs, using the Qiagen RNeasy FFPE kit and protocol. RNA expression data were obtained via Nanostring assay. TMA construction and IHC analyses were conducted at Tissue Pathology Laboratory (TPL) and the Immunohistochemical Core Laboratory (ICL) at UNC Chapel Hill. Nanostring assays were performed in the laboratory of Dr. Melissa Troester.

Table 3.1. Molecular characterization of breast tumors in CBCS III

	Specific Aims	Aim 1		Aim 2	
Specific Aims	Pathway	a. Intrinsic	b. Estrogen-Mediated	a. Genotoxic	b. Growth-Factor Dependent
	Biomarker comparison	Luminal vs. Basal	ER+ vs. ER-	p53+ vs. p53-	EGFR+ vs. EGFR-
Molecular Characterization	Single gene IHC	Nielsen[37] (2004)	Hammond[62] (2010)	Williams[63] (2017)	10%
	Single gene mRNA	N/A	Continuous	NA	Continuous
	Multigene IHC signature	Nielsen[37] (2004)	NA	NA	NA
	Multigene mRNA signature	Parker[64] (2009)	NA	Troester[65] (2006)	NA

Note. NA-Not Applicable

3.3.1 Intrinsic subtype, multigene mRNA and IHC biomarkers

Breast cancer intrinsic subtype was measured using the RNA-based "PAM50 signature"[64]. Here, differential expression of the 50-gene signature is used to categorize breast cancers into 4 intrinsic subtypes: Luminal A, Luminal B, HER2E, and Basal-like. Each case was classified based upon highest Pearson correlation with a centroid defined for each subtype. For the analyses outlined in this proposal, we combined Luminal A, Luminal B, and HER2E tumors, since each have suspected luminal epithelial origins and may represent a single etiologic subtype[8]. Together, these tumors represent the Luminal breast cancer intrinsic subtype and will be compared to Basal-like breast tumors.

The joint expression of five clinical markers was used to define IHC intrinsic subtypes. These markers include: ER, PR, HER2, CK 5/6, and EGFR. Each marker is designated as positive (+) or negative (-), based on clinicopathologic cut points for

overexpression. Automated quantification of ER and PR protein expression was determined by Genie classifier and the Aperio nuclear v9 algorithm; HER2 and EGFR quantification was determined using the Aperio membrane v9 algorithm (Aperio Technologies, Vista, CA). Definiens Tissue Studio® was used to quantify CK5/6 staining. Intrinsic subtypes were designated as follows: Luminal breast cancers were defined as (ER+ and/or PR+, regardless of HER2 status); and Basal-like tumors were defined as (ER-, PR-, HER2-, EGFR+ and/or CK5/6+). We will compare Luminal and Basal-like breast tumors.

3.3.2 Estrogen-mediated biomarkers

We will use a cut point of ≥ 1% of tumor cells with nuclear protein staining for ER to define borderline/positive ER status, as recommended by the American Society of Clinical Oncology (ASCO) and College of American Pathologists (CAP)[62]. Specifically, we will define tumors with 1% to < 10% expression as 'borderline positive' and tumors with ≥ 10% expression as 'positive'. Automated quantification of ER protein expression was determined by a Genie classifier and the Aperio nuclear v9 algorithm (Aperio Technologies, Vista, CA). In addition, we will examine quantiles and continuous measures of ER protein staining in relation to smoking exposure. Specifically, we will examine whether smoking exposure modulates protein expression of ER in breast tumors. We will also examine whether smoking exposure modulates ER mRNA expression.

3.3.3 Genotoxic biomarkers

We will use the 48-gene signature identified by Troester et al. to classify breast tumors as p53 wildtype or p53 mutant[65]. The 48-gene predictor is applied to each case and classified based upon highest Pearson correlation with a centroid defined for either subtype.

The mutant or wildtype designations characterize downstream biologic activity following p53 loss or activation, respectively.

We will use a cut point of ≥ 10% of tumor cells with nuclear protein staining for p53 as described by Williams et al.[63]. Automated quantification of p53 protein expression was determined by a Genie classifier and the Aperio nuclear v9 algorithm (Aperio Technologies, Vista, CA). As described above for ER protein staining, we will also examine continuous measures of p53 protein staining in relation to smoking exposure.

3.3.4 Growth-factor dependent biomarkers

We will explore the relationships between categorical and continuous measures of EGFR protein and RNA and measures of smoking exposure. Automated quantification of EGFR protein expression was determined by a Genie classifier and the Aperio membrane v9 algorithm (Aperio Technologies, Vista, CA).

3.4 Exposure Assessment

History of smoking exposure was obtained during a nurse-administered in-person interview and includes data on smoking duration, frequency, and dose. Self-reported smoking is considered a valid measure of smoking exposure, with increased accuracy obtained during in-person interview formats[66]. Women in CBCS were considered ever smokers if they smoked at least 100 cigarettes during their lifetimes. CBCS investigators collected data on smoking history defined as ‘ever’ or ‘never’ (history); smoking status defined as ‘current’, ‘former’, or ‘never’ (status); age at smoking initiation measured in years (initiation); smoking duration measured as the total number of years of smoking between initiation and current use or cessation (duration); number of cigarettes smoked per day (dose); and age at smoking

cessation, where applicable (recency). Together, these data will be used to derive categorical, ordinal, and continuous measures of smoking (e.g. pack-years).

3.5 Covariate Assessment

We selected potential confounders of the relationship between smoking and breast cancer risk based on a review of the literature. These include: 1. First-degree family history of breast cancer defined as breast cancer diagnosis for mother or full female siblings[16]; 2. Alcohol consumption defined as number of drinks consumed per week[15,17,67]; 3. Breast feeding characterized by age at first breast feeding[13]; 4. Body Mass Index (BMI kg/m2)[13]; 5. Income to family ratio calculated as the household income divided by the number present in the household[68]; 6. Parity defined as number of full-term births[13,15]; 7. Years of oral contraceptive use[10]; 8. Years of hormone replacement therapy use[10]; 9. Physical activity; 10. Age; and 11. Race.

To identify a minimal adjustment set among the eleven potential confounders, we conducted directed acyclic graph analysis using DAGitty software (version 2.3). The smallest minimal adjustment set is as follows: alcohol, breast feeding, family history, income, parity, physical activity, age, and race. For continuous variables, we will explore various coding modalities to identify the coding scheme that results in the best fit and most parsimonious model. Covariates will be included in regression models as confounders based on whether there is a 10% or greater change in estimate when added to the model.

3.6 Data Analysis

This study examines the relationships between smoking, breast cancer intrinsic subtype, and biomarkers linked to pathogenesis. Specifically, we will examine smoking's temporal and dose-dependent associations with PAM50 subtype (Aim 1a) and breast tumor

expression of biomarkers linked to estrogen-mediated (Aim 1b), genotoxic (Aim 2a), and growth-factor mediated (Aim 2b) pathways. The examination of temporal and dose-dependent patterns of smoking in relation to biomarker-defined subtypes may identify exposure-time-windows that are associated with specific mechanistic events.

3.6.1 Linear and logistic regression models

We will use linear regression to model the relationship between smoking exposure and continuous measures of protein or RNA expression. We will test assumptions of linearity by examining higher order polynomials of smoking exposure. Smoking exposure will be modeled as a continuous measure of pack-years, defined as the product of the number of cigarettes smoked per day (dose) and the total number of years smoked (duration). We will examine the association between pack-years and biomarker expression defined as weighted percent cells of positive and the b) H-score. An H-score is a weighted measure of the number of weakly (1), moderate (2), and strongly stained (3) cells in the tissue sample and the ranges between 0 and 300. In addition to measures of percent positivity we will also examine pack-years in relation to quantiles of gene expression using generalized logit models with ordinal outcomes.

3.6.2 Binary outcomes and categorical measures of smoking

For each binary breast cancer classification, we will use generalized logit models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for the relationship between smoking exposure and subtype-specific breast cancer risk. This analysis will consider cumulative measures of smoking exposure. Breast cancer characterizations are described in Table 3.1.

3.6.3 Cumulative smoking exposure and time-windows analysis

For each binary breast cancer classification, we will use generalized logit models to estimate ORs and 95% CIs for risk of breast cancer subtype defined as a trend per pack-year for a) cumulative pack-years smoked and b) within exposure-time-windows defined by time since exposure. c) We will also use latency functions and associated graphs to visualize the temporal relation between smoking exposure and subtype-specific breast cancer risk. Evidence of an association between smoking and breast cancer risk proximal to time of diagnosis may allow us to infer that smoking is associated with a cancer promotion event for a given biomarker pathway. By contrast, if smoking exposure is associated with breast cancer risk at a distal point from date of diagnosis, we may infer that smoking is associated with cancer initiation or promotion events. Previous case-control studies of smoking and breast cancer risk have described an increased risk of disease among women who quit smoking 5 to 10 years prior to date of case/control selection[42]. For our analysis, we selected three time-windows that would accommodate available sample size: <10 years; 10-20 years; > 20 years. We will conduct likelihood ratio tests to determine whether our time windows analyses improve model fit, when compared cumulative exposure models.

3.6.4 Parametric latency functions of smoking and breast cancer risk

For each binary breast cancer classification, we will use generalized logit models with parametric latency functions to estimate ORs and 95% CIs for risk of breast cancer subtype. We emphasize recency of exposure in this analysis, and we hypothesize that associations between smoking exposure proximal or distal to date of diagnosis may differ across biomarker-defined subtypes. The inclusion of a parametric latency function in our logistic regression model will allow us to calculate time weighted exposure estimates using

maximum likelihood estimation (MLE); the highest weights are assigned during the time period where smoking is associated with the greatest risk of subtype-specific breast cancer. That is, our modeling strategy will allow us to identify peak risk according to recency of exposure for each biomarker-defined breast cancer type. We will examine hierarchical models with bilinear and log-normal latency functions. If our B-spline from analysis 2c demonstrates a multimodal relationship between smoking exposure and breast cancer risk, we will explore different probability distributions for our parametric latency function.

3.7 Power Analysis

For all power calculations, we treat smoking exposure as a time-invariant variable, dichotomized as 'ever smoker' vs. 'never smoker'. Based on smoking prevalence in CBCS phases I and II, we estimate that approximately 50% of women in phase III will have a history of ever smoking. For each binary breast cancer classification, we calculated the power to detect a statistically significant association for a theoretical range of case only odds ratio comparing smoking exposure and odds of having breast cancer subtypes: luminal vs. basal, ER+ vs. ER-, p53+ vs. p53-, and EGFR+ vs. EGFR-. IHC data will be available for the entire CBCS III study population (n=2,000) and RNA data will be available for a subset of study participants (n=1,000). (Recall that RNA data include gene signatures and counts of single RNA transcripts.) Table 3.2 describes the expected distributions of subtypes in CBCS III. Early estimates from phase III demonstrate a distribution of 4:1 for Luminal vs. Basal-like; 4:1 for ER+ vs. ER-; 1: 2 for p53+ vs. p53-; and 1:3 for EGFR+ vs. EGFR-. Figure 1 displays power distributions for a theoretical range of odds ratios for IHC and RNA measurements and each case-case comparison. Power calculations were performed in SAS 9.4 (SAS Institute Inc., Cary, NC).

Table 3.2. Expected distributions of subtypes in CBCS III.

Specific Aim	Contrast	Ratio*
Aim 1a.	Luminal vs. Basal	4:1
Aim 1b.	ER+ vs. ER-	4:1
Aim 2a.	p53+ vs. p53-	1:2
Aim 2b.	EGFR+ vs. EGFR-	1:3

*Rounded to the nearest whole number.

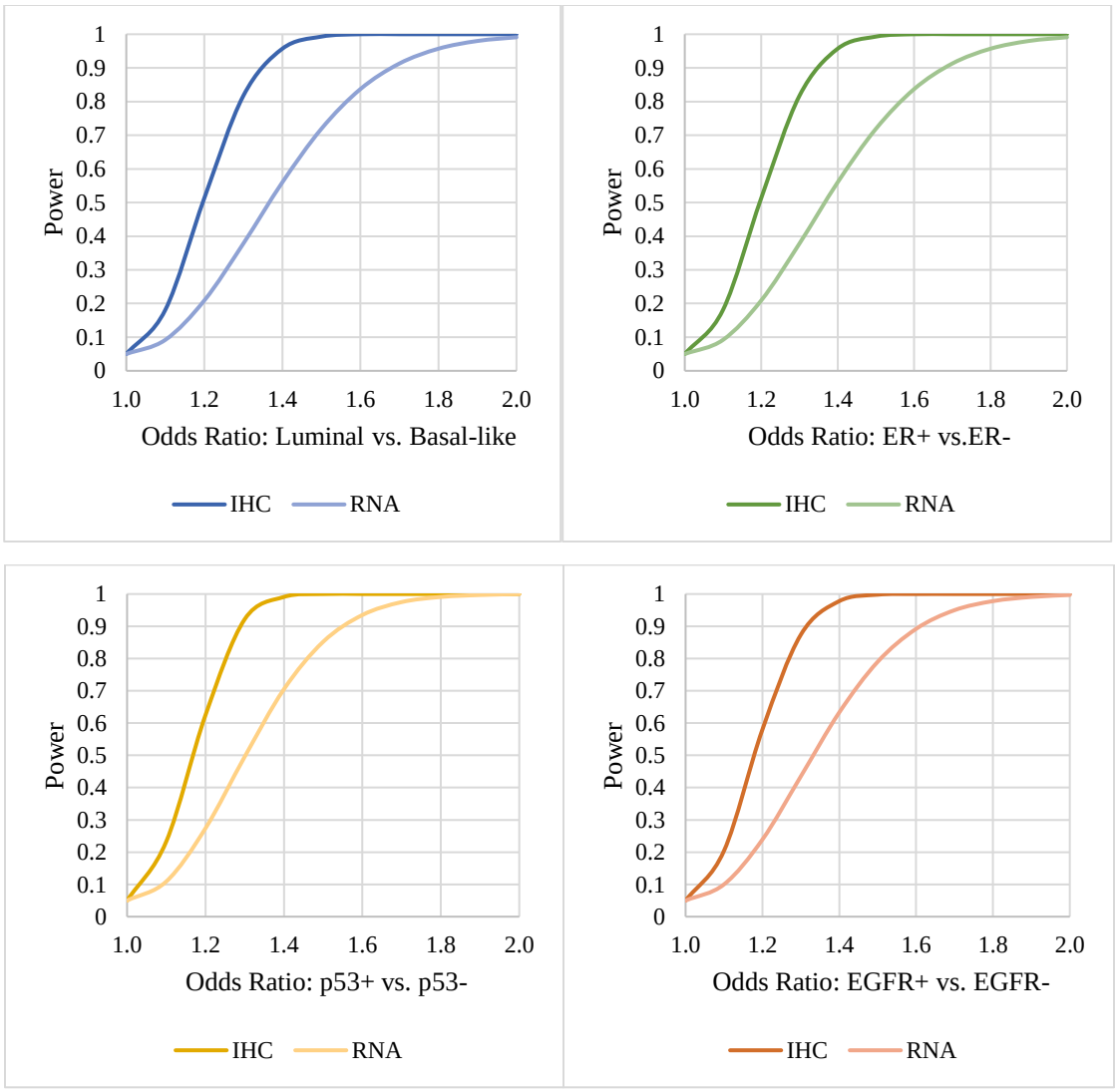


Figure 3.1. Power distributions for theoretical case-case odds ratio.

3.8 Summary

3.8.1 Limitations

Although our study has several benefits, it is important to acknowledge limitations. The CBCS III study design did not include the recruitment of a control group that could be used to assess the baseline exposure distribution. Thus, our case-case odds ratios use a subset of cases as the referent group and the distribution of smoking exposure in this group may not reflect that of the population from which the case groups arise. To address this potential pitfall, we will perform a sensitivity analysis using controls from CBCS phases I and II. Controls from these study time periods were enrolled between 1993 and 2001. Cases in phase III were enrolled between 2008 and 2013. We will use generalized logistic regression functions with polytomous outcomes to calculate both case-control and case-case odds ratios. If the case-case odds ratios in our sensitivity analysis are similar to those which are observed in our phase III case-only analysis, we may infer that controls from the early phases are exchangeable for controls that would have been collected for phase III and are suitable to be used in phase III to evaluate associations between risk factors and subtype-specific breast cancer risk. If the case-case odds ratios from the primary and sensitivity analyses are not comparable, we will report the case-case odds ratios and include a detailed description of this limitation.

3.8.2 Strengths

With the unique compilation of observational and breast tumor biomarker data, CBCS is an idea resource to examine the association between smoking and breast tumor biomarkers linked to pathogenesis. By examining temporal relationships between smoking and breast cancer risk, we may identify exposure-time-windows that are most critical for disease

development. Further, examining temporal patterns of smoking in relation to biomarker-defined breast cancers may allow us to infer whether the expression of a given biomarker is associated with smoking exposures that are proximal or distal to time of diagnosis. Together, the strengths of this study may help to elucidate associations masked in prior investigations.

3.9 Addendum

During the conceptualization of the dissertation research, we planned to evaluate protein biomarker intensity in the form of an H-Score. However, due to properties of the staining protocol and automated quantification, it was difficult to distinguish cells that were stained at 2+ intensity vs. 3+ intensity. Thus, we focused our analyses on counts of cells that stained positive for the antigen.

CHAPTER 4: SMOKING AND ESTROGEN-MEDIATED BIOMARKERS

4.1 Introduction

Epidemiologic studies have demonstrated distinct risk factor profiles for breast cancer subtypes classified according to estrogen-receptor (ER) status[11,13]. Defined by an immunohistochemical (IHC) threshold of 1 to 10 percent staining of examined tumor cells, ER-positive (+) breast tumors account for more than 70% of all breast cancer cases diagnosed in the United States (US), making this disease group an important public health focus[4,62]. Pre-diagnostic smoking exposure has been linked to the ER+ subtype in some epidemiologic studies, with increased risks ranging between 10% to 50%[44,49,69]. Further, prospective studies of breast cancer survivors have suggested that smoking exposure prior to diagnosis may influence survival outcomes, particularly among women with ER+ disease, and presumably through reduced efficacy of anti-estrogenic therapies. For example, in a prospective study of Swedish breast cancer patients, ER+ women who smoked before diagnosis and who were treated with an aromatase inhibitor (AI) had a 3-fold increased risk of experiencing distant metastases or death, when compared to ER+/AI-treated non-smokers [Hazard Ratio (HR) and 95% CI (3.0 and 1.4 to 6.1)][70]. Together, these epidemiologic and clinical findings could be used to suggest that smoking exposure may be linked to estrogen metabolism in breast tumors and subsequent regulation of ER.

Although growing evidence suggests smoking as a possible risk factor for breast cancer, the mechanistic events leading to breast tumor initiation or promotion have not been

clearly defined[71]. Thus, if smoking is linked to breast cancer via estrogen-mediated pathways, it is plausible that quantitative levels of ER may differ between breast tumors of smokers and non-smokers. However, IHC assays are highly sensitive for the detection of ER leading to saturated signals and suppression of ER's dynamic range for gene expression[72]. RNA assays, which are less susceptible to saturation, may allow for improved assessments of smoking and quantitative ESR1 expression. In addition, multigene scores such as the PAM50 Luminal gene signature[64,72] reflect cross-sectional measures of estrogen-signaling in breast tumors, and may offer improved resolution when examining smoking in relation to ER expression.

Binary classifications of breast tumors with respect to ER status represent two distinct classes of disease characterized by age-incidence patterns and cells of origin (i.e., ER+ vs. ER–)[7]. As such, subtype-specific evaluations of smoking and breast cancer risk have demonstrated possible links to the ER+ etiologic type, but may mask quantitative associations between smoking and continuous measures of ER within tumors. In this study, we sought to evaluate smoking and its association with binary classifications and quantitative measures for ER protein, ESR1 mRNA, and a multigene score that serves as a cross-sectional measure of estrogen-signaling patterns in tumors; we examined temporal and dose-dependent measures of smoking in relation to each biomarker. Findings from our study may have implications for future studies that seek to evaluate smoking exposure in relation to hypothesized etiologic biomarkers in breast tumors.

4.2 Methods

4.2.1 Study Population

Phase III of the Carolina Breast Cancer Study (CBCS III) is a population-based case-only study that combines epidemiology and molecular biology to examine environmental and genetic risk factors for molecular subtypes of breast cancer. To be eligible for inclusion, patients must have been female and received a first and primary diagnosis of breast cancer between May 1, 2008 and October 31, 2013. The patient also must have resided in the 44-county study region and been between the ages of 20 and 74 at the time of diagnosis. To examine potential risk factor differences by age and race, the CBCS employed a randomized recruitment strategy that was designed to oversample young and African American women[73].

Breast cancer cases were identified by a rapid case ascertainment system, implemented through collaboration between Lineberger Comprehensive Cancer Center (LCCC) and the North Carolina Central Cancer Registry (NCCCR). Briefly, CBCS contacted the patient's primary physician to obtain permission to invite the patient into the study, yielding an overall response rate of 70% and a total of 2,998 women. Study participants were asked to consent to a nurse-administered in-person interview that took place in the study participant's home or another pre-arranged location. The average time between study enrollment and interview was 6 months. The nurse administered questionnaire included items on family and personal medical history, reproductive history, smoking, alcohol, diet, medication use and occupational history. Upon consent, the nurse also collected a blood sample and objective anthropometric measurements of height (m), weight (kg, waist (m), and hip (m) circumference.

4.2.2 Study Design

Outcome Assessment

The CBCS includes protein and RNA expression data on genes involved in estrogen-signaling. At the time of interview, investigators asked study participants for permission to obtain formalin-fixed, paraffin-embedded (FPPE) tumor blocks or tissue slides from the hospital where the diagnostic surgery was to be performed. Tumor blocks were used to construct tissue microarrays (TMAs) for IHC staining, where each patient's tumor was represented by 1 to 8 cores on the microarray. RNA was extracted from the same tumor specimens used to construct the TMAs, using the Qiagen RNeasy FFPE kit and protocol. CBCS includes data for 1970 women included in the IHC analysis and 1,011 women included in the RNA analysis.

Estrogen receptor protein. Automated quantification of ER protein was determined by a Genie classifier and the Aperio nuclear v9 algorithm (Aperio Technologies, Vista, CA)[61]. We calculated percent positivity for ER as the product of positively stained tumor cells for each core, multiplied by its core-specific weight, summed across all cores per patient (ER WT%). We assigned a cut point of ≥ 10% for 'ER positive' tumors; 1% to < 10% for 'ER borderline' tumors; and < 1% for 'ER negative' tumors. For the ER binary classification, 'ER borderline' tumors were combined with 'ER negative tumors'.

