January 2011

Optimal Tumor Sampling For Immunostaining Of Biomarkers In Breast Carcinoma

Juliana Tolles

Optimal Tumor Sampling for Immunostaining of Biomarkers in Breast Carcinoma

A Thesis Submitted to the
Yale University School of Medicine
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Medicine

Juliana Tolles

2011

OPTIMAL TUMOR SAMPLING FOR IMMUNOSTAINING OF BIOMARKERS IN BREAST CARCINOMA. Juliana Tolles, Yalai Bai, Maria Baquero, Lyndsay N. Harris, David L. Rimm, Annette M. Molinaro. Division of Biostatistics, Yale University School of Public Health, New Haven, CT.

Biomarkers, such as estrogen receptor, are used to determine therapy and prognosis in breast carcinoma. Immunostaining assays of biomarker expression have a high rate of inaccuracy, for example estimates are as high as 20% for estrogen receptor. Biomarkers have been shown to be heterogeneously expressed in breast tumors and this heterogeneity may contribute to the inaccuracy of immunostaining assays. Currently, no evidence-based standards exist for the amount of tumor that must be sampled in order to correct for biomarker heterogeneity.

The purpose of this study is to determine the optimal number of 20X fields that are necessary to estimate a representative measurement of expression in a whole tissue section for selected biomarkers: estrogen receptor (ER), human epidermal growth factor receptor 2 (HER2), AKT, extracellular signal-regulated kinase (ERK), ribosomal protein S6 kinase 1 (S6K1), glyceraldehyde 3-phosphate dehydrogenase (GAPDH), cytokeratin, and microtubule-associated protein-Tau (MAP-Tau).

Two collections of whole tissue sections of breast carcinoma were immunostained for biomarkers. Expression was quantified using Automated Quantitative Analysis (AQUA). Simulated sampling of various numbers of fields (ranging from $1-35$) was performed for each marker. The optimal number was selected for each marker via resampling techniques and minimization of prediction error over an independent test set.

The optimal number of 20X fields varied by marker, ranging between $3-14$ fields. More heterogeneous markers, such as MAP-Tau, required a larger sample of 20X fields to produce representative measurement. The clinical implication of these findings is that small core needle breast biopsies may be inadequate to represent whole tumor

biomarker expression for many markers. Also, for biomarkers newly introduced into clinical use, especially if therapeutic response is dictated by level of expression, the optimal size of tissue sample must be determined on a marker-by-marker basis.

Acknowledgements

This work was supported by the National Center for Research Resources (Grant TL 1 RR024137 to JT), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research (CTSA Grant UL1 RR024139 to AMM), Doris Duke Charitable Foundation (Grant 2007080 for JT), the Yale University Biomedical High Performance Computing Center and NIH (Grant RR19895 for instrumentation) and the USAMRMC Breast Cancer Research Program (Grant W81XWH-06-1-0746 to MB).

I would like to thank my mentors, Dr. Molinaro and Dr. Rimm, for their guidance, patience, and encouragement. This project would not have been possible without Maria Baquero and Yalai Bai, who performed the data collection for this study. Lastly, I want to thank my family for their love and support.

Contents

Introduction

Biomarkers have become essential for therapeutic decision-making and prognostication in breast carcinoma. Estrogen Receptor (ER) is the prototypical biomarker for this cancer; as early as the 1970s, investigations suggested that ER was an independent predictor of both breast carcinomas' response to therapy and the likelihood of tumor recurrence. In 1974, a workshop convened by the Breast Cancer Task Force of the National Cancer Institute reviewed the results of 436 treatment trials in 380 patients in an effort to determine whether assays for ER expression in breast carcinoma could predict clinical response to hormonal therapies. Hormonal therapies were not in widespread use at the time and included surgical ablation of estrogen-producing organs, anti-estrogens, estrogens, glucocorticoids, and androgens [1]. The committee found that 55-60% of the patients with tumors that tested positive for ER responded to hormonal therapy (response was defined as a minimum 50% reduction in size of at least 50% of tumors), whereas only 8% of patients with ER-negative tumors responded. Shortly afterward, Knight et al. found that ER-negative tumors were associated with a higher rate of metastasis and lower rate of overall survival in a cohort of 145 cases [2]. This effect was independent of axillary node status, tumor size, and tumor location. Knight et al. did not control the effect of adjuvant hormonal therapies in this study and, they acknowledged that the differences in survival between ER-positive and ER-negative subjects might be explained by differences in therapeutic response.