ESR1 mRNA. ESR1 was quantified using Nanostring technology. Briefly, total ESR1 mRNA counts were assayed using an ESR1-specific molecular probe, which hybridizes to RNA fragments in solution. Hybrids are then counted using microscopic imaging, yielding raw mRNA counts. Quality control and data normalization were performed using the NanostringNorm R package[74]. Data were first normalized to the geometric means of 6

internal positive controls and subsequently to the geometric means of 5 reference genes. Normalized ESR1 counts were log2 transformed, yielding a bimodal Gaussian distribution of the data. We used the mclust R package to classify the two distributions as ESR1– or ESR1+, reflecting low and high expression, respectively[75]. ESR1- tumors had log2 values ranging between 0 to 8.35 and ESR1+ tumors had log2 values ranging between 8.38 to 15.64.

Luminal Score. The PAM50 Luminal gene signature includes 8 highly correlated genes associated with Luminal type breast cancers, and which are characterized by ER expression[64,72]. The 8 genes include: BAG1, ESR1, FOXA1, GPR160, NAT1, MAPT, MLPH, and PGR. Each gene was quantified and normalized according to procedures for ESR1, as described above. To calculate the Luminal Score (LS), we took the average of the normalized values of the 8 genes. Normalized and transformed values for LS followed a bimodal Gaussian distribution. We used the mclust R package to classify the Luminal Score as LS– or LS+, reflecting low and high scores, respectively. LS– tumors had log2 values ranging between 3.26 to 7.57 and LS+ tumors had log2 values ranging between 7.58 to 11.37. ESR1 mRNA and the 8 genes embedded in the Luminal Score were assayed along with other genes included in 1 of 3 Nanostring batches or code sets. Thus, all Nanostring analyses were adjusted for 'code set' in order to minimize potential batch effects.

Exposure Assessment

History of smoking exposure was obtained during the nurse-administered in-person interview and includes data on smoking duration, frequency, and dose. Self-reported smoking is considered a valid measure of smoking exposure, with increased accuracy obtained during in-person interview formats[66]. Women in CBCS were considered ever smokers if they smoked at least 100 cigarettes during their lifetimes. CBCS investigators collected data on

smoking history defined as ‘ever’ or ‘never’ (history); smoking status defined as ‘current’, ‘former’, or ‘never’ (status); age at smoking initiation measured in years (initiation); smoking duration measured as the total number of years of smoking between initiation and current use or cessation (duration); number of cigarettes smoked per day (dose); and age at smoking cessation, where applicable. Pack-years were defined as a cumulative measure of the number of cigarette packs smoked per day, divided by smoking duration in years. Similarly, pack-decades were defined as a cumulative measures of cigarette packs smoked per day, over 10-year intervals.

Covariate Assessment

We selected potential confounders of the relationship between smoking and breast cancer risk based on established risk factors for breast cancer and study design variables. Potential confounders include: first-degree family history of breast cancer defined as breast cancer diagnosis for mother or a full female sibling[16]; alcohol consumption defined as having any history of alcohol use[15,17,67]; ever having breast fed [13]; body mass index (BMI kg/m2)[13]; parity defined as number of full-term births[13,15]; history of oral contraceptive use[10]; hormone replacement therapy use[10]; menopausal status; meeting physical activity guidelines; age; and race.

Participants were also asked for permission to obtain pathology reports and medical records from the treating facilities. Clinical and pathological data abstracted from medical records and pathology reports included tumor size, stage, and node status; these tumor characteristics were considered as potential confounders of the relationship between smoking and ER expression. For all cases, a single pathologist (Dr. Joseph Geradts) determined tumor grade.

4.2.3 Data Analysis

For each binary breast cancer classification, we used generalized logit models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for categorical measures of smoking and the ER, ESR1, and LS-defined subtypes. To evaluate temporal and dose-dependent associations between smoking and subtype, we first estimated the associations via logistic regression for a one unit increase in pack-decades, defined as the number of cigarette packs smoked per day over a 10-year period. We compared this cumulative exposure model to an exposure-time-windows model (i.e., piecewise logistic regression model) for three windows, with respect to time of diagnosis: < 10 years; 11-20 years; and > 20 years. We used a likelihood ratio test (LRT) to compare the deviances between the two models, the difference of which follows a chi-square distribution with 2 degrees of freedom.

To evaluate the hypothesis that odds of ER, ESR1, and LS-positive subtypes vary with time since smoking exposure, we used a generalized logit model with a lognormal latency function to calculate time weighted exposure estimates for the 40-year period preceding breast cancer diagnosis. The latency period between smoking exposure and breast cancer occurrence is thought to be as much as 40 years and the lognormal distribution has been used to describe variation in risk with time since exposure other studies of cancer etiology[76]. Specifically, the lognormal latency function can be used to describe the rise, peak, and decline in risk or log-odds with respect to time since exposure. The highest weights are assigned during the time interval where smoking is associated with the greatest odds of ER+, ESR1+, or LS+ breast cancer and may signify the most etiologically-relevant time interval for smoking exposure. The macro used to model the lognormal latency function is described in Richardson (2009)[77].

We used linear regression to model the relationship between continuous measures of ER ESR1, LS and categorical measures of smoking, adjusted for age, race, and Nanostring code set (where applicable). We calculated the estimated value of continuous biomarker expression for each individual, based on coding of the smoking exposure and covariate pattern. To isolate the effect of each smoking measure, we subtracted the effects of age, race, and Nanostring code set from the overall expression level. Expression levels for each biomarker were described according to interquartile range and visualized using box plots within categories of smoking.

All analyses were conducted using SAS 9.4 (SAS Institute Inc, Cary, NC) and R version 3.3.3.

4.3 Results

Figure 4.1 illustrates the relationships between categorical and continuous measures for luminal score, ESR1 mRNA, and ER protein. (A) The EM-algorithm identified distinct clusters for ESR1 and LS, reflecting low and high expression for each. We compared binary classifications for ER protein as measured by IHC to those for ESR1 and LS and observed moderate to good values for sensitivity (se) and specificity (sp) (ER vs. ESR1: se=92%, sp=86% and ER vs. LS: se=89%, sp=85%). (B) Weighted percent ER protein (ER WT%) was positively correlated with log2 values for ESR1 mRNA (r=0.70, $p < 0.01$). ESR1 mRNA appeared to have a greater dynamic range compared to ER WT% where values for ER+ tumors tended to saturate the upper end of the percentage range.

Figure 4.2 displays estimated odds ratios and 95% confidence intervals for smoking and binary breast cancer classifications for ER, ESR1, and LS, adjusted for traditional breast cancer risk factors. (A) ER+ breast tumors were most common among ever smokers

compared to never smokers (OR = 1.51 95% CI: 1.15, 1.97). When stratified by smoking status at time of diagnosis, current smokers were twice as likely to be ER+ compared to never smokers (OR=1.89 95% CI: 1.33, 2.70); former smokers had an elevated, though statistically non-significant, odds of ER+ breast cancer (OR = 1.25 95% CI: 0.91, 1.73). In addition, smoking duration of 20 years or more was linked to elevated odds of ER+ breast cancer (OR = 1.79 95% CI: 1.26, 2.56). We also observed elevated odds of the ER+ subtype for shorter intervals of smoking, though estimates were statistically non-significant. Women who smoked '<1/2' or '1/2 to 1' packs of cigarettes per day had increased odds of ER+ breast cancer [(OR = 1.48 95% CI: 1.04, 2.10) and (OR = 1.57 95% CI: 1.09, 2.26), respectively]. However, for the highest category for smoking dose (> 1 pack/day), we observed an elevated but statistically non-significant odds for ER+ tumors (OR = 1.44 95% CI: 087, 2.37). With respect to 'time since smoking', smoking within 5 years of breast cancer diagnosis was associated with a 60% increased odds of having ER+ breast cancer (OR 1.59 95% CI 1.15, 2.20). (B, C) In general, we observed similar patterns of association between smoking measures and the ESR1+ and LS+ subtypes. Notably, the magnitudes of the ORs were slightly higher for the RNA-based measures, particularly for smoking duration and time since smoking exposure.

Table 4.1 presents estimated ORs for cumulative smoking exposure overall and within exposure time-windows. Our cumulative exposure models demonstrated that a 1-unit increase in pack-decades was associated with a 10% to 18% increase in the odds of having a 'positive' subtype: ER+ (OR = 1.09 95% CI: 0.99, 1.20), ESR1+ (OR = 1.18 95% CI: 1.04, 1.34), and LS+ (OR = 1.18 95% CI: 1.04, 1.35). Moreover, for the exposure time-windows models, total pack-decades smoked within 10 years of a breast cancer diagnosis was

associated with the greatest odds of having ER+, ESR1+, or LS+ breast cancer when compared to exposure accumulated between 10 and 20 or greater than 20 years prior to diagnosis. However, results from our likelihood ratio test suggest that the exposure time-windows model provides improved fit over the cumulative exposure model for LS-defined subtypes (LRT = 6.39, 2 df), but did not substantially improve data fit for the ER (LRT = 0.94, 2df) and ESR1 subtypes (LRT = 2.59, 2 df).

Figure 4.3 illustrates variation with time-since-exposure for the association between pack-decades and LS+ breast cancer for the 40-year period preceding breast cancer diagnosis. Our latency model with lognormal weighted exposures demonstrated increased odds of the LS+ subtype for pre-diagnostic smoking proximal to time of diagnosis. A likelihood ratio test comparing the lognormal latency model to the cumulative exposure model for the same 40-year period did not suggest that our latency model provided improved fit for the data (LRT = 4.2, 2 df). However, the dose-response parameter estimate in our latency model was statistically significant, thereby suggesting the peak in odds proximal to diagnosis may be the most etiologically relevant time point for smoking and ER+ breast cancer occurrence.

Table 4.2 and Table 4.3 present estimated biomarker expression values for ER protein, ESR1 mRNA, and the luminal score, adjusted for age, race, and Nanostring code set (where applicable). In general, ER protein levels did not vary across smoking exposures for breast cancer cases overall or when restricted to ER+ cases. Compared to never smokers, we observed the highest levels of ESR1 mRNA and the luminal score among current smokers [(mean (log2) = 10.0 vs. 9.4, p = 0.02) and (mean (log2) = 8.8 vs. 8.5, p = 0.01), respectively]. When restricted to ER+ breast cancer cases, we still observed higher levels of

ESR1 among current smokers, however the luminal score value for this group was attenuated. Figure 4.4, Figure 4.5, and Figure 4.6 visualize distributions for each biomarker among 'Never', 'Former', and 'Current' smokers.

4.4 Discussion

Findings from our study lend support to the hypothesis that smoking could be linked to breast cancer via estrogen-mediated pathways. In our case-only study of nearly 2,000 patients, we observed increased odds of the ER+ subtype for temporal and dose-dependent measures of smoking. We also demonstrate that these associations hold for ER-related subtypes characterized by ESR1 mRNA and a multigene luminal score (LS). Increased odds of ER+, ESR1+, and LS+ subtypes was most apparent among women who were self-reported current smokers at time of diagnosis. Logistic regression models with latency parameters allowed us to simultaneously model dose, duration, and time of exposure to demonstrate that the most etiologically relevant period for smoking and ER-defined breast cancer may be during pre-diagnostic smoking exposure proximal to time of diagnosis. In addition, we observed that current smoking was associated with increased quantitative levels for ESR1, but not ER protein, which may suggest that RNA more sensitively captures biological differences when compared to ER protein expression.

Contemporary epidemiologic studies have demonstrated positive associations between smoking and ER+ breast cancer with estimates ranging between 10%-50% increased risk among current or former smokers[42,48,49,69]. Our case-only analysis demonstrated a near doubling for the odds of having ER+ vs. ER- breast cancer among current smokers. These findings are complemented by our previous case-control analysis in the Carolina Breast Cancer Study, which showed increased risk of ER+ disease among smokers and

heterogeneity of ORs for the Luminal (ER+) and Basal-like (ER-) subtypes[69,78]. Notably, we observed a positive association between smoking and ER+ risk but no association between smoking and the ER- subtype in the case-control setting – a pattern which was observed in other studies performed in US populations. However – in stark contrast – older studies of smoking and breast cancer risk in Swedish, Swiss, and Australian populations have demonstrated positive associations between smoking and the ER- breast cancer subtype[52,54,55]. These conflicting observations may reflect temporal differences in exposure, behavioral patterns, or may also be an artifact of differing methods used to assay ER expression (e.g., ligand-binding, immunoreactivity). Thus, a careful investigation of the relationship between smoking and ER-defined breast cancer subtypes should consider era, methodological approaches, and characteristics of population of interest.

In both clinical and research settings, immunohistochemistry has been used as the standard for estrogen-receptor quantification in breast tumors[79]. IHC is highly sensitive for the detection of ER protein, serving as an excellent marker of antigenicity, but may be less than ideal for accurate quantification of positively stained cells due to assay saturation and challenges of ER quantification by digital imaging. Our study addresses this potential limitation by using ESR1 mRNA counts assayed via Nanostring technology to characterize breast tumors as ESR1+ and ESR1-. Unlike ER expression values for percent positivity, the log2 transformed ESR1 mRNA counts in our study follow a bimodal Gaussian distribution, identifying two distinct classes of breast tumors reflecting low and high expression of ESR1. However, percent agreement between ER and ESR1 subtype classifications was good. And we observed similar patterns of association between smoking measures and the ER+ and ESR1+ subtypes, where current smoking, long smoking duration of more than 20 years, and

smoking within 5 years of a breast cancer diagnosis was associated with 2 to 3 times the odds of having a positive (+) subtype.

In addition, our study benefits by the incorporation of a multigene luminal score embedded in the PAM50 signature, used to classify breast tumors according to intrinsic subtype[64,72]. The 8 genes included in the luminal score are highly correlated with Luminal subtypes, which are characterized by high estrogen-receptor expression. Multigene scores may offer improved resolution over single gene markers as these cross-sectional assessments are often predictive, prognostic, and may have etiologic relevance. We observed similar patterns of association between smoking and the LS subtypes though the magnitudes of association were slightly higher for the ESR1 and LS mRNA classifications.

Although the prevalence of cigarette smoking has steadily decreased since the 1950s, approximately 50% of women in the United States report a history of ever smoking and 14% are self-reported current smokers[21]. For protracted exposures in studies of etiology, it is important to evaluate measures of dose, duration, and temporality in order to fully evaluate associations with the outcome. Women with the longest smoking histories in our study were older compared to never smokers and were also most likely to be self-reported current smokers at time of diagnosis. As such, traditional metrics for smoking in studies of cancer etiology are confounded by age, dose, and duration of exposure thereby creating a challenge in understanding how combination of dose and timing influence biomarker expression in breast tumors.

In the present study, we use time constant (cumulative) and time-varying (latency) models to simultaneously evaluate dose, duration, and timing of exposure; we observed that pre-diagnostic smoking proximal to time of diagnosis may be associated with increased odds

of ER+, ESR1+, and LS+ subtypes. We also observed higher quantitative levels of ESR1 among current smokers and women who smoked within 5 years of breast cancer diagnosis. In the absence of longitudinal measures for smoking and biomarker expression, this observation may allow us to hypothesize that smoking could be linked to promotion events, as opposed to initiation events that would likely be linked to more distal smoking exposures. Indeed, several clinical studies have reported decreased efficacy of both estrogenic therapy in postmenopausal women and antiestrogenic therapy among women with breast cancer, suggesting that smoking may interfere with estrogen metabolism. Researchers have also suggested that fluctuations in endogenous estrogens may influence intrinsic subtyping. Thus, it is plausible that exogenous exposures that influence estrogen metabolism may modulate estrogen-receptor expression, which may have implications for intrinsic subtyping of breast tumors among smokers. Interestingly, we did not observe associations between smoking and quantitative ER expression. Protein and mRNA reflect distinct processes in gene expression and also require differing methods for evaluation and quantification. The high sensitivity of ER IHC assays may yield this method less suitable for quantification of ER protein.

With the unique compilation of observational and biomarker data, CBCS is an ideal resource to examine the association between smoking and ER-defined breast cancer subtypes. Findings from our study add a unique contribution to the body of literature by considering multiple methods to characterize ER-defined breast tumors and by incorporating measures of time, duration and dose to identify etiologically relevant exposure periods. We identified what may be an etiologically relevant exposure period for smoking proximal to time of diagnosis. And we also suggest that RNA measures may provide improved resolution of gene expression for studies seeking to evaluate the etiology of ER+ breast tumors. Future

work should seek to examine smoking in relation to other proposed biomarkers of breast carcinogenesis.

4.5 Addendum

When developing the analytic plan for Aim 1, we identified the possibility of evaluating smoking exposure in relation to the multigene luminal score, though it was not described in our initial methodology plan. In addition, we also examined associations between smoking and breast cancer intrinsic subtype; patterns of association comparing Luminal and Basal-like subtypes were similar to that for our luminal score comparisons (i.e., LS+ vs. LS-). We elected to present results for the luminal score measure and have omitted results for the intrinsic subtype comparisons. With respect to estimated odds ratios, we adjusted for node status, tumor size, tumor stage, and tumor grade in addition to covariates described in Chapter 3. We hypothesized that these tumor characteristics may influence the expression of ER and ER-related genes.

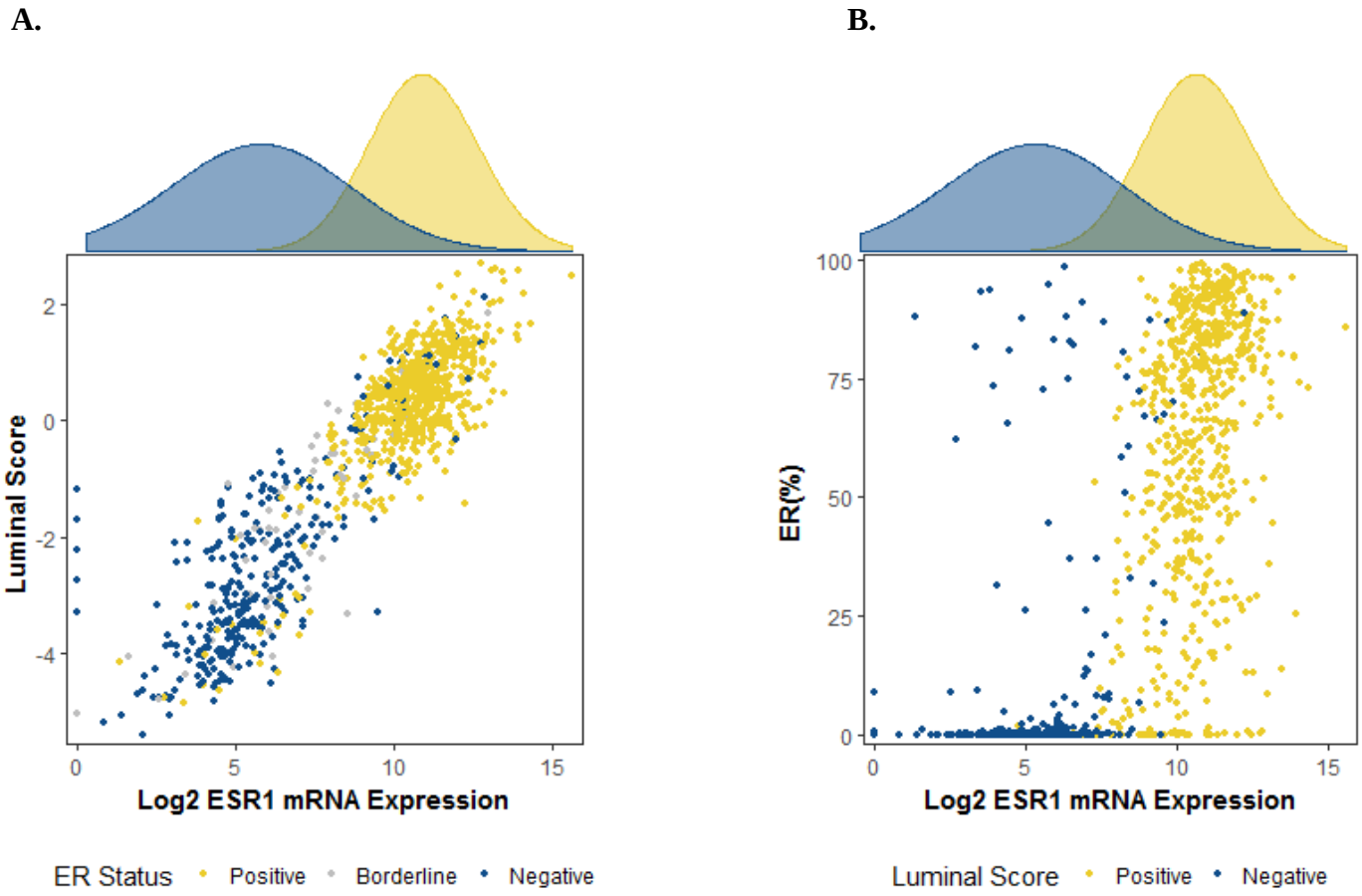


Figure 4.1. Relationships between ER IHC status, *ESR1* mRNA expression (log2), and luminal score (median-centered).

Note. (A) ER positive breast tumors are colored yellow (≥10% weighted positivity); ER borderline tumors are colored gray (1% to < 10% weighted positivity); and ER negative tumors are colored dark blue (< 1% weighted positivity). (B) Scatterplot showing the relationship between ER weighted percent positivity (%), *ESR1* mRNA expression (log2), and luminal score binary classifications (i.e., LS+ and LS-). ER IHC and *ESR1* mRNA values were positively correlated (r=0.70, p <0.01). An expectation-maximization (EM) algorithm identified two distinct clusters for *ESR1* expression (*ERS1-*, dark blue; *ESR1+*, yellow).

A.

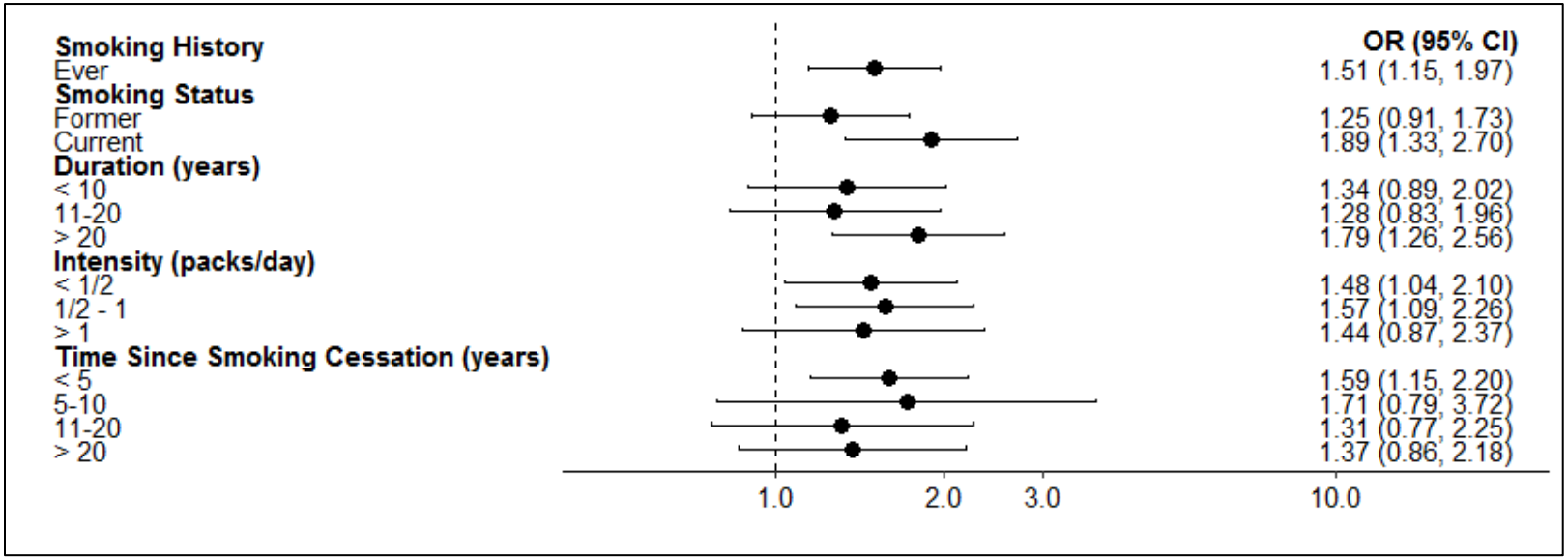
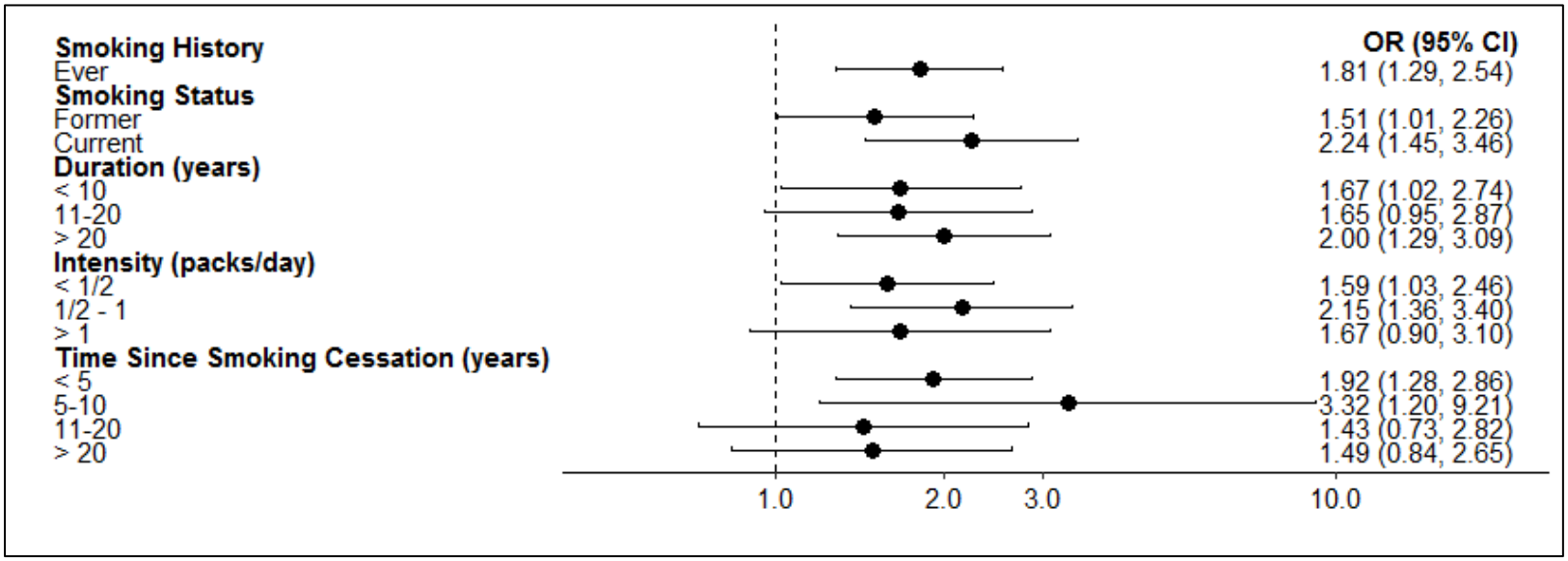


Figure 4.2. Categorical smoking metrics and association with ER, ESR1, and LS breast cancer subtypes

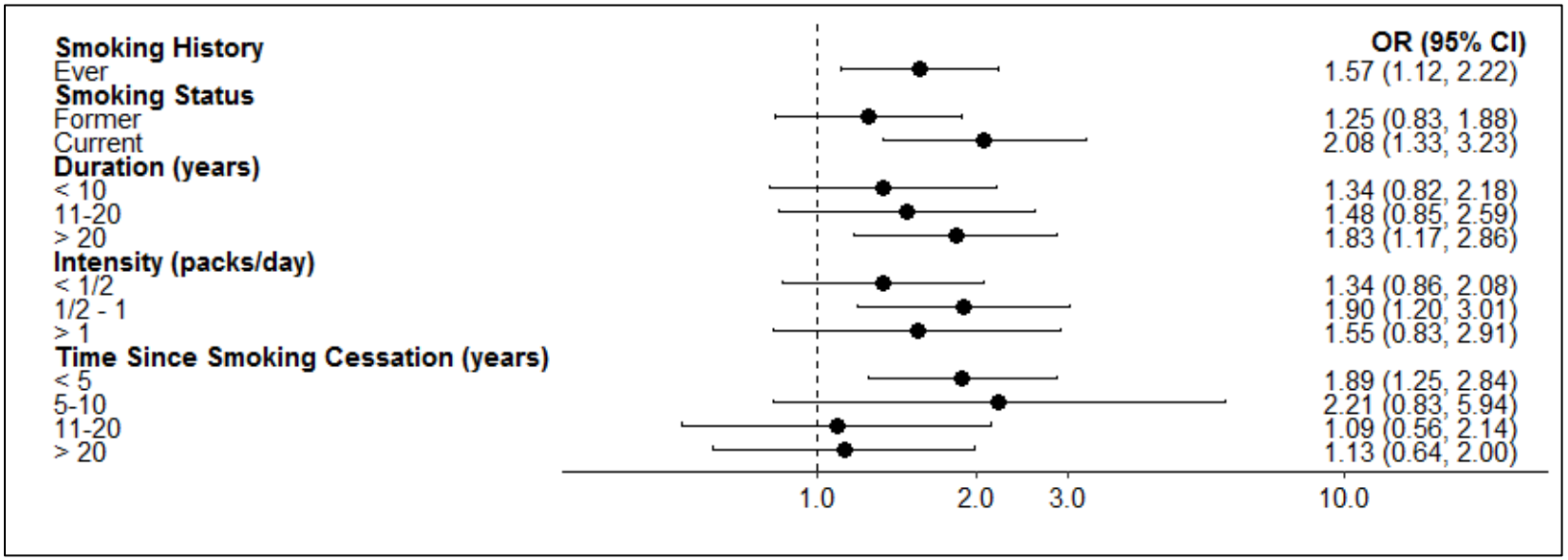
Note. (A) The association between smoking and **ER+ breast cancer.** Contrast compares ER+ cases to ER- cases with never smokers as referent group; (B) The association between smoking and *ESR1*+ breast cancer. Contrast compares *ESR1*+ cases to *ESR1*- cases with never smokers as referent group; (C) The association between smoking and LS+ breast cancer. Contrast compares LS+ cases to LS- cases with never smokers as referent group. Odds ratios and 95% confidence intervals were derived from unconditional logistic regression models, adjusted for: Nanostring batch, age, race, menopausal status, parity, breastfeeding, family history of breast cancer, alcohol use, body mass index (kg/m^2), physical activity, oral contraceptive use, hormone replacement therapy use, node status, stage, tumor size, and tumor grade.

B.



Note. (A) The association between smoking and ER+ breast cancer. Contrast compares ER+ cases to ER- cases with never smokers as referent group; (B) The association between smoking and ***ESR1+* breast cancer.** Contrast compares *ESR1*+ cases to *ESR1*- cases with never smokers as referent group; (C) The association between smoking and LS+ breast cancer. Contrast compares LS+ cases to LS- cases with never smokers as referent group. Odds ratios and 95% confidence intervals were derived from unconditional logistic regression models, adjusted for: Nanostring batch, age, race, menopausal status, parity, breastfeeding, family history of breast cancer, alcohol use, body mass index (kg/m^2), physical activity, oral contraceptive use, hormone replacement therapy use, node status, stage, tumor size, and tumor grade.

C.



Note. (A) The association between smoking and ER+ breast cancer. Contrast compares ER+ cases to ER- cases with never smokers as referent group; (B) The association between smoking and *ESR1*+ breast cancer. Contrast compares *ESR1*+ cases to *ESR1*- cases with never smokers as referent group; (C) The association between smoking and **LS+ breast cancer.** Contrast compares LS+ cases to LS- cases with never smokers as referent group. Odds ratios and 95% confidence intervals were derived from unconditional logistic regression models, adjusted for: Nanostring batch, age, race, menopausal status, parity, breastfeeding, family history of breast cancer, alcohol use, body mass index (kg/m^2), physical activity, oral contraceptive use, hormone replacement therapy use, node status, stage, tumor size, and tumor grade.

Table 4.1. Estimated odds ratios and 95% confidence intervals for cumulative smoking exposure and ER-defined breast cancer subtypes.

	ER OR per pack-decade (95% CI)	**ESR1** OR per pack-decade (95% CI)	**LS** OR per pack-decade (95% CI)
Cumulative Exposure	1.09 (0.99, 1.20)	1.18 (1.04, 1.34)	1.18 (1.04, 1.35)
Time Since Exposure			
0 - 10 years	1.54 (0.74, 3.19)	2.15 (0.82, 5.64)	2.99 (1.11, 8.08)
10 - 20 years	0.83 (0.37, 1.89)	0.87 (0.30, 2.55)	0.79 (0.26, 2.35)
20+ years	1.08 (0.88, 1.32)	1.07 (0.83, 1.39)	1.01 (0.78, 1.31)
Test of heterogeneity			
LRT, 2 df[a]	0.94	2.59	6.39
p-value	0.63	0.27	0.04

Abbreviations: ER-Estrogen-receptor. IHC-Immunohistochemistry. LRT- Likelihood Ratio Test. LS-Luminal Score. PS-Proliferation Score.

Note. Smoking exposure was modeled as the number of packs smoked per decade. Odds ratios and 95% confidence intervals were derived from unconditional logistic regression models, adjusted for: Nanostring code set, age, race, menopausal status, parity, breastfeeding, family history of breast cancer, alcohol use, body mass index (kg/m^2), physical activity, oral contraceptive use, hormone replacement therapy use, node status, stage, tumor size, and tumor grade.

a - LRT comparing cumulative and exposure-time-windows model, with 2 degrees of freedom.

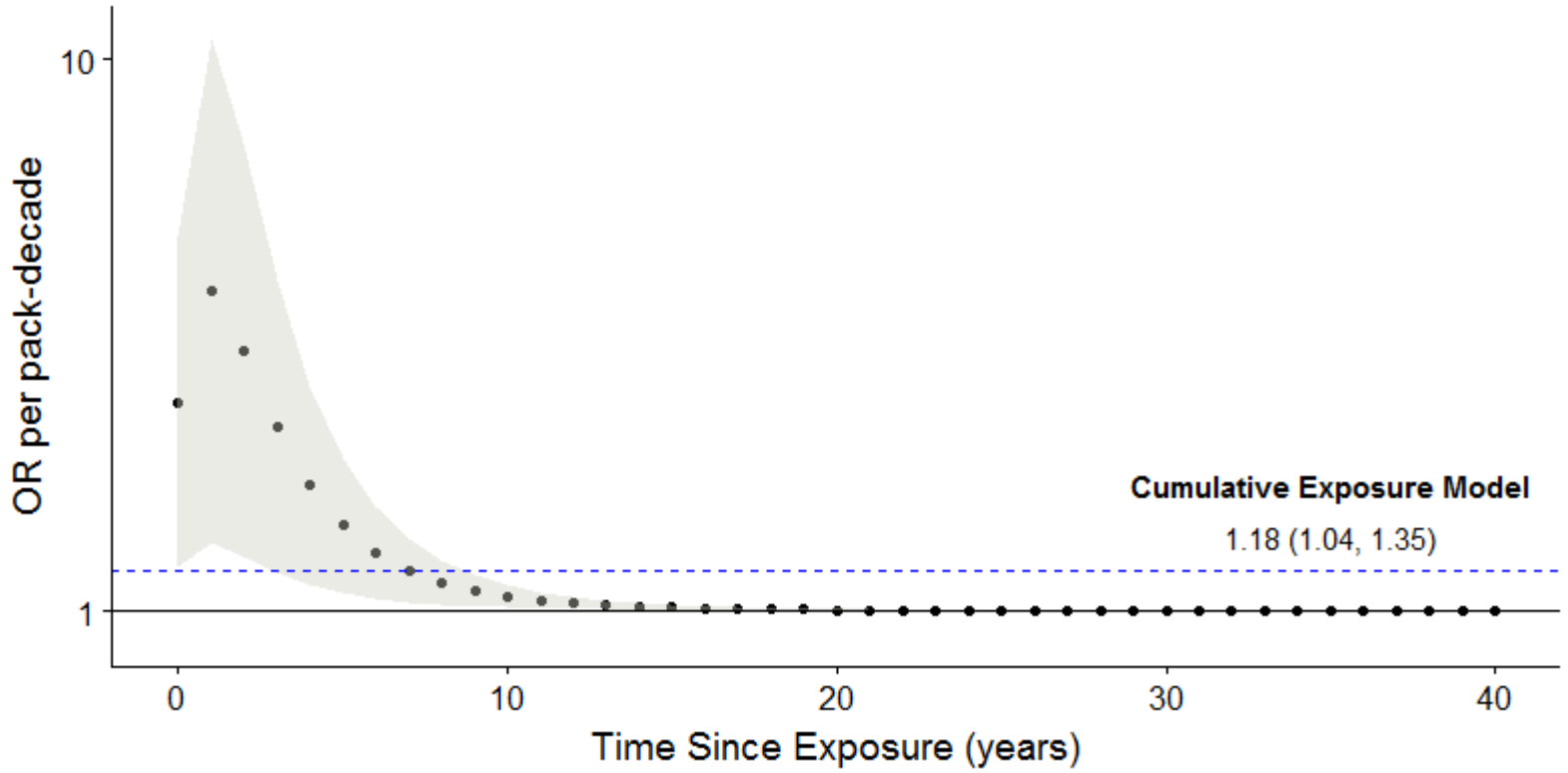


Figure 4.3. Temporal associations between pack-decades of cigarettes smoked and luminal score positive (LS+) breast cancer.

Note. Logistic regression models were adjusted for age and race. The dashed blue line indicates the estimated odds ratio for cumulative smoking exposure (pack-decades) for the model described in Table 2 (OR and 95% CI = 1.18 (1.04, 1.35)). The solid dark gray dots indicate point estimates for the association between pack-decades and LS+ breast cancer for each year preceding breast cancer diagnosis over a period of 40 years, with exposure time points weighted using a lognormal distribution. The light gray bands represent 95% confidence intervals surrounding point estimates.

Table 4.2. Estimated biomarker expression values for the effect of categorical smoking measures.

	ER (WT%)		ESR1 mRNA (Log2)		Luminal Score (Log2)	
Metric	Mean ± SD	p	Mean ± SD	p	Mean ± SD	p
Smoking						
Never	56.6 ± 10.2	REF	9.4 ± 0.7	REF	8.5 ± 0.4	REF
Ever	57.5 ± 9.6	0.05	9.7 ± 0.7	0.02	8.6 ± 0.4	0.04
Smoking status						
Never	56.6 ± 10.2	REF	9.4 ± 0.7	REF	8.5 ± 0.4	REF
Former	55.2 ± 9.4	0.11	9.4 ± 0.7	0.15	8.5 ± 0.4	0.38
Current	60.8 ± 8.9	0.14	10.0 ± 0.6	0.02	8.8 ± 0.4	0.01
Duration (years)						
Never	56.6 ± 10.3	REF	9.4 ± 0.7	REF	8.5 ± 0.4	REF
≤10 years	59.0 ± 9.5	0.22	9.7 ± 0.7	0.32	8.6 ± 0.4	0.24
11-20 years	60.0 ± 9.0	0.13	10.0 ± 0.6	0.05	8.7 ± 0.4	0.21
> 20 years	55.5 ± 9.8	0.20	9.6 ± 0.7	0.05	8.6 ± 0.4	0.07
Dose (packs per day)						
Never	56.6 ± 10.2	REF	9.4 ± 0.7	REF	8.5 ± 0.4	REF
<1/2	58.5 ± 9.7	0.35	9.8 ± 0.6	0.13	8.7 ± 0.4	0.14
1/2 – 1	53.0 ± 9.7	0.03	9.6 ± 0.7	0.08	8.5 ± 0.4	0.18
>1	56.6 ± 10.2	0.47	9.5 ± 0.6	0.08	8.5 ± 0.3	0.08
Time since smoking (years)						
Never	56.5 ± 10.3	REF	9.4 ± 0.7	REF	8.5 ± 0.7	REF
< 5 years	60.2 ± 9.2	0.13	9.9 ± 0.7	0.01	8.9 ± 0.7	0.01
5-10 years	60.4 ± 10.5	0.14	10.2 ± 0.8	0.09	8.8 ± 0.8	0.22
11-20 years	55.5 ± 9.3	0.41	9.4 ± 0.7	0.51	8.3 ± 0.7	0.98
> 20 years	51.9 ± 7.9	0.32	9.0 ± 0.6	0.74	8.1 ± 0.6	0.84

Note. Among all cases. Estimated means and standard deviations were derived from linear regression models, adjusted for age, race, and Nanostring code set (where applicable). P-values are presented for the respective linear regression parameter estimates where 'Never' smokers serve as referent group.

Table 4.3. Estimated biomarker expression values for the effect of categorical smoking measures.

	ER (WT%)		ESR1 mRNA (Log2)		Luminal Score (Log2)	
Metric	Mean ± SD	p	Mean ± SD	p	Mean ± SD	p
Smoking						
Never	67.3 ± 5.7	REF	10.1 ± 0.5	REF	8.9 ± 0.2	REF
Ever	64.5 ± 5.4	0.16	10.3 ± 0.5	0.03	9.0 ± 0.2	0.27
Smoking status						
Never	68.5 ± 5.7	REF	10.1 ± 0.5	REF	8.9 ± 0.2	REF
Former	65.2 ± 5.4	0.50	10.1 ± 0.5	0.27	8.9 ± 0.2	0.73
Current	66.2 ± 4.6	0.09	10.5 ± 0.4	0.01	9.1 ± 0.2	0.13
Duration (years)						
Never	68.6 ± 5.9	REF	10.1 ± 0.5	REF	8.9 ± 0.2	REF
≤10 years	67.1 ± 5.5	0.38	10.3 ± 0.5	0.40	9.0 ± 0.2	0.82
11-20 years	68.1 ± 5.4	0.87	10.5 ± 0.5	0.12	9.0 ± 0.2	0.44
> 20 years	63.6 ± 5.1	0.10	10.2 ± 0.5	0.05	9.0 ± 0.2	0.23
Dose (packs per day)						
Never	68.5 ± 5.7	REF	10.1 ± 0.5	REF	8.9 ± 0.2	REF
<1/2	65.6 ± 5.2	0.09	10.3 ± 0.5	0.16	8.9 ± 0.2	0.77
1/2 – 1	66.0 ± 5.4	0.59	10.1 ± 0.5	0.36	8.9 ± 0.2	0.67
>1	64.8 ± 5.1	0.49	10.5 ± 0.4	0.01	9.1 ± 0.2	0.04
Time since smoking (years)						
Never	68.4 ± 5.7	REF	10.1 ± 0.5	REF	9.0 ± 0.2	REF
< 5 years	66.4 ± 4.9	0.33	10.5 ± 0.5	0.01	9.1 ± 0.2	0.04
5-10 years	69.5 ± 6.0	0.96	10.4 ± 0.6	0.72	8.9 ± 0.3	0.92
11-20 years	66.5 ± 5.5	0.70	10.2 ± 0.5	0.53	8.9 ± 0.2	0.96
> 20 years	61.8 ± 4.1	0.09	9.8 ± 0.4	0.79	8.8 ± 0.2	0.59

Note. Restricted to ER+ breast cancer cases. Estimated means and standard deviations were derived from linear regression models, adjusted for age, race, and Nanostring code set (where applicable). P-values are presented for the respective linear regression parameter estimates where 'Never' smokers serve as referent group.

A. Among All Cases

B. Among ER+ Cases

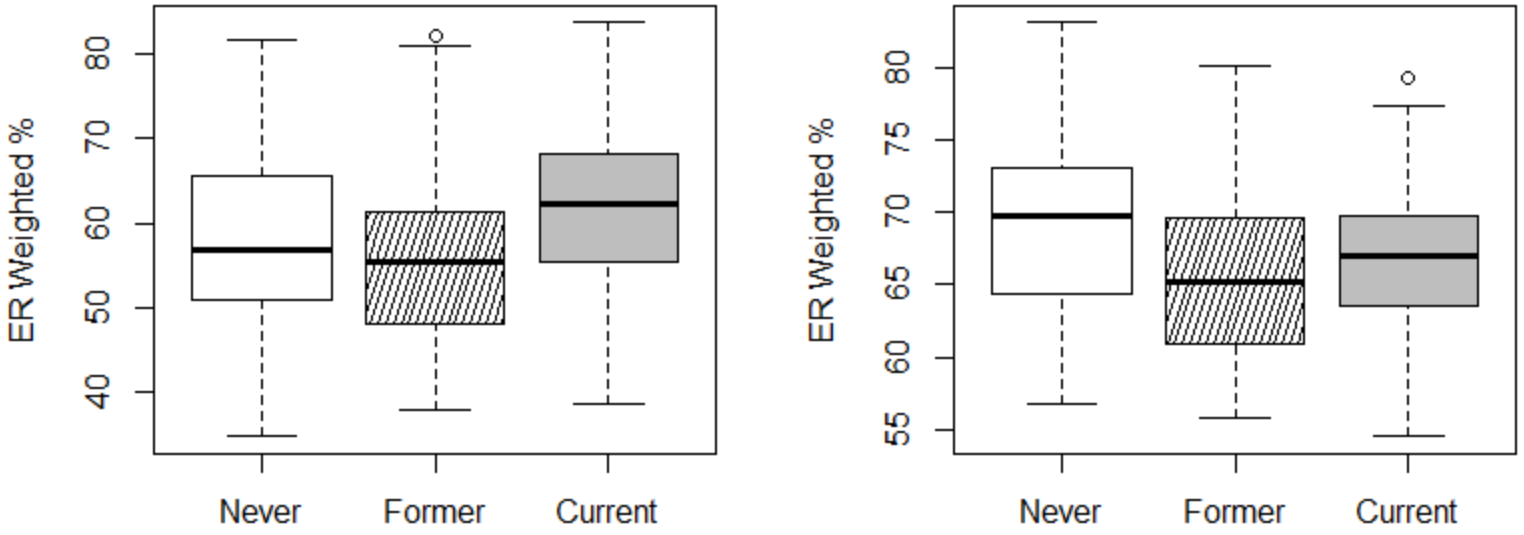


Figure 4.4. Distribution of ER protein by never, former, or current smoking status.