These findings motivated a series of large randomized, controlled clinical trials of hormonal therapies, the results of which justified a change in the standard of care to include ER testing for all breast carcinomas. The first of these trials, the National Surgical Adjuvant Breast and Bowel Project, began in 1977. It included over 1800 subjects with breast carcinoma at 68 institutions. It found that the addition of tamoxifen, an ER antagonist in breast tissue, to the standard chemotherapy regimen of L-phenylalanine mustard and 5-fluorouracil improved disease-free survival and survival for patients with node-positive cancers that expressed ER [3, 4]. ER positivity was defined as an ER protein level above a threshold of 10 fmol as measured by ligand-binding assays (LBAs; described further below).

In 1989, a randomized, controlled trial involving over 2600 subjects, demonstrated that hormonal therapy with tamoxifen increased disease-free survival in patients with node-negative, ER-positive tumors [5]. Decreases in local recurrence, tumors of the opposite breast, and treatment failure at metastatic sites all contributed to this result. The authors concluded that tamoxifen therapy was justified in all subjects who met inclusion criteria for the study: women under the age of 70, with operable tumors expressing ER levels $\geq$ 10 fmol, who met the set of common National Surgical Adjuvant Breast and Bowel Project inclusion criteria. Stated more generally, the clinical implication of this study was that ER status should inform the choice of therapy for all patients with breast carcinoma, regardless of the presence or absence of lymph nodes positive for carcinoma.

Today, there are two broad classes of hormonal therapy. The first class is selective estrogen-receptor modulators (SERMs) such as tamoxifen, which, depending on the tissue type, have either agonistic or antagonistic effects on ER. Although tamoxifen is the most widely used compound in this class, SERMs include other drugs, such as fulvestrant. The second class consists of aromatase inhibitors (AIs), such as anastrozole, which block conversion of adrenally-produced estrogen precursors into estrogen [6]. Other methods of hormonal therapy, such as the surgical ablation of estrogen-producing organs, are not in widespread use. Importantly, a meta-analysis of 78 randomized clinical trials involving over 42,000 patients found that hormonal therapies do not increase survival time for patients with ER-negative tumors, suggesting that the indiscriminate treatment of all breast carcinomas with hormonal therapy is inadvisable [7].

In the last two decades, methods for detecting biomarker expression have benefitted from technological improvements. In 1999, Harvey et al. demonstrated that an immunohisochemichal assay for ER, which employed a mouse monoclonal antibody directed against the epitope, predicted disease-free survival with greater accuracy than ligand-binding assays (LBAs) [8]. Additionally, the immunohistochemical assay had several technical advantages over LBAs. LBAs require large quantities of fresh-frozen tissue, whereas immunohistochemistry can be applied to formalin-fixed paraffin-imbedded specimens. LBAs also require homogenization of tissue, rendering it impossible to determine the relative composition of tumor and benign cells in the isolate; immunohistochemistry can be performed on histologically intact tissue, allowing distinguishing morphological features to be left intact. Thus, immunohistochemistry has become the standard assay for measuring breast tumor expression of ER.

Although ER was the first biomarker used to guide the management of breast carcinoma, the measurement of several other markers has become part of the standard of care for this disease. In 1983, Clark et al. found, in a study of 189 women receiving adjuvant therapy for breast carcinoma, that positive staining of tumors for progesterone receptor (PR) predicted increased length of disease-free survival. The analysis demonstrated that this effect was independent of ER expression and of the type of adjuvant therapy used (regimens including hormonal therapy vs. those without hormonal therapy) [9]. These results were confirmed by subsequent studies and, like ER, immunohistochemical assays became the preferred method of detection for PR [6].

The next pivotal marker for breast carcinoma, HER2, was discovered in the 1990s. A case-control study conducted by Press et al. of 210 women with node-negative breast carcinoma found that tumors' overexpression of the cell surface receptor HER2 predicted likelihood of cancer recurrence [10]. Much like ER, HER2 was first detected by immunohistochemistry applied to formalin-fixed sections; however, in this study, expression levels were also quantified by computer image analysis of immunohistochemically stained tissue sections. Interestingly, a dose-dependent effect was uncovered: subjects with any of level of HER2 overexpression were 3 times as likely to have a cancer recurrence, whereas those with "high" levels of overexpression were 9.5 times as likely to have a cancer recurrence. Later studies confirmed this finding and validated *in situ* hybridization as a alternative technique for detecting HER2 overexpression [11]. Subsequent work found that HER2 positivity predicts a lesser likelihood of response to hormonal therapies, non-anthracycline agents, and non-taxane agents. Mostly importantly, the presence of HER2 in a breast carcinoma predicts a greater likelihood of response to trastuzumab, a humanized monoclonal antibody targeted against HER2. Trastuzumab has been shown to improve survival in both metastatic and early-stage breast cancer [12, 13].