Note. Boxplots displaying the distribution of weighted percent ER (WT%) overall (A) and among ER+ breast tumors (B). ER WT% values were estimated from a linear regression model adjusted for age and race.

A. Among All Cases

B. Among ER+ Cases

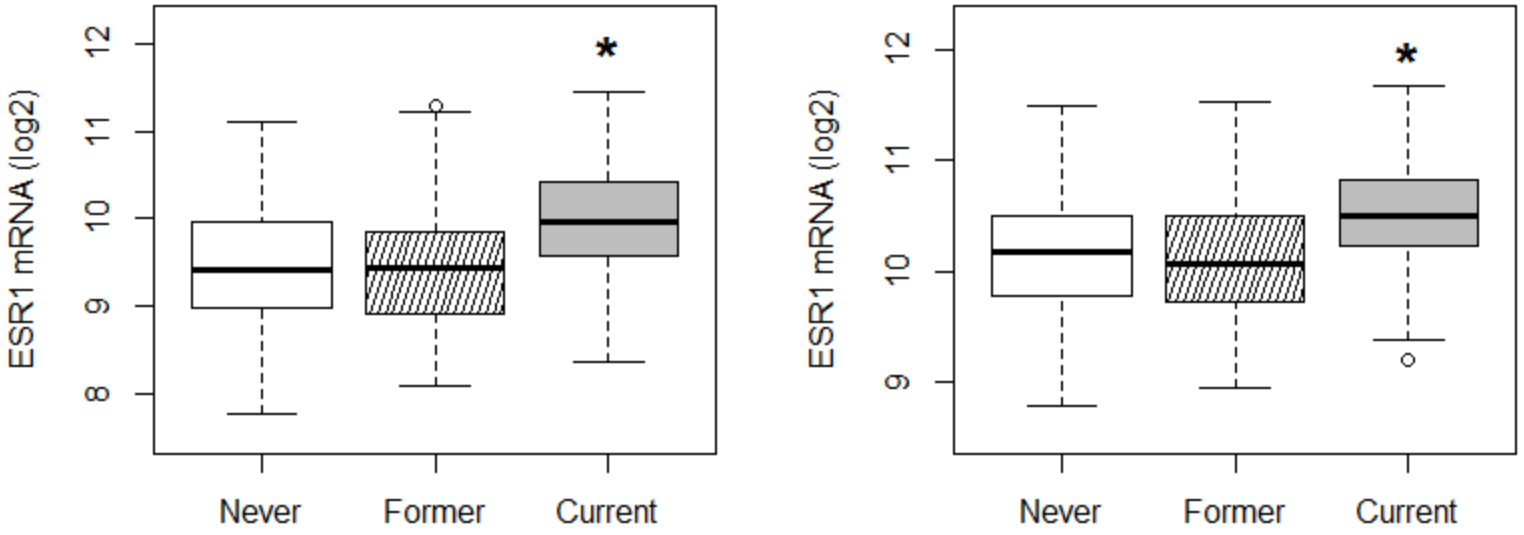


Figure 4.5. Distribution of ESR1 mRNA by never, former, or current smoking status.

Note. Boxplots displaying the distribution of ESR1 overall (A) and among ER+ breast tumors (B). ESR1 values were estimated from a linear regression model adjusted for age, race, and Nanostring code set. ***** = $p < 0.05$, where 'Never' smokers serve as the referent group.

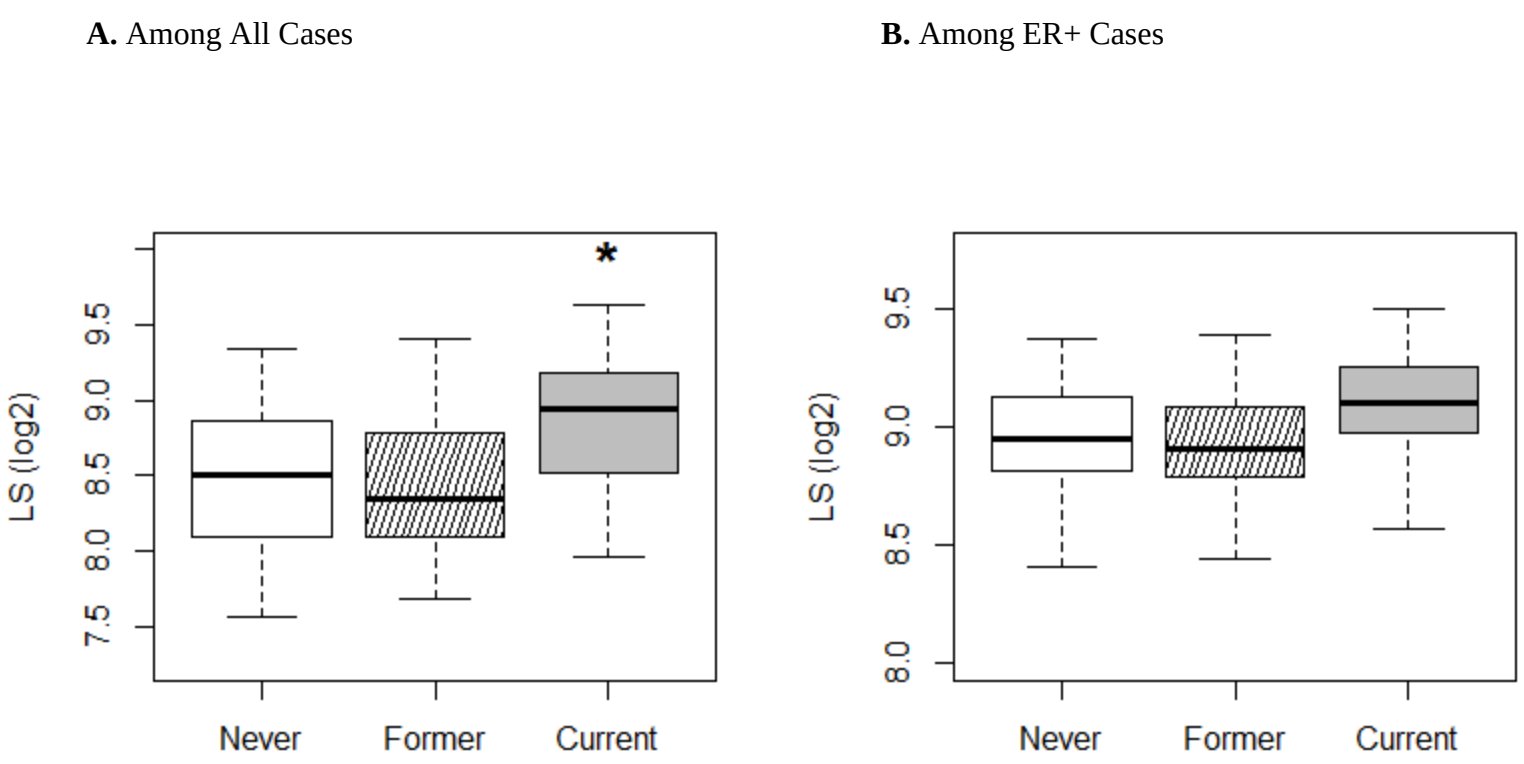


Figure 4.6. Distribution of the luminal score by never, former, or current smoking status.

Note. Boxplots displaying the distribution of the luminal score (LS) overall (A) and among ER+ breast tumors (B). LS values were estimated from a linear regression model adjusted for age, race, and Nanostring code set. ***** = $p < 0.05$, where 'Never' smokers serve as the referent group.

CHAPTER 5: SMOKING, P53, EGFR, AND RELATED BIOMARKERS

5.1 Background

Although growing evidence suggests smoking as a possible risk factor for breast cancer, the mechanistic events leading to disease have not been clearly defined[71]. P53 – the most frequently mutated gene in breast cancer[80] – functions as a tumor suppressor and its mutation or overexpression in breast tissue may serve as an etiologic marker for DNA damage from smoking. Indeed, results from the Carolina Breast Cancer Study demonstrated that breast cancer patients who were self-reported smokers at time of diagnosis had a higher prevalence of specific p53 mutations, when compared with their non-smoking counterparts (Conway et al.)[57]. In addition to gene sequencing technologies, researchers have also employed immunohistochemical (IHC) staining to detect p53 protein expression in breast tumors. However, the few studies that have examined smoking in relation to p53 IHC subtypes have yielded mixed results[58,59]. Elevated nuclear staining of p53 positively correlates to the presence of missense mutations, which stabilize the p53 protein, and may not adequately capture other genetic changes that could influence protein expression (e.g., deletion, frameshift)[81]. Thus, single and multigene RNA measures that reflect biologic activity of p53 function in breast tissues may provide improved resolution in studies of smoking and breast cancer etiology[65].

While the DNA damaging effects of cigarette smoke are hypothesized to elicit disease[82,83], it is plausible that other chemicals unrelated to mechanisms of DNA damage contribute to breast cancer initiation or progression. Nicotine, though not considered to be a

carcinogen, has been implicated in the development of breast cancer in animal and tissue culture studies[27,29,84,85]. Mouse xenograft models have demonstrated the transformation of normal mammary cells to neoplastic cells upon exposure to nicotine derivatives[27]. And similar to findings from lung cancer studies[28], breast epithelial cells treated with nicotine show increased rates of cell proliferation through activation of EGFR[29], providing a measurable clinical endpoint for the hypothesis that nicotine may be associated breast cancer phenotypes (i.e., EGFR+ breast cancer). Further, if smoking is associated with differential proliferative activity in breast tumors, a measure of cell proliferation may also serve as a suitable outcome to examine potential associations with smoking exposure. Thus, examining smoking exposure in relation to EGFR expression and measures of cell proliferation may highlight mechanisms for these putative markers of smoking exposure in breast tissue. Finally, several studies have linked smoking to the ER+ breast cancer subtype, which represents a distinct etiologic class of disease characterized by overexpression of the estrogen-receptor[7]. Thus, a proper evaluation of proposed markers should be examined in light of established ER+ and ER- etiologic types.

In this study, we sought to evaluate smoking and its association with binary and continuous measures for p53 and EGFR gene expression, overall and stratified on ER+ or ER- breast cancer subtypes. We incorporated immunohistochemical data and RNA-based multigene signatures as measures of p53 and cell proliferation signaling patterns in tumors.

5.2 Methods

5.2.1 Study Population

Phase III of the Carolina Breast Cancer Study (CBCS III) is a population-based case-only study that combines epidemiology and molecular biology to examine environmental and

genetic risk factors for molecular subtypes of breast cancer. To be eligible for inclusion, patients must have been female and received a first and primary diagnosis of breast cancer between May 1, 2008 and October 31, 2013. The patient also must have resided in the 44-county study region and been between the ages of 20 and 74 at the time of diagnosis. To examine potential risk factor differences by age and race, the CBCS employed a randomized recruitment strategy that was designed to oversample young and African American women[73].

Breast cancer cases were identified by a rapid case ascertainment system, implemented through collaboration between Lineberger Comprehensive Cancer Center (LCCC) and the North Carolina Central Cancer Registry (NCCCR). Briefly, CBCS contacted the patient's primary physician to obtain permission to invite the patient into the study, yielding an overall response rate of 70% and a total of 2,998 women. Study participants were asked to consent to a nurse-administered in-person interview that took place in the study participant's home or another pre-arranged location. The average time between study enrollment and interview was 6 months. The nurse administered questionnaire included items on family and personal medical history, reproductive history, smoking, alcohol, diet, medication use and occupational history. Upon consent, the nurse also collected a blood sample and objective anthropometric measurements of height (m), weight (kg, waist (m), and hip (m) circumference.

5.2.2 Outcome Assessment

The CBCS includes protein and RNA expression data on genes involved in estrogen-signaling. At the time of interview, investigators asked study participants for permission to obtain formalin-fixed, paraffin-embedded (FPPE) tumor blocks or tissue slides from the hospital where the diagnostic surgery was to be performed. Tumor blocks were used to

construct tissue microarrays (TMAs) for IHC staining, where each patient's tumor was represented by 1 to 8 cores on the microarray. RNA was extracted from the same tumor specimens used to construct the TMAs, using the Qiagen RNeasy FFPE kit and protocol. CBCS includes data for 1970 women included in the IHC analysis and 993 women included in both the IHC and RNA analyses.

ER, p53, and EGFR Protein Quantification. Automated quantification of ER, p53, and EGFR protein was determined by a Genie classifier and Aperio algorithms specified for nuclear or membrane proteins (Aperio Technologies, Vista, CA)[61]. We calculated percent positivity for each marker as the product of positively stained tumor cells for each core, multiplied by its core-specific weight, summed across all cores per patient. We assigned a binary cut point of ≥ 10% for ER positive, p53 positive, and EGFR positive tumors. Specifically, and with respect to protein expression cut points, ER positive and ER negative tumors are referred to as ER+ and ER-; p53 positive and p53 negative tumors are referred to as p53 IHC+ and p53 IHC- and, finally, EGFR positive and EGFR negative tumors are referred to as EGFR+ and EGFR-.

MDM2 and EGFR mRNA Quantification. MDM2 (a negative regulator of p53) and EGFR were quantified using Nanostring technology. Briefly, total mRNA counts were assayed using target-specific molecular probes, which hybridize to RNA fragments in solution. Hybrids were then counted using microscopic imaging, yielding raw mRNA counts. Quality control and data normalization were performed using the NanostringNorm R package[74]. Data were first normalized to the geometric means of 6 internal positive controls and subsequently to the geometric means of 5 reference genes. Normalized mRNA counts were log2 transformed. MDM2 and EGFR were assayed along with other genes included in 1

of 3 Nanostring batches or code sets. Thus, all Nanostring analyses were adjusted for 'code set' to minimize potential batch effects. We assessed potential binary cut points for mRNA measures of MDM2 and EGFR using an expectation-maximization algorithm that is used to classify mixed Gaussian distributions[75]. However, results from this analysis suggested that the distribution of values for each biomarker followed a single normal distribution and were not suitable for dichotomization (Figure 6.5 and Figure 6.6). Thus, we considered only continuous measures for MDM2 and EGFR mRNA.

P53 mRNA Signature. We used the 48-gene signature identified by Troester et al. to classify breast tumors as p53 wildtype or p53 mutant[65]. The 48-gene predictor was applied to each case and classified based upon highest Pearson correlation with a centroid defined for either subtype. The mutant or wildtype designations characterize downstream biologic activity following p53 loss or activation, respectively. The p53 signature captures changes in p53-dependent gene expression due to both missense and non-missense mutations and other defects in regulators of p53. For all logistic regression analyses that examine associations between smoking and binary classifications for the p53 mRNA signature, p53 mutant tumors are referred to as 'p53 Mut' and p53 wildtype tumors are referred to as 'p53 Wt'. mRNA transcripts used to define the p53 signature were quantified using Nanostring technology.

Proliferation mRNA Signature. The PAM50 proliferation signature includes 11 highly correlated genes associated with cell proliferation in breast tumors[64,72]. The 11 genes include: BIRC5, CCNB1, CDC20, NUF2, CEP55, NDC80, MKI67, PTTG1, RRM2, TYMS, AND UBE2C. A description of the proliferation signature is described in Nielsen (2010)[72]. Briefly, each tumor is assigned a continuous score based on an algorithm that incorporates the average of expression levels for the 11 genes in the proliferation signature and linear

regression coefficients for clinical variable including breast cancer intrinsic subtype and tumor size. Tumors are classified as having high, medium, or low proliferation. For all logistic regression analyses that examine associations between smoking and binary classifications for the proliferation score (PS), tumors classified as high are referred to as 'PS positive or PS+'; and tumors that are classified as medium or low are referred to as 'PS negative or PS-'. Single mRNA transcripts included in the proliferation signature were quantified using Nanostring technology.

5.2.3 Exposure Assessment

History of smoking exposure was obtained during the nurse-administered in-person interview and includes data on smoking duration, frequency, and dose. Self-reported smoking is considered a valid measure of smoking exposure, with increased accuracy obtained during in-person interview formats[66]. Women in CBCS were considered ever smokers if they smoked at least 100 cigarettes during their lifetimes. CBCS investigators collected data on smoking history defined as 'ever' or 'never' (history); smoking status defined as 'current', 'former', or 'never' (status); age at smoking initiation measured in years (initiation); smoking duration measured as the total number of years of smoking between initiation and current use or cessation (duration); number of cigarettes smoked per day (dose); and age at smoking cessation, where applicable.

5.2.4 Data Analysis

For each binary breast cancer classification, we used generalized logit models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for categorical measures of smoking and the p53 and EGFR-defined subtypes. We used linear regression to model the relationship between continuous measures of single or multigene markers and categorical

measures of smoking, adjusted for age, race, and Nanostring code set (where applicable). We calculated the estimated value of continuous biomarker expression for each individual, based on coding of the smoking exposure and covariate pattern, adjusted for age, race, and Nanostring code set. Expression levels for each biomarker were described according to interquartile range and visualized using box plots within categories of smoking.

All analyses were conducted using SAS 9.4 (SAS Institute Inc, Cary, NC) and R version 3.3.3.

5.3 Results

Our analysis includes data on 1,970 women diagnosed with invasive breast cancer. Each patient had immunohistochemical data for tumor protein expression of ER, p53, and EGFR. In a subset of our study population (n=993) we measured mRNA levels of single gene (i.e., MDM2 and EGFR) and multigene markers (i.e., p53 signature and proliferation signature) that reflect signaling pathways of interest. Baseline patient characteristics were not substantially different between the two analytic groups (Table 5.1).

5.3.1 Quantitative P53 Protein and MDM2 mRNA Expression

Average p53 protein expression levels were slightly higher among p53 mutant (Mut) tumors (mean (SD), 7.3% (10.9)) compared to p53 wildtype (Wt) tumors [mean (SD), 3.9% (6.2)], where "Mut" and "Wt" signify tumors classified per the 48-gene signature by Troester et al[65] (Figure 5.1). p53 wildtype tumors generally had low p53 protein expression levels, with 80% having values below ten percent (i.e., proportion of tumor cells that stained positive for the p53 antigen). We also assessed expression levels for MDM2 – a negative regulator of p53[86]. P53 protein and MDM2 mRNA expression were not strongly correlated, but had a slight inverse association ($r=-0.18$, $p < 0.0001$). The density graph for MDM2 mRNA

expression values stratified by p53 mutant and wildtype mRNA binary classifications had substantial overlap, thereby suggesting the existence of a unimodal distribution (Figure 5.2).

5.3.2 Smoking and P53 Protein Binary Subtypes

Considering all patients regardless of ER status, we observed no apparent association between smoking and p53 IHC mutant and wildtype tumors (Table 5.2). However, when stratified by ER status, women with a history of smoking who were diagnosed with ER- breast tumors were more likely to also be classified as p53 IHC+ (i.e., ER-/p53 IHC+). Among ER- cases, current and former smokers had increased odds of p53 IHC+ tumors [Current: OR = 1.34 95% CI (0.77, 2.34); Former: OR = 1.49 95% CI (0.91, 2.44)]. In addition, women who smoked 20 or more years had greater odds of ER-/p53 IHC+ tumors when compared to their ER- counterparts with no history of smoking (OR = 1.86 95% CI (1.06, 3.27)).

5.3.3 Smoking and P53 mRNA Binary Subtypes

Although we observed evidence of associations between smoking and ER/p53 IHC defined subtypes (i.e., ER-/p53 IHC+), breast tumors classified as Wildtype (Wt) or Mutant (Mut) per the p53 RNA signature were not differentially linked to smoking exposure overall or within ER subtype (Table 5.3). Notably, when stratified by ER status, our RNA-based analyses were restricted to smaller sample sizes, thereby yielding less precise estimates.

5.3.4 Smoking, Quantitative p53 Protein and MDM2 mRNA Expression

We assessed quantitative levels of p53 protein and MDM2 mRNA expression in relation to ER subtype and smoking status at time of diagnosis using linear regression models (Figure 5.3). ER- breast cancer cases who were former smokers at time of diagnosis had higher p53 protein expression when compared to their never smoker counterparts [mean (%)

= 23.1 vs 18.1, $p < 0.05$]. However, we did not observe a complementary trend for MDM2 mRNA expression when comparing former and never smokers [mean (log2) = 9.3 vs 9.3, p =0.81]. Similarly, considering duration >20 years, we observed statistically significant higher p53 protein levels but no trend for MDM2 mRNA expression (data not shown).

5.3.5 Smoking and EGFR Protein Binary Subtypes

In general, we did not observe a consistent association between smoking and binary subtypes for EGFR protein overall or when stratified by ER status (Table 3). However, there was a slight suggestion of an association between smoking dose and EGFR positivity. The highest category for smoking dose was associated with decreased odds of the EGFR+ subtype among women with ER+ tumors (OR = 0.67 95% CI (0.46, 0.97)) and increased odds of the EGFR+ subtype among ER- tumors (OR = 1.79 95% CI (0.99, 1.04)).

5.3.6 Smoking and Quantitative EGFR protein and EGFR mRNA

In addition, we did not observe an association between smoking and quantitative EGFR protein expression overall or within ER-defined subtypes (Figure 5.7). However, quantitative levels of EGFR mRNA were lowest among ER- cases who were former smokers compared to never smokers [mean (%) = 8.0 vs 8.4, $p < 0.05$]. (Figure 5.8).

5.3.7 Smoking and Proliferation Binary Subtypes

Smoking was consistently linked to decreased odds of having a PS+ breast tumor irrespective of ER status (Table 5.5). Specifically, measures for current smoking (OR = 0.67 95% CI (0.47, 0.94)), smoking duration > 20 years (OR = 0.70 95% CI (0.49, 0.99)), and smoking cessation within 5 to 10 years of diagnosis (OR = 0.51 95% CI (0.28, 0.94)) had substantially lower odds of harboring the PS+ phenotype. In general, we observed similar

trends for analyses stratified by ER subtype. However, the ER-specific analyses yielded less precise estimates and were not statistically significant.

5.4 Discussion

In this study, we examined multigene signatures for p53, proliferation, and single-gene EGFR as potential biomarkers of smoking exposure that serve as surrogate markers for DNA damage or growth factor dependent mechanisms. For p53 and EGFR, we observed null associations for smoking exposure when considering all cases. However, when stratifying on ER status, we observed temporal and dose-dependent associations between smoking and p53+ and EGFR+ molecular phenotypes among ER- cases, particularly for smoking duration > 20 years and those who smoked more than 1 pack of cigarettes per day. In addition, binary classifications for our multigene proliferation score was differentially linked to smoking exposure where current smoking, long smoking duration, and smoking cessation within specific time-windows were linked to decreased odds of the PS+ phenotype.

Cigarette smoke contains more than 70 carcinogens that have been evaluated by the International Agency for Research on Cancer (IARC), and which comprise eight chemical classes [20]. Two of the largest classes – the polycyclic aromatic hydrocarbons (PAHs) and the *N*-nitrosamines – are thought to be responsible for cancer initiation in lung tumors through development of DNA-adducts.[21]. Historically, investigators have proposed that PAHs store in breast adipose tissue to form PAH-DNA adducts; the resultant adducts are believed to lead to somatic mutations and subsequent tumor initiation[82,87,88]. P53 mutation has been evaluated as a probable marker of DNA damage in breast tissue following cigarette smoke exposure in several studies; however, results have been inconclusive. Conway et al. demonstrated that current smokers at the time of diagnosis had a higher prevalence of any p53 mutation in

exons 4-8, the DNA-binding region of the gene, when compared to former and never smokers[57]. However, in a similar analysis from the Long Island Breast Cancer Study Project (LIBCSP), Mordukhovich et al. reported null associations for current smoking and p53 gene mutation in the same exonic region[60].

In the present study, we used immunohistochemical staining to examine smoking in relation to p53 overexpression in breast tumors. Consistent with earlier reports from phases I and II of Carolina Breast Cancer Study, we found null associations between smoking and p53-expression (i.e., dark nuclear staining in 10% or greater of examined tumor cells)[58]. However, we note that the same analytic subset demonstrated associations between smoking and p53 genetic mutations in a separate analysis[57]. Furthermore, in a case-case analysis from the LIBCSP, Gammon et al. reported that smokers were twice as likely to be diagnosed with p53 IHC+ breast cancer when compared to never smokers, where p53 IHC+ was defined as moderate to strong staining in 10% or greater of tumor cells. The discordant results across studies may be explained by several factors including differences in assay protocol, selection of the p53 IHC cut point, and differences among the study populations.

In addition, neither CBCS nor LIBCSP evaluated smoking and p53 IHC+ cut points in relation to ER expression, whose dichotomization defines two distinct classes of breast tumors. In the present study, though we observed no association between smoking and p53-IHC breast cancer subtypes in breast cancer overall, temporal and dose-dependent patterns were observed among ER- cases. Among ER- cases, smoking duration of 20 years or more was associated with increased odds of the p53+ phenotype and higher levels of p53 protein expression. Interestingly, we also observed higher odds of the p53+ phenotype and higher quantitative levels of p53 protein among former smokers, but not current smokers. Thus, the

observed associations may reflect distal events in a woman's smoking history such that smoking at diagnosis has limited influence on modulation of p53 expression at diagnosis. Further, observed association may be sensitive to the distribution of the ER+ and ER- tumors in a given study population.

We did not observe similar associations between smoking and p53-mRNA binary subtypes. While the RNA signature captures downstream biological activity following p53 loss, overexpression of p53 protein most commonly reflects missense mutations that lead to accumulation of the protein in cells. Thus, the effect of smoking in breast tissue may be specific to p53 point mutations as hypothesized by research that highlighted the increased presence of DNA adducts in breast tissues of smokers. For our present analysis and similar future analyses, p53 protein may serve as a preferred proxy for smoking-related DNA-damage.

We also examined EGFR as a potential marker of growth-factor mediated activity in the relationship between smoking and breast carcinogenesis. Until recently, nicotine - the most abundant and pharmacologically active component of cigarette smoke - has not been implicated in the development of breast cancer, though early studies demonstrated measurable levels of nicotine in the breast fluid of smokers[23,24]. However, *in vitro* studies of nicotine exposure to breast epithelium has been linked to increased rates of cell proliferation, migration, and overexpression of EGFR, providing insight to a possible biological mechanism for the increased risk of breast cancer among smokers; in vitro studies of breast epithelial cells exposed to nicotine have also demonstrated higher expression of EGFR compared with unexposed cells[29,84,85,89] – and notably, it is linked to a poor clinical marker of disease. Similar to results for p53, when considering EGFR alone we observed no

differentiation of smoking exposure between EGFR+ and EGFR- subtypes. However, when stratified by ER status, we observed inverse associations between high smoking dose and the EGFR+ phenotype among ER- cases.

Interestingly, smoking was linked to lower odds of the PS+ phenotype and, conversely, increased odds of the PS- phenotype. This observation suggests that smoking may be linked to lower proliferative activity in breast tumors. Though several *in vitro* studies have linked cigarette smoke – and nicotine exposure specifically - to increased cell proliferation in breast tissue, studies have also demonstrated associations between smoking and lower rates of cell proliferation in other cell types and is thought to impede cell regeneration and wound healing[90,91]. Further, we hypothesized that both EGFR and the cell proliferation signature could serve as biomarkers related to cell growth. Though we found no clear links between smoking and the EGFR subtypes, it may be reasonable to investigate EGFR as a mediator of cell proliferation in future analyses.

Smoking has been linked to breast tumors that arise via genotoxic and growth-factor dependent mechanisms. In our study, we modeled those mechanisms as overexpression of single gene measures for P53 and EGFR, and differential expression of multigene signatures for p53 and cell proliferation. In addition, we also considered the expression of each marker by ER status Though we found no association between smoking and p53 or EGFR-defined subtypes independent of ER status, results from our study suggest that distal smoking exposure may be linked to differential expression of each among ER- cases. In addition, we observed consistently lower associations between smoking and the PS+ subtype, thereby suggesting that smoking may be linked to decreased proliferative activity. Future studies

should seek to replicate these findings in larger study samples and consider smoking in relation to additional biomarkers that may be related to DNA-damage or cell proliferation.

5.5 Addendum

When developing the analytic plan for Aim 2, we identified the possibility of evaluating smoking exposure in relation to the multigene proliferation score, though it was not described in our initial methodology plan. This added analysis complements the hypothesis that smoking couple be linked to differential cell proliferation patterns in breast tumors.

Table 5.1. Age, race, and smoking characteristics of CBCS III study participants.

	Overall		**Nanostring Sampled**	
Characteristics	n	%	n	%
Race				
AA	969	49.19	488	49.1
Non-AA	1001	50.81	505	50.9
Age				
<50	1039	52.74	506	51
≥50	931	47.26	487	49
Smoking History				
Never	1084	55.05	523	52.7
Ever	885	44.95	469	47.3
Missing	1		1	
Smoking Status				
Current	360	18.28	209	21.1
Former	525	26.66	260	26.2
Never	1084	55.05	523	52.7
Missing	1		1	
Duration of smoking (active)				
Never	1084	55.08	523	52.7
≤10 years	250	12.7	143	14.4
11-20 years	208	10.57	98	9.88
> 20 years	426	21.65	228	23
Missing	2		1	
Amount smoked (active)				
Never	1084	55.05	523	52.7
< 1/2 pack	344	17.47	184	18.5
1/2-1 pack	353	17.93	187	18.9
> 1 pack	188	9.55	98	9.88
Missing	1		1	
Time Since Smoking				
Never	1084	55.05	523	52.7
< 5 years	451	22.91	260	26.2
5-10 years	67	3.4	31	3.13
11-20 years	147	7.47	67	6.75
> 20 years	220	11.17	111	11.2
Missing	1		1	

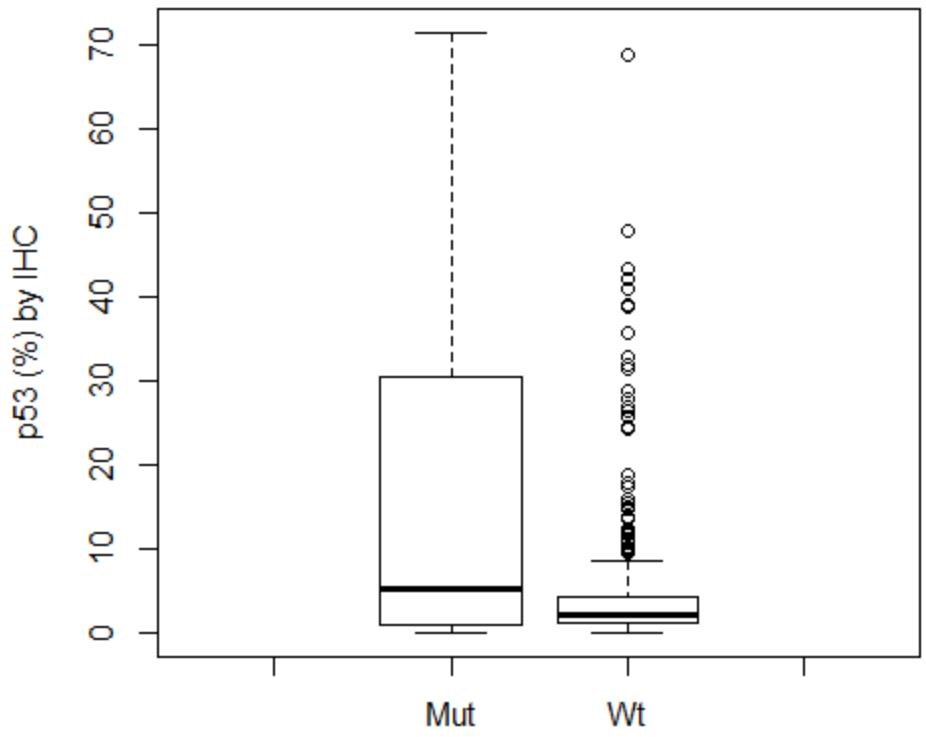


Figure 5.1. P53 protein expression (%), by p53 mRNA signature binary classification.

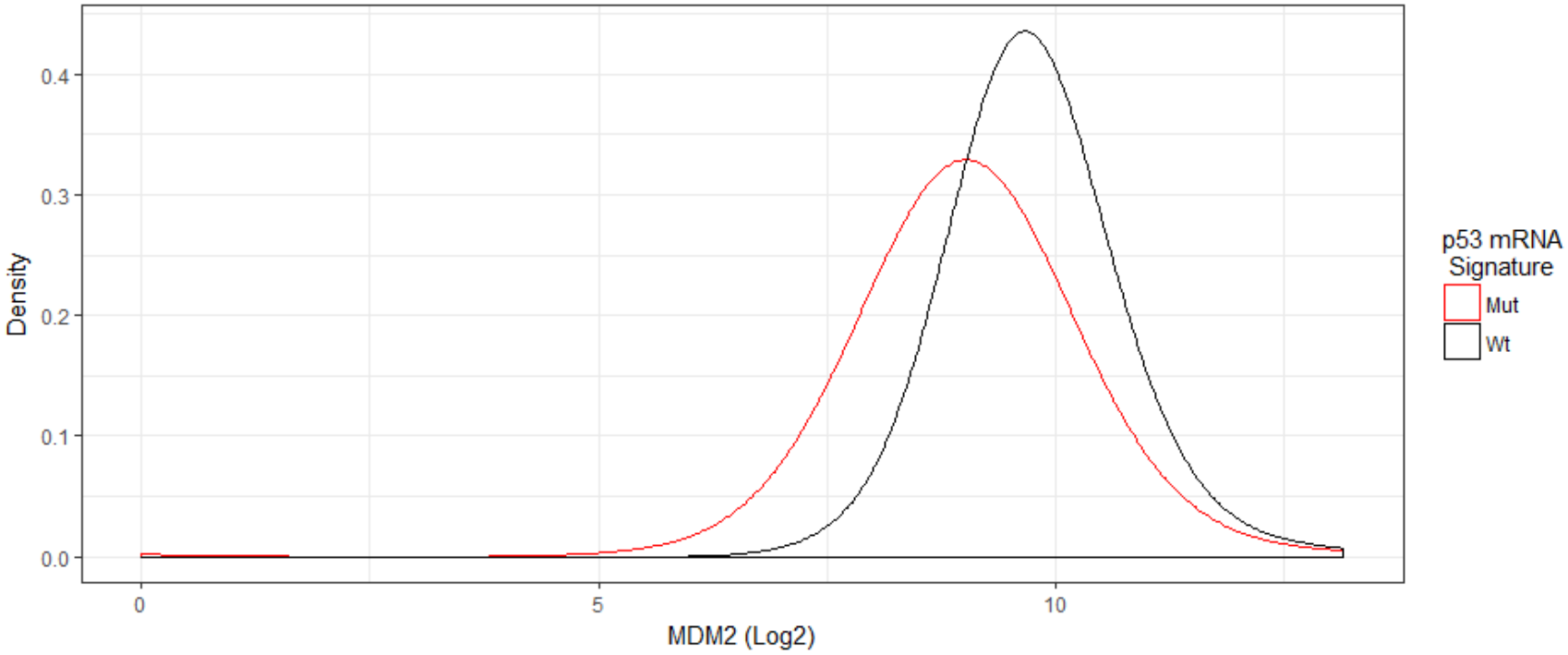


Figure 5.2. Density graphs for MDM2 mRNA expression (Log2), by p53 mRNA signature binary classification.

Table 5.2. Estimated odds ratios and 95% confidence intervals for p53 protein-defined breast cancer subtypes (adjusted for age and race).

	Overall	**ER+ Cases**	**ER- Cases**
	P53 IHC+ vs. P53 IHC-	P53 IHC+ vs. P53 IHC-	P53 IHC+ vs. P53 IHC-
Smoking history			
Ever	0.94 (0.72, 1.22)	0.84 (0.56, 1.28)	1.43 (0.95, 2.15)
Smoking status			
Former	0.93 (0.68, 1.27)	0.73 (0.44, 1.21)	1.49 (0.91, 2.44)
Current	0.95 (0.67, 1.34)	1.02 (0.60, 1.74)	1.34 (0.77, 2.34)
Duration (years)			
< 10	1.02 (0.69, 1.52)	0.87 (0.46, 1.65)	1.78 (0.91, 3.47)
11-20	0.64 (0.39, 1.00)	0.51 (0.23, 1.16)	0.75 (0.38, 1.48)
> 20	1.07 (0.77, 1.48)	1.00 (0.61, 1.65)	1.86 (1.06, 3.27)
Dose (packs/day)			
< 1/2	0.91 (0.64, 1.30)	0.65 (0.35, 1.20)	1.72 (0.98, 3.01)
1/2 – 1	1.17 (0.83, 1.65)	1.20 (0.72, 1.99)	1.87 (1.02, 3.42)
> 1	0.61 (0.36, 1.01)	0.61 (0.28, 1.33)	0.61 (0.29, 1.32)
Time since exposure (years)			
< 5	0.91 (0.66, 1.25)	0.92 (0.55, 1.52)	1.34 (0.80, 2.22)
5-10	0.59 (0.26, 1.35)	0.23 (0.03, 1.75)	0.83 (0.27, 2.49)
11-20	1.10 (0.67, 1.80)	0.65 (0.27, 1.58)	1.80 (0.81, 3.99)
> 20	1.02 (0.66, 1.58)	1.00 (0.53, 1.87)	1.81 (0.82, 4.00)

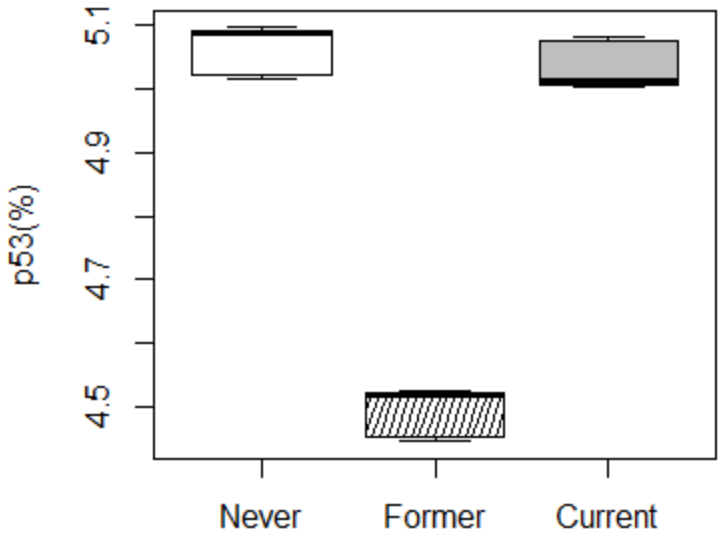
Note. ER+ tumors are defined as tumors with ≥ 10 percent of examined tumor cells that stained positive for ER antigen; ER- tumors are defined as tumors with < 10 percent staining of cells. Similarly, p53 IHC+ tumors are defined as tumors with ≥ 10 percent of examined tumor cells that stained positive for p53 antigen; p53 IHC- tumors are defined as tumors with < 10 percent staining of cells.

Table 5.3. Estimated odds ratios and 95% confidence intervals for p53 Mut or Wt (mRNA) breast cancer subtypes (adjusted for age and race).

	Overall	**ER+ Cases**	**ER- Cases**
	Mut vs. Wt	Mut vs. Wt	Mut vs. Wt
Smoking history			
Ever	1.11 (0.86, 1.45)	1.08 (0.75, 1.54)	0.79 (0.37, 1.66)
Smoking status			
Former	1.11 (0.81, 1.53)	1.17 (0.74, 1.84)	1.07 (0.47, 2.43)
Current	1.11 (0.80, 1.56)	0.99 (0.64, 1.54)	0.43 (0.12, 1.50)
Duration (years)			
< 10	1.26 (0.86, 1.86)	1.10 (0.64, 1.87)	1.37 (0.52, 3.68)
11-20	0.94 (0.60, 1.47)	1.02 (0.55, 1.88)	0.68 (0.19, 2.49)
> 20	1.11 (0.80, 1.54)	1.09 (0.69, 1.72)	0.50 (0.16, 1.56)
Dose (packs/day)			
< 1/2	1.15 (0.81, 1.63)	1.05 (0.65, 1.70)	1.07 (0.43, 2.69)
1/2 – 1	1.13 (0.80, 1.61)	1.14 (0.70, 1.86)	0.47 (0.13, 1.67)
> 1	1.01 (0.64, 1.60)	0.99 (0.52, 1.89)	0.81 (0.22, 3.00)
Time since exposure (years)			
< 5	1.14 (0.73, 1.77)	1.07 (0.57, 2.02)	0.89 (0.24, 3.34)
5-10	1.26 (0.74, 2.17)	1.80 (0.76, 4.24)	0.84 (0.18, 3.99)
11-20	0.85 (0.40, 1.79)	0.67 (0.23, 1.94)	0.71 (0.08, 6.02)
> 20	1.11 (0.81, 1.51)	1.02 (0.67, 1.55)	0.75 (0.30, 1.88)

Note. ER+ tumors are defined as tumors with ≥ 10 percent of examined tumor cells that stained positive for ER antigen; ER- tumors are defined as tumors with < 10 percent staining of cells. P53 mutant and wildtype designations are defined using a multigene RNA signature[65].

A. Among ER+ Cases



B. Among ER- Cases

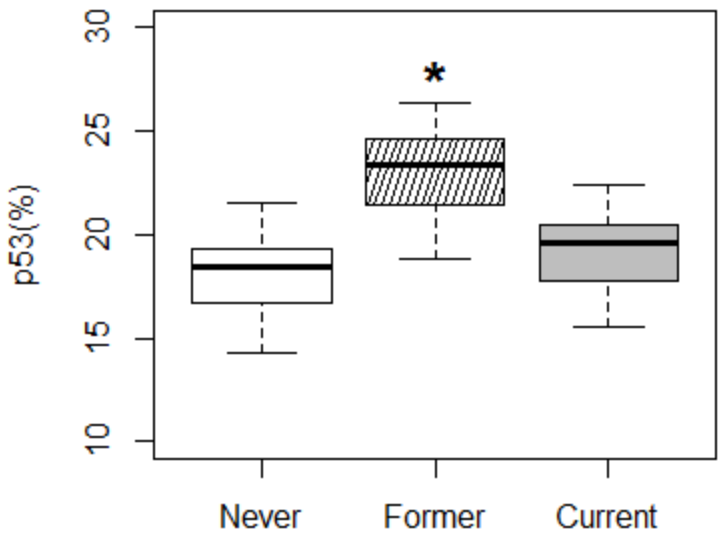


Figure 5.3. Boxplots displaying the distribution of weighted percent p53 protein (%).

Note. Boxplots for ER+ (**A**) and ER- (**B**) cases. p53(%) values were estimated from a linear regression model adjusted for age and race. Mean (SD): (**A**) Never [5.1 (0.03)], Former [4.5 (0.03)], Current [5.04 (0.03)]; (**B**) Never [18.1 (1.75)], Former [23.1 (1.84)], Current [19.3 (1.74)].

A. Among ER+ Cases

B. Among ER- Cases

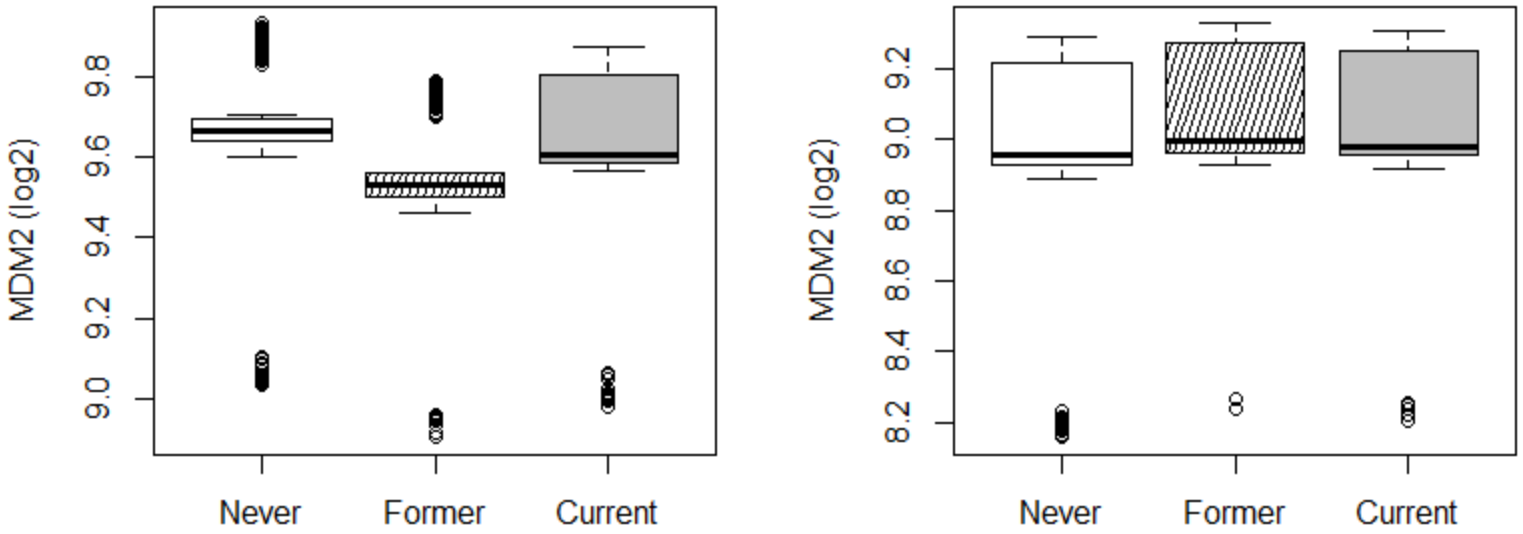


Figure 5.4. Boxplots displaying the distribution of MDM2 mRNA (log2)

Note. Boxplots for ER+ (**A**) and ER- (**B**) cases. MDM2 values were estimated from a linear regression model adjusted for age, race, and Nanostring code set. Mean (SD): (**A**) Never [9.7 (0.2)], Former [9.5 (0.2)], Current [9.6 (0.2)]; (**B**) Never [9.0 (0.3)], Former [9.0 (0.2)], Current [9.0 (0.3)].

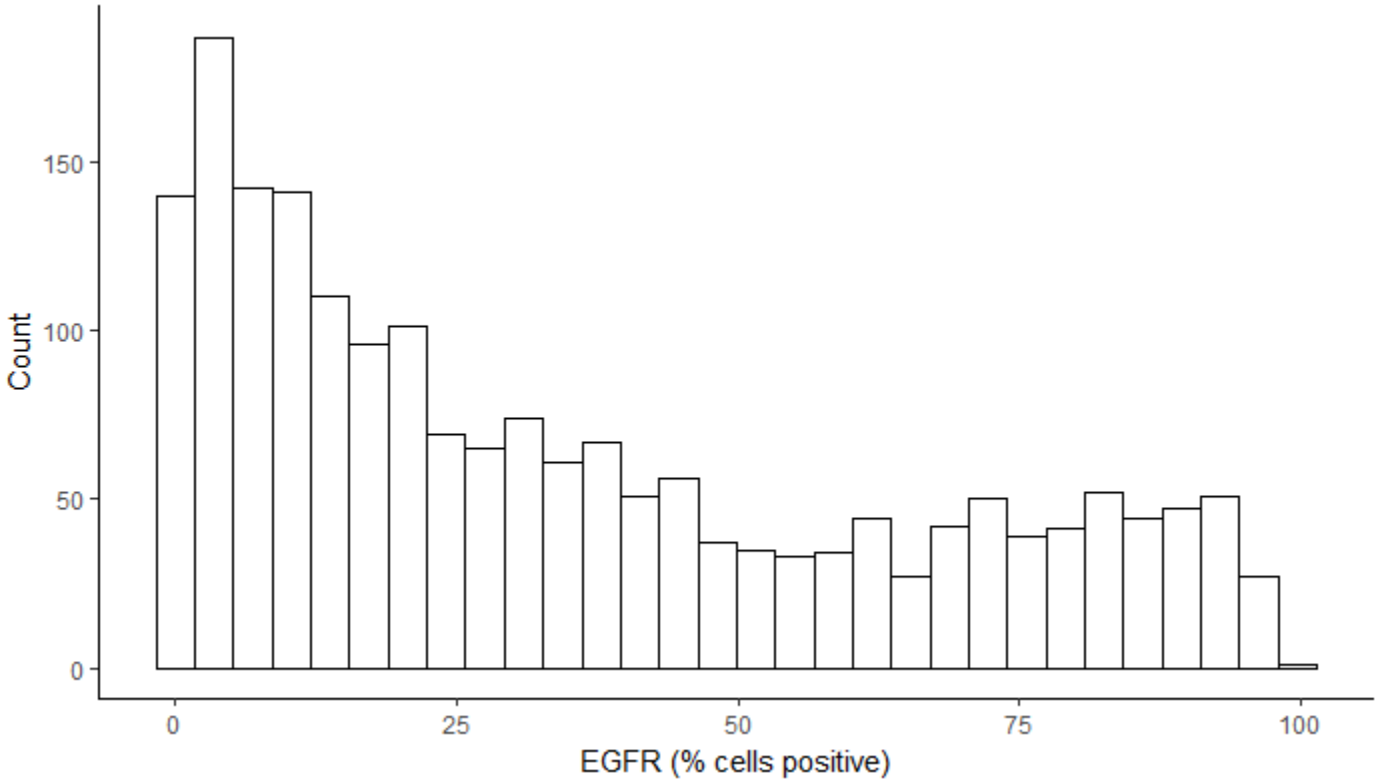


Figure 5.5. Histogram of EGFR protein expression values among 1,964 breast tumors in CBCS III

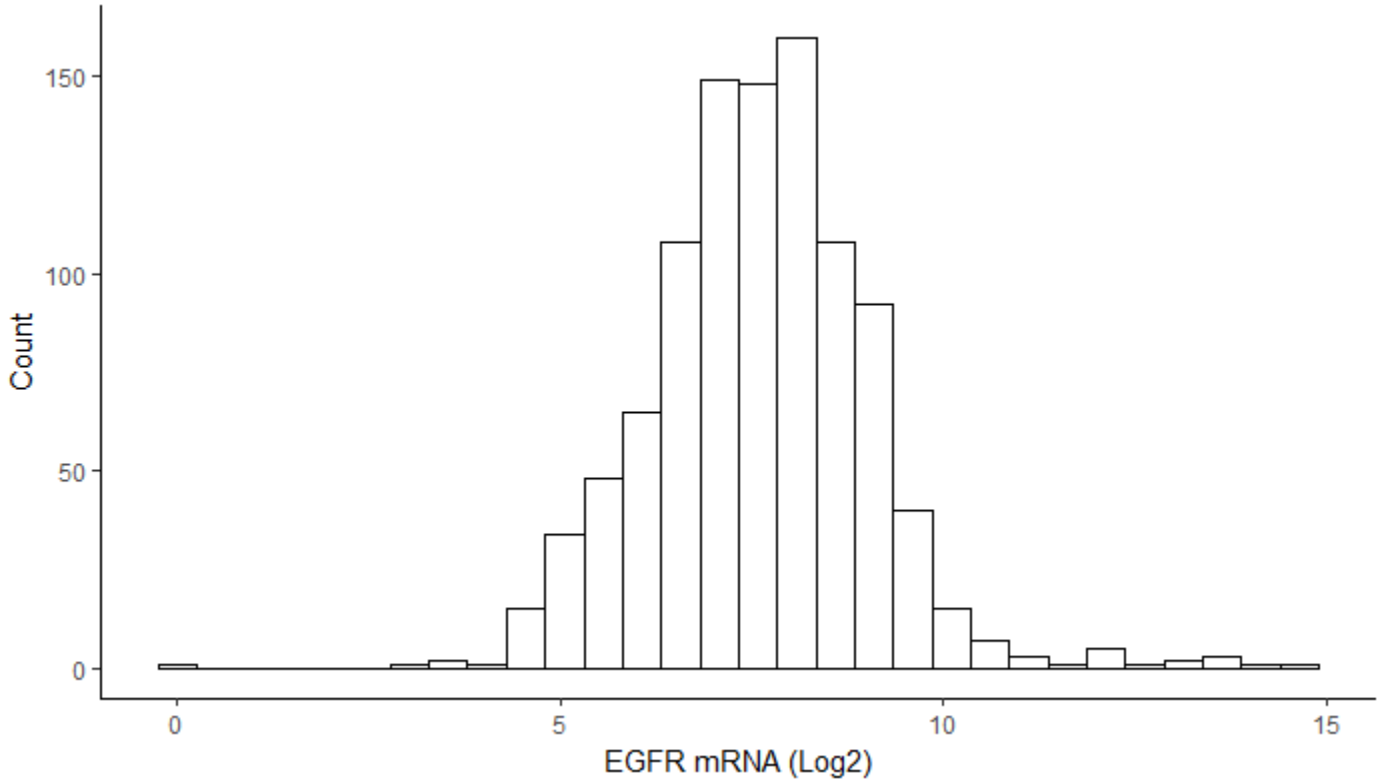


Figure 5.6. Histogram of EGFR mRNA expression values among 1,011 breast tumors in CBCS III.

Table 5.4. Estimated odds ratios and 95% confidence intervals for EGFR IHC-defined breast cancer subtypes (adjusted for age and race).

	Overall	**ER+ Cases**	**ER- Cases**
	EGFR+ vs. EGFR-	EGFR+ vs. EGFR-	EGFR+ vs. EGFR-
Smoking history			
Ever	0.95 (0.77, 1.16)	1.03 (0.82, 1.30)	1.14 (0.61, 2.11)
Smoking status			
Former			
Current	0.99 (0.78, 1.26)	1.09 (0.83, 1.43)	1.20 (0.55, 2.60)
Duration (years)	0.89 (0.68, 1.16)	0.96 (0.71, 1.29)	1.07 (0.48, 2.41)
< 10			
11-20	1.17 (0.85, 1.62)	1.26 (0.88, 1.80)	1.54 (0.53, 4.51)
> 20	0.95 (0.68, 1.33)	1.05 (0.72, 1.53)	1.18 (0.40, 3.49)
Dose (packs/day)	0.84 (0.65, 1.08)	0.91 (0.69, 1.21)	0.93 (0.42, 2.06)
< 1/2			
1/2 – 1	1.06 (0.80, 1.40)	1.21 (0.88, 1.67)	0.96 (0.44, 2.09)
> 1	1.01 (0.77, 1.33)	1.13 (0.83, 1.54)	1.19 (0.48, 2.97)
Time since exposure (years)	0.70 (0.50, 0.97)	0.67 (0.46, 0.97)	1.79 (0.99, 1.04)
< 5			
5-10	0.88 (0.69, 1.12)	0.96 (0.73, 1.28)	0.89 (0.44, 1.79)
11-20	1.06 (0.60, 1.89)	1.06 (0.57, 1.98)	NE
> 20	1.06 (0.71, 1.58)	1.32 (0.83, 2.09)	1.57 (0.36, 6.90)

Note. ER+ tumors are defined as tumors with ≥ 10 percent of examined tumor cells that stained positive for ER antigen; ER- tumors are defined as tumors with < 10 percent staining of cells. Similarly, EGFR+ tumors are defined as tumors with ≥ 10 percent of examined tumor cells that stained positive for EGFR antigen; EGFR- tumors are defined as tumors with < 10 percent staining of cells.

A. Among ER+ Cases

B. Among ER- Cases

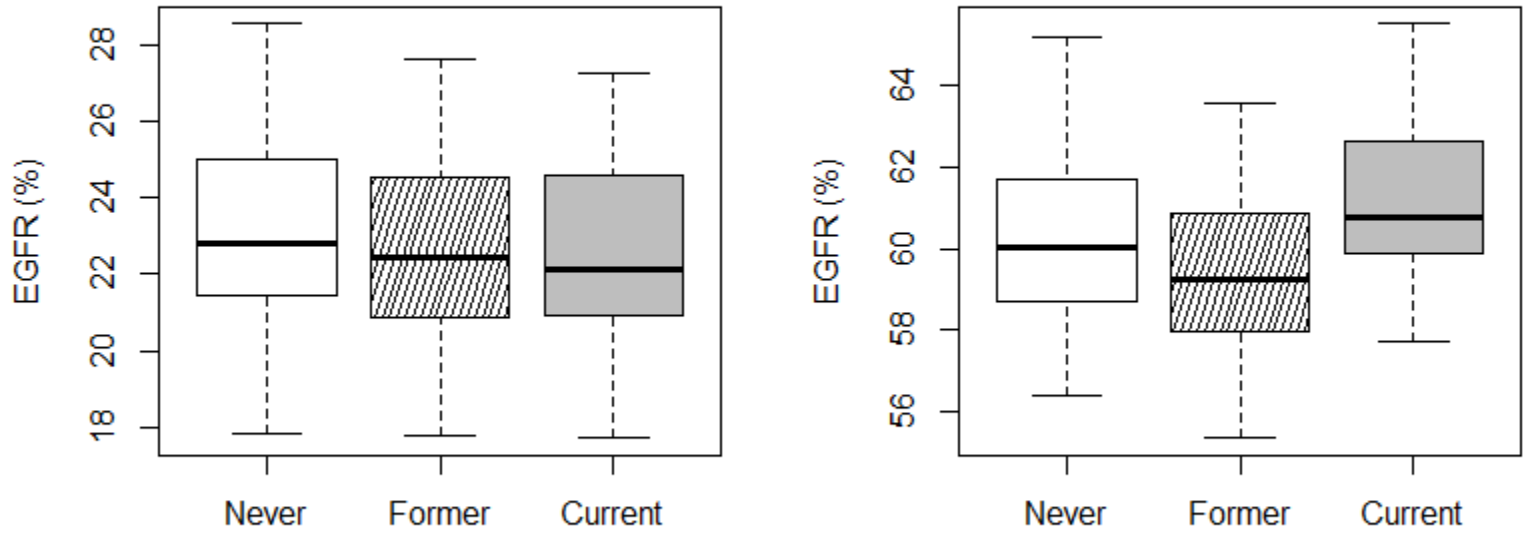


Figure 5.7. Boxplots displaying the distribution of weighted percent EGFR protein (%)

Note. Boxplots for ER+ (**A**) and ER- (**B**) cases. EGFR (%) values were estimated from a linear regression model adjusted for age and race. Mean (SD): (**A**) Never [22.9 (2.3)], Former [22.5 (2.4)], Current [22.5 (2.3)]; (**B**) Never [60.3 (1.9)], Former [59.5 (2.0)], Current [61.2 (1.8)].

A. Among ER+ Cases

B. Among ER- Cases

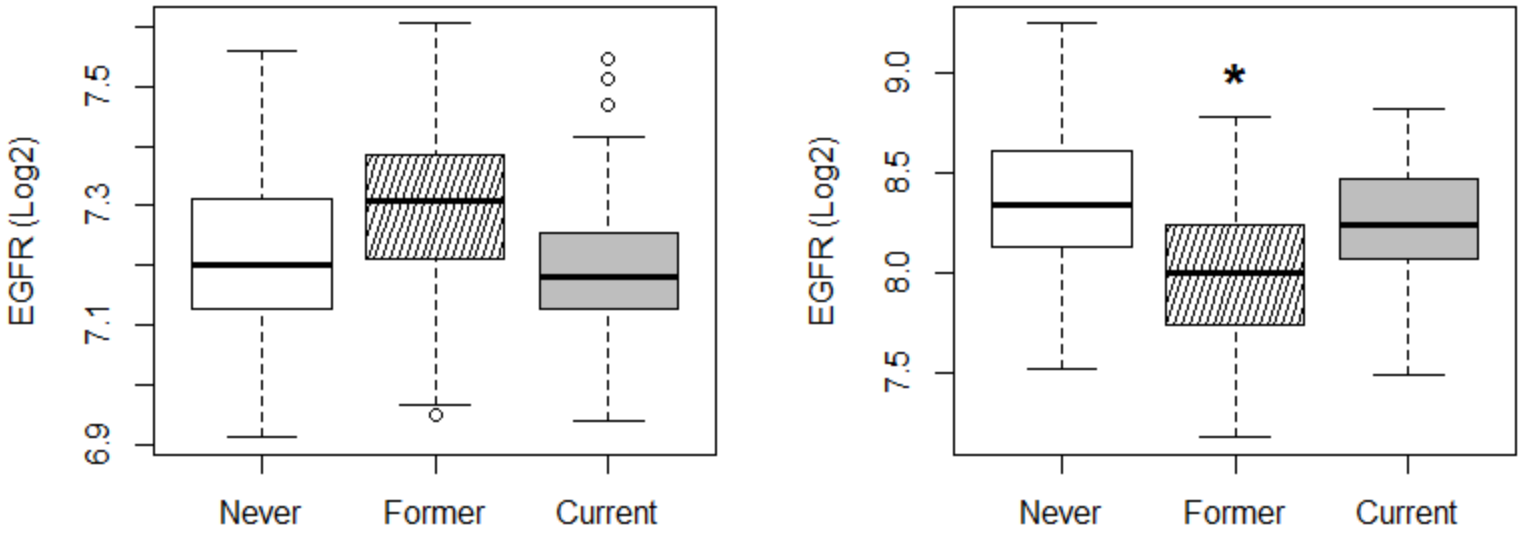


Figure 5.8. Boxplots displaying the distribution of EGFR mRNA (log2)

Note. Boxplots for ER+ (**A**) and ER- breast tumors (**B**). EGFR values were estimated from a linear regression model adjusted for age, race, and Nanostring code set. Mean (SD): (**A**) Never [7.2 (0.1)], Former [7.3 (0.1)], Current [7.2 (0.1)]; (**B**) Never [8.4 (0.4)], Former [8.0 (0.4)], Current [8.2 (0.3)].

Table 5.5. Estimated odds ratios and 95% confidence intervals for proliferation score breast cancer subtypes (adjusted for age and race).

	Overall	**ER+ Cases**	**ER- Cases**
	PS+ vs. PS-	PS+ vs. PS-	PS+ vs. PS-
Smoking history			
Ever	0.69 (0.53, 0.90)	0.74 (0.50, 1.12)	0.70 (0.45, 1.11)
Smoking status			
Former	0.71 (0.51, 0.99)	0.79 (0.48, 1.32)	0.59 (0.35, 1.02)
Current	0.67 (0.47, 0.94)	0.70 (0.42, 1.17)	0.89 (0.48, 1.66)
Duration (years)			
< 10	0.69 (0.46, 1.04)	0.83 (0.46, 1.50)	0.68 (0.34, 1.34)
11-20	0.67 (0.42, 1.08)	0.66 (0.32, 1.37)	0.66 (0.31, 1.41)
> 20	0.70 (0.49, 0.99)	0.73 (0.43, 1.23)	0.75 (0.41, 1.36)
Dose (packs/day)			
< 1/2	0.71 (0.50, 1.02)	0.86 (0.51, 1.46)	0.65 (0.36, 1.18)
1/2 – 1	0.66 (0.46, 0.96)	0.64 (0.36, 1.13)	0.82 (0.43, 1.57)
> 1	0.69 (0.42, 1.14)	0.72 (0.33, 1.56)	0.64 (0.28, 1.43)
Time since exposure (years)			
< 5	0.83 (0.52, 1.33)	1.12 (0.56, 2.22)	0.64 (0.29, 1.41)
5-10	0.51 (0.28, 0.94)	0.30 (0.09, 1.01)	0.64 (0.25, 1.63)
11-20	1.13 (0.53, 2.39)	1.43 (0.48, 4.31)	1.12 (0.28, 4.47)
> 20	0.65 (0.47, 0.89)	0.69 (0.43, 1.12)	0.70 (0.41, 1.23)

Note. ER+ tumors are defined as tumors with ≥ 10 percent of examined tumor cells that stained positive for ER antigen; ER- tumors are defined as tumors with < 10 percent staining of cells. PS+ and PS- designations are defined using a multigene RNA signature[72].

CHAPTER 6: DISCUSSION

6.1 Summary

In this dissertation, we examined relationships between smoking and breast tumor expression of protein and mRNA biomarkers that reflect biologic activity for estrogen-mediated, genotoxic, and growth-factor dependent mechanisms of carcinogenesis. Using a case-only study design, we employed logistic and linear regression to evaluate binary and quantitative outcomes for biomarker expression. We used both single gene and multigene markers to characterize breast tumors, in efforts to explore associations between smoking and cross-sectional assessments of signaling pathways. Finally, we also incorporated latency functions to characterize protracted smoking exposure in our retrospective analysis to identify temporal associations between smoking and odds of binary breast cancer subtypes.

6.2 Main Findings

6.2.1 Smoking and estrogen-receptor expression in breast tumors

Findings from our study lend support to the hypothesis that smoking may be linked to breast cancer via estrogen-mediated pathways. In our case-only study of nearly 2,000 patients, we observed increased odds of the ER+ subtype for temporal and dose-dependent measures of smoking. We also demonstrated that these associations hold for ER-related subtypes characterized by ESR1 mRNA and a multigene luminal score (LS)[72] that is comprised of expression values for eight genes involved in estrogen-signaling pathways. Notably, results from our study suggested that time of smoking exposure was an important

predictor of having a tumor that has high expression of genes involved in estrogen-signaling. Increased odds of ER+, ESR1+, and LS+ subtypes was most apparent among women who were self-reported current smokers at time of diagnosis. We also observed higher quantitative levels of ESR1 mRNA among current smokers, thereby suggesting that pre-diagnostic smoking proximal to date of diagnosis may modulate estrogen-receptor expression in breast tumors. Though we did not observe this same association with quantitative measures of protein, we suggest that mRNA may be a more sensitive measure of the biological activity for the estrogen-receptor.

6.2.2 Smoking and biomarkers for p53, EGFR, and cell proliferation

Though the literature on smoking and breast cancer risk is vast, few studies have assessed smoking in relation to specific biomarkers beyond ER. One longstanding hypothesis suggests that cigarette smoke may cause DNA-damage in breast tissue through the development of DNA adducts[59,60]. We selected overexpression of the p53 tumor suppressor gene as a marker of one DNA damage pathway. In addition, we evaluated EGFR, given a recent hypothesis implicating upregulated EGFR expression as biomarker linked to nicotine exposure in tissue culture models[29]. However, in our analysis we did not find evidence of links between smoking and protein measures for p53-defined or EGFR-defined breast cancer subtypes. Specifically, we observed no association between smoking measures and breast cancer subtypes defined by binary cut points for p53 and EGFR protein (i.e., p53 IHC+/p53 IHC- and EGFR+/EGFR-), where p53 IHC+ and EGFR+ subtypes were defined as greater than or equal to ten percent of examined tumor cells.

Guided by the literature and findings in Aim 1, we recognized ER as a biomarker whose dichotomization distinguished two distinct classes of breast tumors[7,8]. So, we

considered ER subtype when evaluating the association between smoking and our chosen biomarkers with proposed links to smoking and DNA-damage or growth factor-mediated mechanisms in breast tumors. When we stratified our analyses by ER status, we found the suggestion of a possible link between smoking and p53-positvity (p53 IHC+) among ER- cases. For EGFR, the stratified associations were less clear; differing metrics for smoking were linked to either increased or decreased odds of EGFR-positivity (EGFR+) among either ER+ or ER- cases. In addition, though we observed no association between smoking and quantitative measures for p53 protein or EGFR protein for breast cancer cases overall, differential expression of p53 protein was observed among former smokers, thereby enhancing the argument that past smoking history may be linked to differential expression of p53 protein in breast tumors.

In addition to protein classifications for p53 and EGFR, we also considered single and multigene RNA measures that reflect biologic activity of p53, EGFR, and cell proliferation signaling pathways. We observed no clear patterns for associations between smoking measures and continuous measures for MDM2 (i.e., a negative regulator or p53) or EGFR mRNA expression. These observations suggest that smoking exposure captured at time of breast cancer diagnosis may not be related to modulation of these two biomarkers. In addition, our multigene RNA signature used to classify breast tumors as p53 “wildtype” or “mutant” was not linked to smoking exposure for breast cancer cases overall or when stratified by ER status. As discussed in chapter 5, the p53 RNA signature captures an array of mutational forms linked to defects in p53 regulation and may be too broad a measure for DNA-damage in our study if smoking is specifically linked to p53 missense mutations.

During the conceptualization of this dissertation, we considered EGFR as a marker of a growth-factor mediated mechanism but found inconsistent links to smoking metrics. We later incorporated a cell proliferation signature to evaluate whether smoking was linked to differential proliferative activity. Interestingly, the multigene cell proliferation signature was consistently linked to decreased odds of the PS+ subtype for all smoking measures – thereby suggesting that smoking may be linked to lower cell proliferation rates in breast tumors. Though beyond the scope of the dissertation, future work may seek to evaluate whether EGFR gene expression mediates the relationship between smoking and the cell proliferation signature in breast tumors.

6.3 Breast Tumor Biomarker Expression

Evaluation of biomarkers in tumors can be performed via several methods including the use of high throughput technologies to quantify protein and mRNA. Though these methods allow advantages for large epidemiologic studies seeking to evaluate underlying mechanisms that link exposures to breast cancer risk, several measurement and classification issues emerge. These issues include: (1) classifying continuous variables into categorical or binary groups; (2) challenges with interpretation of biomarkers in relation to etiology vs. progression; (3) discordance between related biomarkers (i.e. protein vs. RNA); and (4) issues related to interpretation of single vs. multigene markers.

6.3.1 Classifying continuous variables into categorical or binary groups

Though we considered continuous gene expression, molecular biomarkers are routinely used in a dichotomous state – most notably estrogen-receptor (ER) protein. Contemporary clinical guidelines indicate an IHC cut point of 1% to define ER-positivity, though tumors with expression values between 1% and 9% are routinely considered as

borderline ER+ and those at 10% or higher as ER+[62,92]. The percentage of cells stained for ER follows a right-skewed distribution due to properties of the assay, which are tuned for high sensitivity to detect ER protein expression[93]. In our study, we used a threshold of 10 % typically define tumors as "strong" positives, yielding binary categories for ER that are suitable for epidemiologic investigations (Figure 6.1).

Notably, our selection of 10% as a binary cut point for protein measures of p53 and EGFR mimics the 10% cut point used for ER protein expression. Although we explored use of a data-driven algorithm for establishing cut points[75], we did not identify two or more distinct classes for either biomarker. Using the 10% cut point, we observed trends for p53 and EGFR protein similar to that for ER, with right-skewed distributions characterized by a proportion of tumors with low biomarker expression and a uniform flat tail for higher expression values (Figure 6.2) (Figure 6.3). However, we acknowledge that more suitable cut points may exist for these two markers and that future work should seek to examine characteristics of IHC assays for p53 and EGFR quantification that may inform cut point identification. When utilizing established clinical biomarkers, dichotomization may rely on standard cut points, but when using other biomarkers, it is sometimes necessary to establish cut points. Standards for these dichotomizations have not been developed. Herein we present some approaches and rationale, but as genomic biomarkers become more widely utilized, standardized methods for categorization will be needed. Procedures may also differ for protein vs. mRNA cut points, as discussed in the following section.

6.3.2 Discordance between related biomarkers

With the inclusion of both protein and mRNA biomarker data, we were positioned to examine whether observed associations with smoking were consistent across marker types.

As previously discussed, smoking was associated with higher quantitative levels of ESR1 mRNA, but not ER protein. We suggested that properties of the IHC assay may be attributed to this observed discordance, where sensitivity of the assay leads to saturated signals and reduced dynamic range for protein expression. It follows that binary cut points may be more appropriate for protein biomarkers in studies of etiology. Other molecular techniques such as those for RNA assessment may be more suitable for biomarker quantification.

Nevertheless, we also sought to evaluate possible binary cut points for ESR1, MDM2, and EGFR mRNA to determine whether these markers performed similarly to their binary protein counterparts. The expectation-maximization algorithm – used to evaluate Gaussian mixture models – yielded two distinct populations of breast tumors for ESR1, but not MDM2 or EGFR. When we considered the distribution of log-transformed values for ESR1 mRNA, we observed a bimodal distribution that suggests a mixture of 2 breast tumor populations (Figure 6.4). The distributions of MDM2 and EGFR mRNA were both normally distributed but unimodal, thereby suggesting single populations of expression values for these two markers (Figure 6.5) (Figure 6.6); thus, these markers were suitable for analyses that examined continuous gene expression, but there was no evidence of a reasonable binary cut point. Further, the identification of a binary cut point for ESR1 mRNA, which had high percent agreement with ER+ and ER- IHC cut points, highlights the unique nature of ER and its role in breast cancer characterization.

6.3.3 Etiology vs. Progression

Tumor biomarkers serve a variety of functions in both clinical and research settings. Given that tumor biomarkers may reflect etiologic effects or effects related to disease progression, it is often challenging to interpret their associations with proposed risk factors.

ER is a particularly complex marker as it may function as a prognostic, predictive, and etiologic marker. Overexpression of ER in breast tumors is generally regarded as signifying a less aggressive form of the disease, but has also been linked to higher rates of recurrence when compared to ER- tumors. It is also used to predict response to anti-estrogenic therapies and has demonstrated associations with distinct risk factor profiles for ER+ and ER- breast cancer subtypes[13].

Our previous work in the Carolina Breast Cancer Study demonstrated distinct associations between smoking and the ER+ subtype, which further suggests ER+ breast tumors comprise a distinct etiologic type. However, in the present study we observed that pre-diagnostic smoking close to time of diagnosis was linked to modulation of quantitative ESR1 mRNA expression levels in breast tumors and that smoking during this same time period was associated with the increased odds of being diagnosed with an ER+ or ESR1+ tumor. Thus, while smoking may be linked to specific etiologic subtypes defined by ER and it may also modulate ESR1 biomarker expression in existing tumors. It is unclear whether this quantitative variation in ER expression is related to etiology, progression, response to therapy, or some combination of these. Future studies should aim to evaluate whether modulation of ER or ESR1 expression levels by smoking is linked to cancer progression events and cancer outcomes.

When we evaluated p53 and EGFR as potential biomarkers of DNA-damage and growth factor mediated effects related to smoking exposure in breast tumors, we did not observe a consistent association between smoking and either biomarker. However, we also noted the importance of estrogen-receptor expression, which separates breast tumors into two distinct populations. Accordingly, we sought to examine the relationships between smoking

and p53 or EGFR gene expression separately for ER+ and ER- breast tumors. Though limited by sample size, we observed possible associations between smoking and p53 or EGFR biomarkers when we restricted our analyses by ER subtype. Thus, the influence of an exposure on a proposed marker of etiology or progression may function within breast cancer types defined by ER status.

The various smoking metrics in our study allowed us to examine temporal associations to infer the most relevant time-windows for smoking and biomarker expression. We used logistic regression models with latency parameters to simultaneously model dose, duration, and time of exposure to demonstrate that the most relevant period for smoking and ER-defined breast cancer may be during pre-diagnostic smoking exposure proximal to time of diagnosis. This finding may have clinical implications for molecular subtyping for breast cancer patients who are smokers at time of diagnosis, particularly for cases with borderline positivity. Future studies should seek to evaluate other proposed breast cancer risk factors in relation to estrogen-receptor expression to evaluate the extent to which exogenous exposures influence breast cancer intrinsic subtyping. In addition, we were limited by smaller sample size in our restricted analyses used to evaluate the relationship between smoking exposure and p53 and EGFR expression. Future studies within the Carolina Breast Cancer Study should seek to combine data from case-only and case-control data in order to bolster sample size.

6.3.4 Single vs. Multigene Biomarkers

In addition to single gene biomarkers, we selected three multigene signatures that reflect estrogen-mediated, DNA-damage response, and cell proliferation signaling pathways. These multigene signatures offer improved resolution over single gene markers may offer

advantages in capturing the biological pathway activity. However, one challenge with multigene signatures is that a variety of approaches are used for binary classifications of tumors. For example, the luminal score is the average expression of the 8 genes included in the signature, where each gene is given equal weight. The p53 signature uses nearest centroid and Spearman correlation to classify tumors as mutant or wildtype. And the proliferation score is the sum of the average of 11 genes, the regression coefficient of the tumor's intrinsic subtype, and a measure of tumor size. Thus, our results may be somewhat sensitive to how the composite scores were devised; when there is no evidence of association for a particular measure, establishing true negative findings likely requires more thorough investigation of additional algorithms. Likewise, it may be helpful to confirm that positive findings do not depend upon the specific algorithm and/or evaluating the stability of classification when using different algorithms. In addition, each signature yields a quantitative value that may be useful in regression analysis. Future research may consider how to incorporate such quantitative scores in analyses and how to interpret these.

6.4 Conclusions

The goal of this dissertation was to examine the association between smoking and breast tumors biomarkers linked to proposed mechanisms of carcinogenesis. We observed potential associations between smoking and each of the proposed mechanistic pathways, but also encountered several challenges related to interpretation of the data. The technical and conceptual issues discussed herein are important considerations for linking exposure to tumor biomarker expression and warrant further investigation to improve future studies that attempt to integrate epidemiologic and molecular biology in population-based studies of cancer etiology. In the era of genomic testing, it will be important to understand how past and

present exposures influence tumor biology to understand whether these exposures have implications for etiology, progression, or both. Smoking, specifically, may influence both the etiology of cancer and – as shown in our study – may be linked to increased expression of ESR1 mRNA among women who are smokers at diagnosis. Future work should evaluate other exposures in relation to breast tumor biomarkers linked to proposed mechanisms of carcinogenesis.

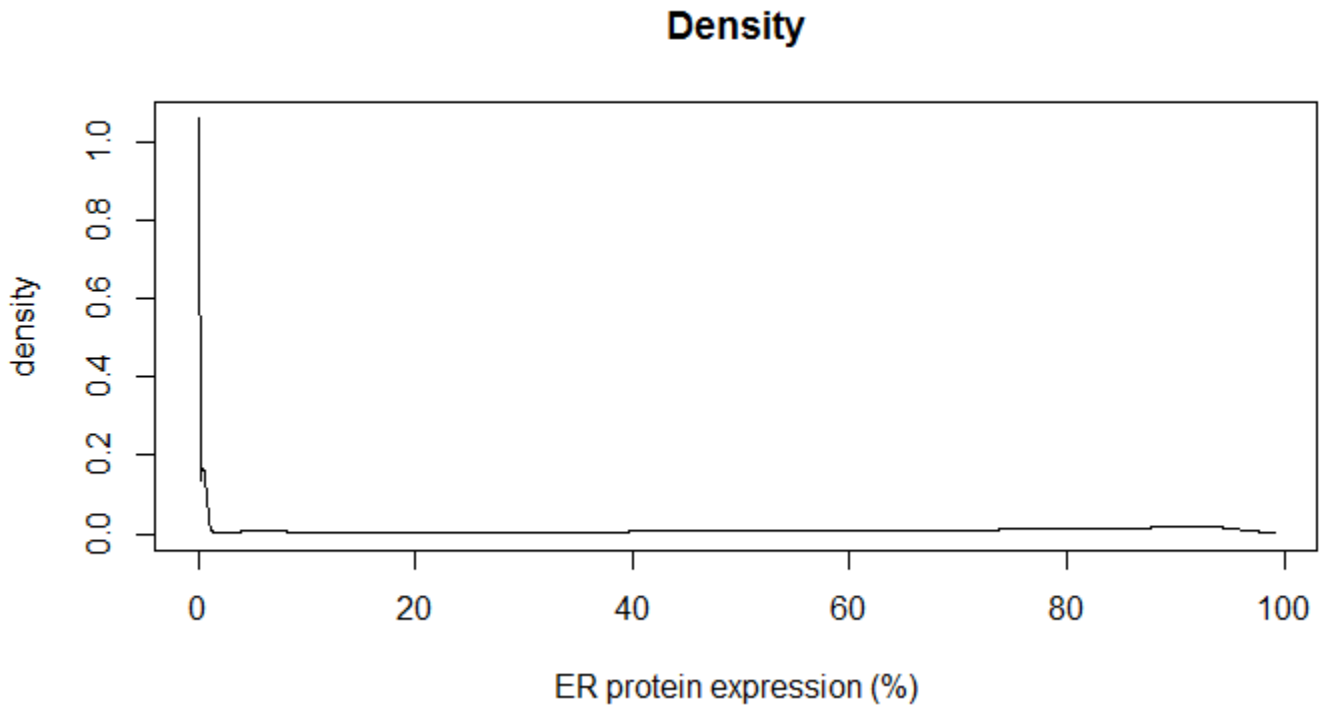


Figure 6.1. Distribution of estrogen-receptor (ER) protein expression values, as measured by immunohistochemistry.

Note. Values are calculated as percentage of cells that stained positive for ER.

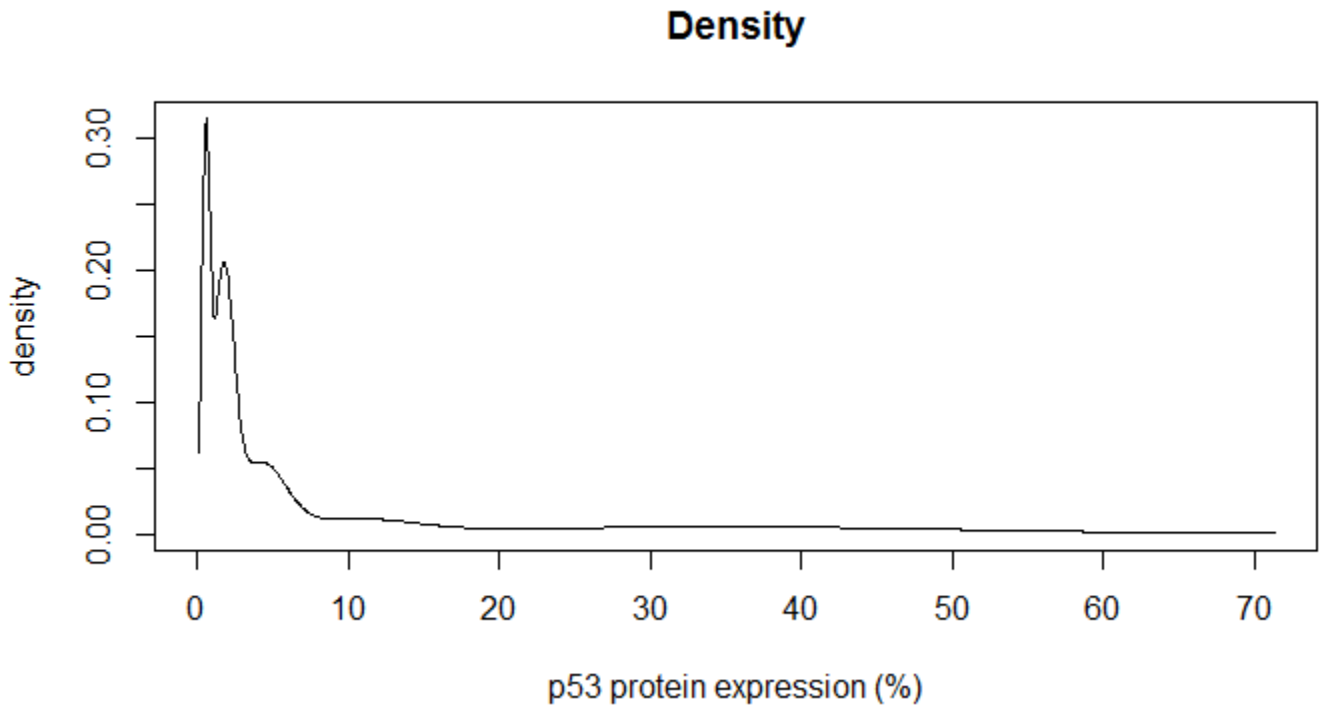


Figure 6.2. Distribution of p53 protein expression values.

Note. Values are calculated as percentage of cells that stained positive for p53.

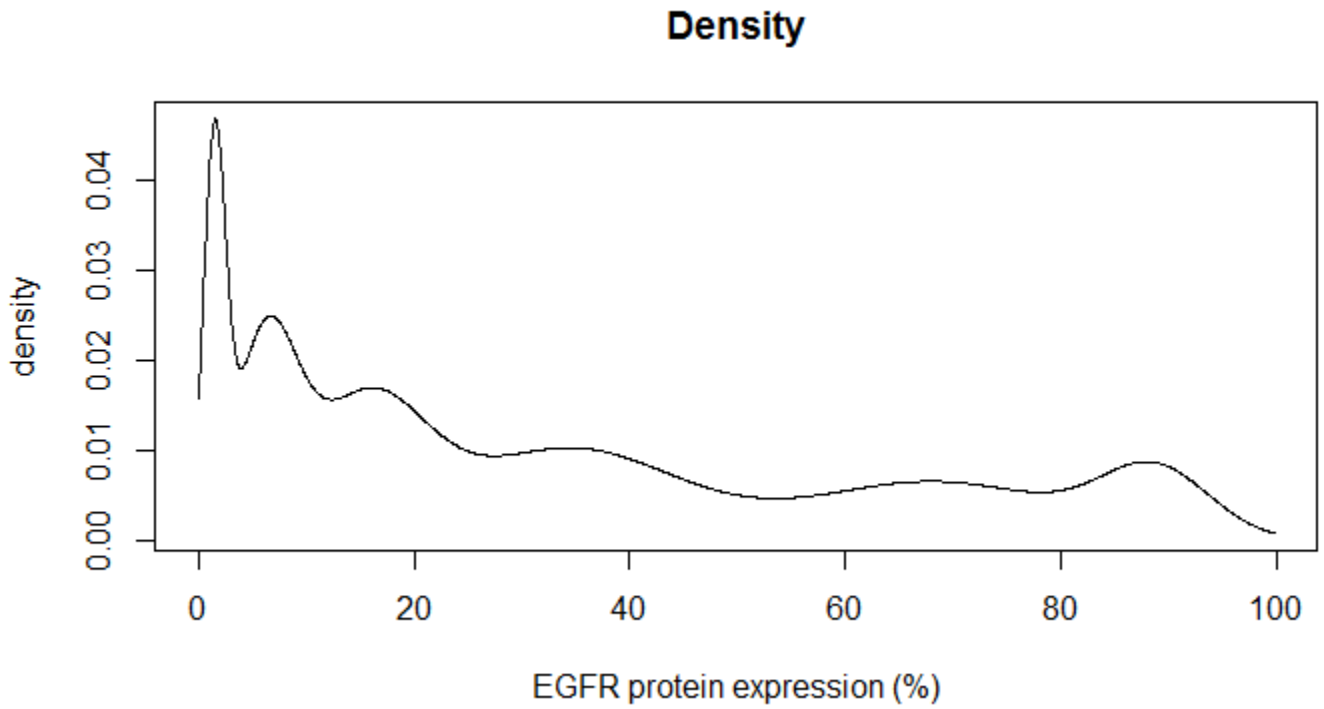


Figure 6.3. Distribution of epidermal growth factor receptor (EGFR) protein expression values.

Note. Values are calculated as percentage of cells that stained positive for EGFR.

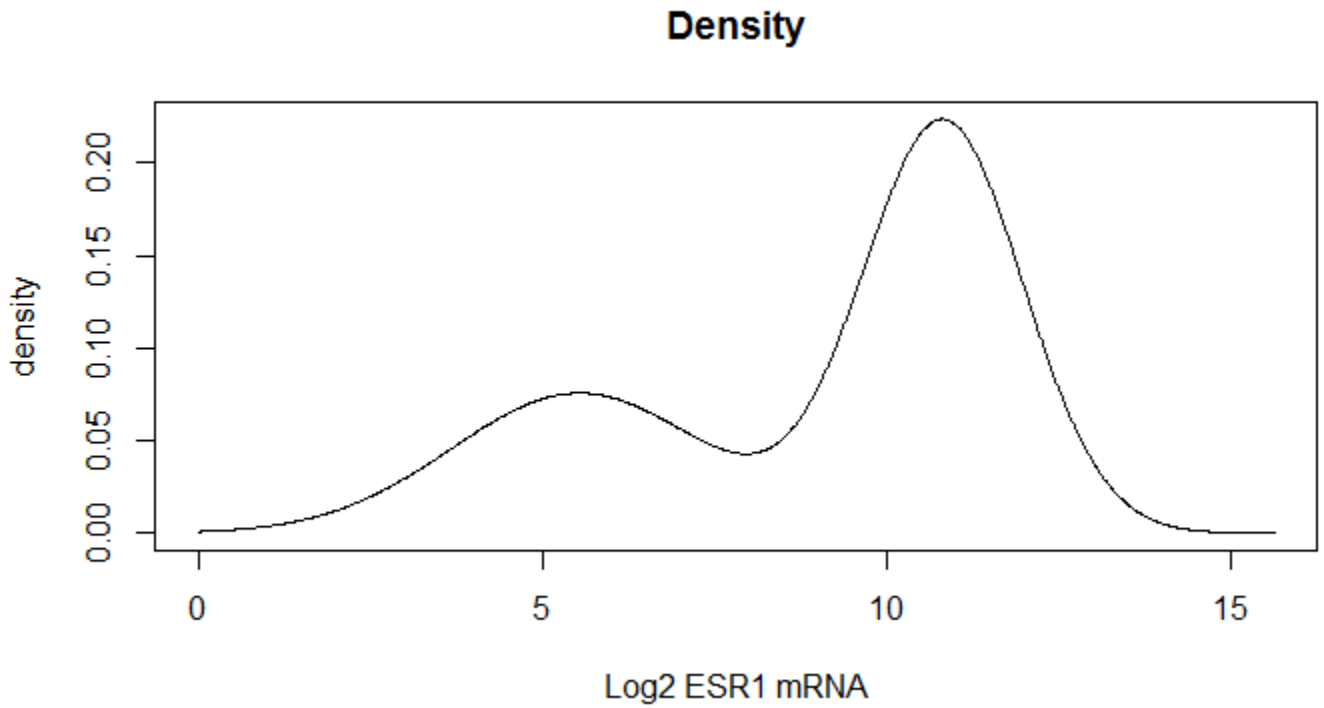


Figure 6.4. Distribution of ESR1 mRNA values (Log2).

Note. The bimodal peaks represent two distinct populations of ESR1-defined tumors.

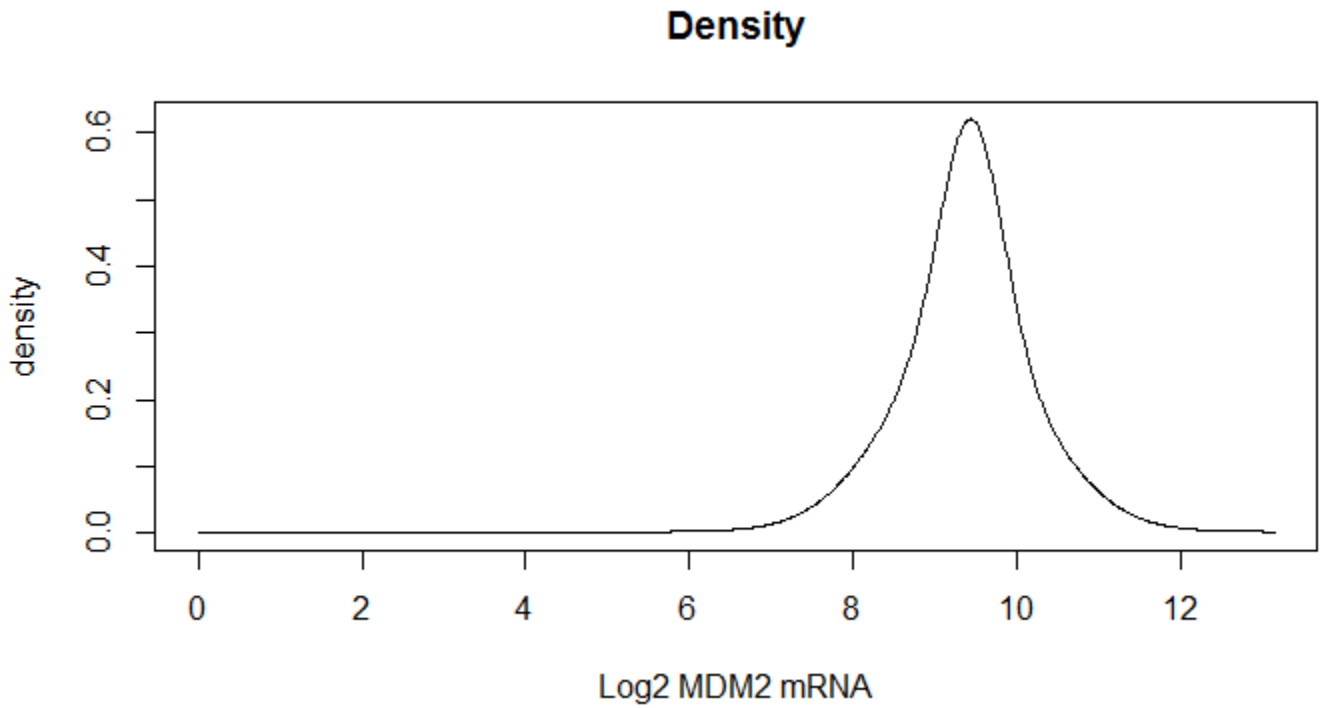


Figure 6.5. Distribution of MDM2 mRNA values (Log2).

Note. The unimodal peak suggests a single population of tumors defined by MDM2.

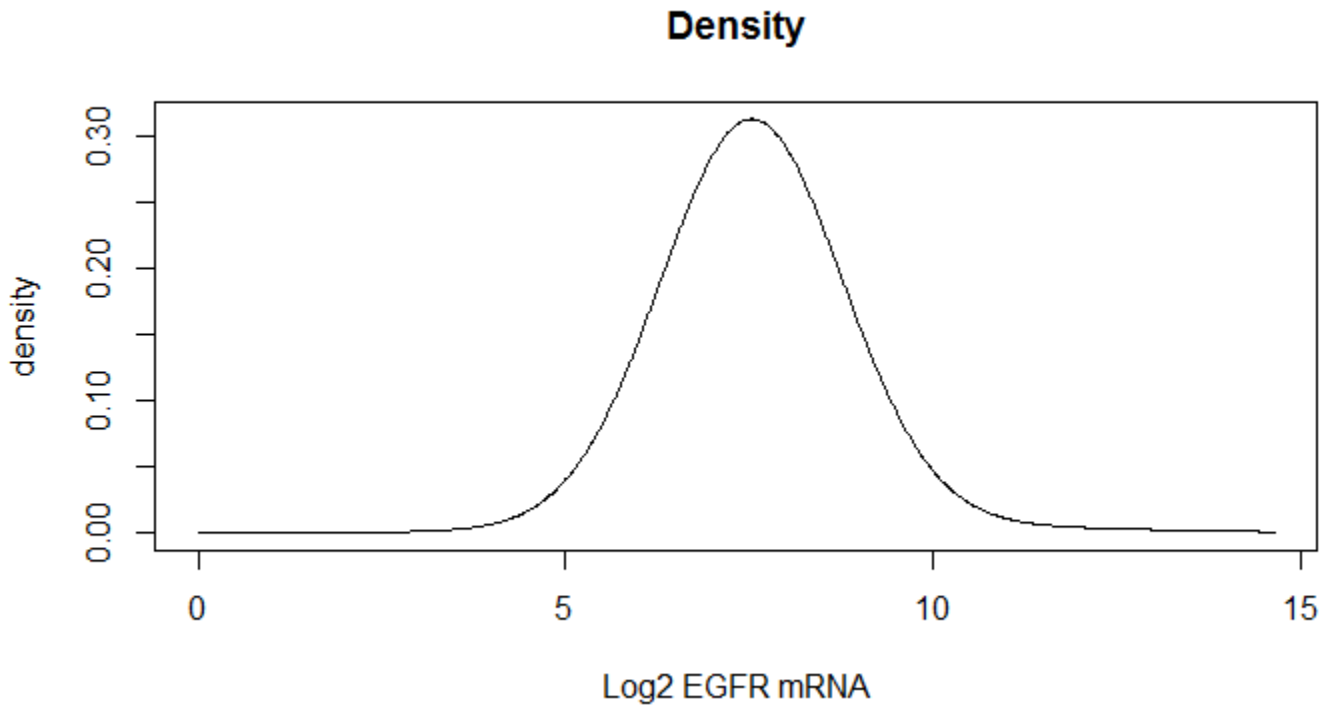


Figure 6.6. Distribution of EGFR mRNA values (Log2).

Note. The unimodal peak suggests a single population of tumors defined by EGFR.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA: a cancer journal for clinicians.* Jan 2016;66(1):7-30.

2. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America.* Jul 8 2003;100(14):8418-8423.

3. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature.* Aug 17 2000;406(6797):747-752.

4. Howlader N, Altekruse SF, Li CI, et al. US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *J Natl Cancer Inst.* May 2014;106(5).

5. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America.* Sep 11 2001;98(19):10869-10874.

6. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature.* Oct 4 2012;490(7418):61-70.

7. Anderson WF, Matsuno R. Breast cancer heterogeneity: a mixture of at least two main types? *J Natl Cancer Inst.* Jul 19 2006;98(14):948-951.

8. Anderson WF, Rosenberg PS, Prat A, Perou CM, Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? *J Natl Cancer Inst.* Aug 2014;106(8).

9. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell.* Aug 14 2014;158(4):929-944.**10.** Kwan ML, Kushi LH, Weltzien E, et al. Epidemiology of breast cancer subtypes in two prospective cohort studies of breast cancer survivors. *Breast Cancer Res.* 2009;11(3):R31.

11. Trivers KF, Lund MJ, Porter PL, et al. The epidemiology of triple-negative breast cancer, including race. *Cancer causes & control : CCC.* Sep 2009;20(7):1071-1082.

12. Yang XR, Chang-Claude J, Goode EL, et al. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. *J Natl Cancer Inst.* Feb 2 2011;103(3):250-263.

13. Millikan RC, Newman B, Tse CK, et al. Epidemiology of basal-like breast cancer. *Breast Cancer Res Treat.* May 2008;109(1):123-139.

14. Phipps AI, Malone KE, Porter PL, Daling JR, Li CI. Reproductive and hormonal risk factors for postmenopausal luminal, HER-2-overexpressing, and triple-negative breast cancer. *Cancer.* Oct 1 2008;113(7):1521-1526.

15. Tamimi RM, Colditz GA, Hazra A, et al. Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. *Breast Cancer Res Treat.* Jan 2012;131(1):159-167.

16. Yang XR, Sherman ME, Rimm DL, et al. Differences in risk factors for breast cancer molecular subtypes in a population-based study. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* Mar 2007;16(3):439-443.

17. Williams LA, Olshan AF, Tse CK, Bell ME, Troester MA. Alcohol intake and invasive breast cancer risk by molecular subtype and race in the Carolina Breast Cancer Study. *Cancer causes & control : CCC.* Feb 2016;27(2):259-269.

18. Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE. Risk factors for breast cancer according to estrogen and progesterone receptor status. *J Natl Cancer Inst.* Feb 4 2004;96(3):218-228.

19. Rusiecki JA, Holford TR, Zahm SH, Zheng T. Breast cancer risk factors according to joint estrogen receptor and progesterone receptor status. *Cancer Detect Prev.* 2005;29(5):419-426.

20. Rosner B, Glynn RJ, Tamimi RM, et al. Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. *Am J Epidemiol.* Jul 15 2013;178(2):296-308.

21. Agaku IT, King BA, Dube SR, Centers for Disease C, Prevention. Current cigarette smoking among adults - United States, 2005-2012. *MMWR. Morbidity and mortality weekly report.* Jan 17 2014;63(2):29-34.

22. Gammon MD, Eng SM, Teitelbaum SL, et al. Environmental tobacco smoke and breast cancer incidence. *Environ Res.* Oct 2004;96(2):176-185.

23. Petrakis NL, Gruenke LD, Beelen TC, Castagnoli N, Jr., Craig JC. Nicotine in breast fluid of nonlactating women. *Science.* Jan 20 1978;199(4326):303-305.

24. Hill P, Wynder EL. Nicotine and cotinine in breast fluid. *Cancer letters.* Apr 1979;6(4-5):251-254.

25. Cheang MC, Martin M, Nielsen TO, et al. Defining breast cancer intrinsic subtypes by quantitative receptor expression. *The oncologist.* May 2015;20(5):474-482.

26. *Breast Cancer Epidemiology.* 2010.

27. Lee CH, Huang CS, Chen CS, et al. Overexpression and activation of the alpha9-nicotinic receptor during tumorigenesis in human breast epithelial cells. *J Natl Cancer Inst.* Sep 8 2010;102(17):1322-1335.

28. Filosto S, Becker CR, Goldkorn T. Cigarette smoke induces aberrant EGF receptor activation that mediates lung cancer development and resistance to tyrosine kinase inhibitors. *Molecular cancer therapeutics.* Apr 2012;11(4):795-804.

29. Nishioka T, Kim HS, Luo LY, Huang Y, Guo J, Chen CY. Sensitization of epithelial growth factor receptors by nicotine exposure to promote breast cancer cell growth. *Breast Cancer Res.* 2011;13(6):R113.

30. Sung H, Garcia-Closas M, Chang-Claude J, et al. Heterogeneity of luminal breast cancer characterised by immunohistochemical expression of basal markers. *Br J Cancer.* Feb 2 2016;114(3):298-304.

31. Stebbing J, Thiyagarajan A, Surendrakumar V, et al. Epidermal growth factor receptor status in early stage breast cancer is associated with cellular proliferation but not cross-talk. *Journal of clinical pathology.* Sep 2011;64(9):829-831.

32. Aziz SA, Pervez S, Khan S, Kayani N, Rahbar MH. Epidermal growth factor receptor (EGFR) as a prognostic marker: an immunohistochemical study on 315 consecutive breast carcinoma patients. *JPMA. The Journal of the Pakistan Medical Association.* Mar 2002;52(3):104-110.

33. Park K, Han S, Shin E, Kim HJ, Kim JY. EGFR gene and protein expression in breast cancers. *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology.* Oct 2007;33(8):956-960.

34. Rimawi MF, Shetty PB, Weiss HL, et al. Epidermal growth factor receptor expression in breast cancer association with biologic phenotype and clinical outcomes. *Cancer.* Mar 1 2010;116(5):1234-1242.

35. Nakajima H, Ishikawa Y, Furuya M, et al. Protein expression, gene amplification, and mutational analysis of EGFR in triple-negative breast cancer. *Breast cancer.* Jan 2014;21(1):66-74.

36. Hoadley KA, Weigman VJ, Fan C, et al. EGFR associated expression profiles vary with breast tumor subtype. *BMC genomics.* 2007;8:258.

37. Nielsen TO, Hsu FD, Jensen K, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Aug 15 2004;10(16):5367-5374.

38. Rothman KJ. Induction and latent periods. *American journal of epidemiology.* Aug 1981;114(2):253-259.

39. Catsburg C, Miller AB, Rohan TE. Active cigarette smoking and risk of breast cancer. *International Journal of Cancer.* 2014.

40. Cui Y, Miller AB, Rohan TE. Cigarette smoking and breast cancer risk: update of a prospective cohort study. *Breast Cancer Res Treat.* Dec 2006;100(3):293-299.

41. Dossus L, Boutron-Ruault MC, Kaaks R, et al. Active and passive cigarette smoking and breast cancer risk: Results from the EPIC cohort. *International Journal of Cancer.* 2014;134(8):1871-1888.

42. Gaudet MM, Gapstur SM, Sun J, Diver WR, Hannan LM, Thun MJ. Active smoking and breast cancer risk: original cohort data and meta-analysis. *J Natl Cancer Inst.* Apr 17 2013;105(8):515-525.

43. Gram IT, Park SY, Kolonel LN, et al. Smoking and Risk of Breast Cancer in a Racially/Ethnically Diverse Population of Mainly Women Who Do Not Drink Alcohol: The MEC Study. *Am J Epidemiol.* Dec 1 2015;182(11):917-925.

44. Nyante S, Gierach G, Dallal C, et al. Cigarette smoking and postmenopausal breast cancer risk in a prospective cohort. *British journal of cancer.* 2014;110(9):2339-2347.

45. Rosenberg L, Boggs DA, Bethea TN, Wise LA, Adams-Campbell LL, Palmer JR. A prospective study of smoking and breast cancer risk among African-American women. *Cancer causes & control : CCC.* Dec 2013;24(12):2207-2215.

46. Xue F, Willett WC, Rosner BA, Hankinson SE, Michels KB. Cigarette smoking and the incidence of breast cancer. *Arch Intern Med.* Jan 24 2011;171(2):125-133.

47. Butler. Active Smoking and Risk of Luminal and Basal-like Breast Cancer Types in the Carolina Breast Cancer Study. In Press.

48. Kabat GC, Kim M, Phipps AI, et al. Smoking and alcohol consumption in relation to risk of triple-negative breast cancer in a cohort of postmenopausal women. *Cancer causes & control : CCC.* May 2011;22(5):775-783.

49. Kawai M, Malone KE, Tang MT, Li CI. Active smoking and the risk of estrogen receptor-positive and triple-negative breast cancer among women ages 20 to 44 years. *Cancer.* Apr 1 2014;120(7):1026-1034.

50. Tariq K, Farhangi A, Rana F. TNBC vs Non-TNBC: A Retrospective Review of Differences in Mean Age, Family History, Smoking History, and Stage at Diagnosis. *Clinical advances in hematology & oncology: H&O.* 2014;12(6):377-381.

51. Turkoz FP, Solak M, Petekkaya I, et al. Association between common risk factors and molecular subtypes in breast cancer patients. *Breast.* Jun 2013;22(3):344-350.

52. Cooper JA, Rohan TE, Cant EL, Horsfall DJ, Tilley WD. Risk factors for breast cancer by oestrogen receptor status: a population-based case-control study. *Br J Cancer.* Jan 1989;59(1):119-125.

53. Gaudet MM, Gammon MD, Bensen JT, et al. Genetic variation of TP53, polycyclic aromatic hydrocarbon-related exposures, and breast cancer risk among women on Long Island, New York. *Breast Cancer Res Treat.* Mar 2008;108(1):93-99.

54. Manjer J, Malina J, Berglund G, Bondeson L, Garne JP, Janzon L. Smoking associated with hormone receptor negative breast cancer. *International journal of cancer. Journal international du cancer.* Feb 15 2001;91(4):580-584.

55. Morabia A, Bernstein M, Ruiz J, Heritier S, Diebold Berger S, Borisch B. Relation of smoking to breast cancer by estrogen receptor status. *International journal of cancer. Journal international du cancer.* Jan 30 1998;75(3):339-342.

56. Nishino Y, Minami Y, Kawai M, et al. Cigarette smoking and breast cancer risk in relation to joint estrogen and progesterone receptor status: a case-control study in Japan. *SpringerPlus.* 2014;3(1):65.

57. Conway K, Edmiston SN, Cui L, et al. Prevalence and spectrum of p53 mutations associated with smoking in breast cancer. *Cancer Res.* Apr 1 2002;62(7):1987-1995.

58. Furberg H, Millikan RC, Geradts J, et al. Environmental factors in relation to breast cancer characterized by p53 protein expression. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* Sep 2002;11(9):829-835.

59. Gammon MD, Hibshoosh H, Terry MB, et al. Cigarette smoking and other risk factors in relation to p53 expression in breast cancer among young women. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* Mar 1999;8(3):255-263.

60. Mordukhovich I, Rossner P, Jr., Terry MB, et al. Associations between polycyclic aromatic hydrocarbon-related exposures and p53 mutations in breast tumors. *Environmental health perspectives.* Apr 2010;118(4):511-518.

61. Allott EH, Cohen SM, Geradts J, et al. Performance of Three-Biomarker Immunohistochemistry for Intrinsic Breast Cancer Subtyping in the AMBER Consortium. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* Mar 2016;25(3):470-478.

62. Hammond ME, Hayes DF, Wolff AC, Mangu PB, Temin S. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Journal of oncology practice / American Society of Clinical Oncology.* Jul 2010;6(4):195-197.

63. Williams LA. TP53 protein levels, RNA-based pathway assessment, and race among invasive breast cancer cases. *In Preparation.* 2017.

64. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Mar 10 2009;27(8):1160-1167.

65. Troester MA, Herschkowitz JI, Oh DS, et al. Gene expression patterns associated with p53 status in breast cancer. *BMC cancer.* 2006;6:276.

66. Patrick DL, Cheadle A, Thompson DC, Diehr P, Koepsell T, Kinne S. The validity of self-reported smoking: a review and meta-analysis. *American journal of public health.* Jul 1994;84(7):1086-1093.

67. Park SY, Kolonel LN, Lim U, White KK, Henderson BE, Wilkens LR. Alcohol consumption and breast cancer risk among women from five ethnic groups with light to moderate intakes: the Multiethnic Cohort Study. *International journal of cancer. Journal international du cancer.* Mar 15 2014;134(6):1504-1510.

68. Starks AM, Martin DN, Dorsey TH, Boersma BJ, Wallace TA, Ambs S. Household income is associated with the p53 mutation frequency in human breast tumors. *PloS one.* 2013;8(3):e57361.

69. Butler EN, Tse CK, Bell ME, Conway K, Olshan AF, Troester MA. Active smoking and risk of Luminal and Basal-like breast cancer subtypes in the Carolina Breast Cancer Study. *Cancer causes & control : CCC.* Jun 2016;27(6):775-786.

70. Persson M, Simonsson M, Markkula A, Rose C, Ingvar C, Jernstrom H. Impacts of smoking on endocrine treatment response in a prospective breast cancer cohort. *Br J Cancer.* Jul 26 2016;115(3):382-390.

71. Alberg AJ, Shopland DR, Cummings KM. The 2014 Surgeon General's Report: Commemorating the 50th Anniversary of the 1964 Report of the Advisory Committee to the US Surgeon General and Updating the Evidence on the Health Consequences of Cigarette Smoking. *American journal of epidemiology.* Jan 15 2014.

72. Nielsen TO, Parker JS, Leung S, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Nov 01 2010;16(21):5222-5232.

73. Weinberg CR, Sandler DP. Randomized recruitment in case-control studies. *Am J Epidemiol.* Aug 15 1991;134(4):421-432.

74. Waggott DM. NanoStringNorm: Normalize NanoString miRNA and mRNA Data. 2015(R package version 1.1.21).

75. Fraley C, Raftery AE, Murphy TB, Scrucca L. Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. *Technical Report No. 597, Department of Statistics, University of Washington.* 2012(mclust Version 4 for R).

76. Terry PD, Miller AB, Rohan TE. Cigarette smoking and breast cancer risk: a long latency period? *International journal of cancer. Journal international du cancer.* Aug 20 2002;100(6):723-728.

77. Richardson DB. Latency models for analyses of protracted exposures. *Epidemiology.* May 2009;20(3):395-399.

78. Newman B, Moorman PG, Millikan R, et al. The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. *Breast Cancer Res Treat.* Jul 1995;35(1):51-60.

79. Yaziji H, Taylor CR, Goldstein NS, et al. Consensus recommendations on estrogen receptor testing in breast cancer by immunohistochemistry. *Applied immunohistochemistry & molecular morphology : AIMM / official publication of the Society for Applied Immunohistochemistry.* Dec 2008;16(6):513-520.

80. Runnebaum IB, Nagarajan M, Bowman M, Soto D, Sukumar S. Mutations in p53 as potential molecular markers for human breast cancer. *Proceedings of the National Academy of Sciences of the United States of America.* Dec 01 1991;88(23):10657-10661.

81. Alsner J, Jensen V, Kyndi M, et al. A comparison between p53 accumulation determined by immunohistochemistry and TP53 mutations as prognostic variables in tumours from breast cancer patients. *Acta oncologica.* 2008;47(4):600-607.

82. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British journal of cancer.* Mar 1954;8(1):1-12.

83. Armitage P, Doll R. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British journal of cancer.* Jun 1957;11(2):161-169.

84. Hirata N, Sekino Y, Kanda Y. Nicotine increases cancer stem cell population in MCF-7 cells. *Biochemical and biophysical research communications.* Dec 3 2010;403(1):138-143.

85. Lee CH, Chang YC, Chen CS, et al. Crosstalk between nicotine and estrogen-induced estrogen receptor activation induces alpha9-nicotinic acetylcholine receptor expression in human breast cancer cells. *Breast Cancer Res Treat.* Sep 2011;129(2):331-345.

86. Moll UM, Petrenko O. The MDM2-p53 interaction. *Molecular cancer research : MCR.* Dec 2003;1(14):1001-1008.

87. Gammon MD, Santella RM, Neugut AI, et al. Environmental toxins and breast cancer on Long Island. I. Polycyclic aromatic hydrocarbon DNA adducts. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* Aug 2002;11(8):677-685.

88. Morris JJ, Seifter E. The role of aromatic hydrocarbons in the genesis of breast cancer. *Medical hypotheses.* Jul 1992;38(3):177-184.

89. Hung C-S. The alpha9 Nicotinic Acetylcholine Receptor is the Key Mediator in Nicotine-enhanced Cancer Metastasis in Breast Cancer Cells. *Journal of Experimental and Clinical Medicine.* 2011;3(6):283-292.

90. Ng TK, Huang L, Cao D, et al. Cigarette smoking hinders human periodontal ligament-derived stem cell proliferation, migration and differentiation potentials. *Scientific reports.* Jan 16 2015;5:7828.

91. Paixao LL, Gaspar-Reis RP, Gonzalez GP, et al. Cigarette smoke impairs granulosa cell proliferation and oocyte growth after exposure cessation in young Swiss mice: an experimental study. *Journal of ovarian research.* Sep 20 2012;5(1):25.

92. Iwamoto T, Booser D, Valero V, et al. Estrogen receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1% to 10% ER-positive by immunohistochemistry. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Mar 1 2012;30(7):729-734.

93. Allred DC. Issues and updates: evaluating estrogen receptor-alpha, progesterone receptor, and HER2 in breast cancer. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc.* May 2010;23 Suppl 2:S52-59.