Taken together, these findings had significant implications for the utility of HER2 measurement in the management of breast carcinoma. An immunohistochemical assay for HER2 expression received FDA approval in 1998 and, in 2001, the ASCO/CAP committee recommended HER2 testing as the standard of care for all newly diagnosed and metastatic breast carcinoma [14, 15]. The most commonly used clinical algorithm employs immunohistochemistry to "screen" specimens and reflex fluorescent *in situ* hybridization testing of high-scoring cases to confirm results [14].

Commercial assays and academic investigations have moved beyond the use of individual biomarkers to the development of biomarker "signatures" to inform prognosis and therapeutic decision-making. Oncotype DX$^{\text{TM}}$ is 21-gene RT-PCR assay that measures markers such as Ki67, HER2 family members, and matrix metalloproteases in order to stratify ER-positive tumors into "low risk," "intermediate risk," and "high risk" groups with predicted recurrence rates of 7%, 14%, and 31% respectively [16]. In a parallel effort, a recent study of biomarkers in ductal carcinoma *in* *situ* demonstrated that patients with these precancerous lesions could be stratified into groups with statistically significant differences in risk for progression to invasive breast carcinoma. This stratification is based on the expression signature of the following biomarkers: ER, PR, Ki67, p53, p-16, HER2, and cyclooxygenase-2. In that study, the markers were detected by immunohistochemistry [17].

Many other putative biomarkers of prognosis and therapeutic response are currently in various stages of pre-clinical investigation [16]. Overexpression of cell cycle markers, such as cyclin D1 and cyclin E, have been linked to decreased survival times for patients with breast carcinoma [18, 19, 20]. The H-ras oncogene has been shown in several studies to be predictive of breast carcinoma progression [21, 22]. Some evidence suggests that loss of p53 expression in breast carcinoma is predictive of resistance to hormonal and adjuvant therapies [23, 24]. Overexpression of certain matrix metalloproteases, believed to be involved in tumor invasion and metastasis, has been associated with poorer clinical outcomes [25, 26, 27]. Although none of these markers are currently FDA-approved (or recommended for clinical use by the most recent ASCO/CAP review of biomarkers in breast cancer [28]), the large number of promising pre-clinical studies suggests that new markers will become part of the standard of clinical care in coming years. Of note, the vast majority of these markers are detected by immunohistochemical methods.

It is therefore concerning that conventional assays for ER and other biomarkers suffer from lack of objective methods of measurement. Immunohistochemical assays have become the standard of care for determining ER and PR status, but the most recent American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) committee review of immunohistochemical assays for breast carcinoma estimated that "up to 20% of ER and PR determinations worldwide may be inaccurate (false negative or false positive)" [6]. A separate ASCO/CAP committee determined that approximately the same percentage of HER2 assays for breast carcinoma were inaccurate, noting that neither the immunohistochemical assay nor the *in situ* hybridization assay for HER2 demonstrated a lower rate of error [14].

National quality assurance audits conducted in the UK and Australia each identified significant variation in rates of ER- and PR-positivity in laboratories across those countries [29, 30]. In Canada, where government health services are provided independently by the provincial governments, it was discovered that, based upon retesting in a central Ontario laboratory, false negative results had been reported in approximately 33% of 1,023 samples that underwent ER assays in Newfoundland laboratories [6]. More than 100 patients in this group died; a subsequent class action lawsuit was filed against the provincial health service for negligence in ER testing.

In Asian countries – such as the Philippines, Vietnam, and Malaysia – a significant rise in the rate of ER-positive breast cancer cases was reported after more stringent standards for methods of conducting ER assays were introduced [6, 31]. The ASCO/CAP committee cited all of the above findings to support its development of a "guideline to improve the accuracy of immunohistochemical estrogen receptor and progesterone receptor testing in breast cancer and the utility of these receptors as predictive markers." The committee hypothesized that misclassifications of ER and PR status were due to a number of factors, which it grouped into three categories: pre-analytic variables, thresholds for positivity, and interpretation criteria.

Pre-analytical variables are variations in events that occur prior to immunohistochemical assays, such as length of cold ischemic time, duration of fixation, and fixative type. In order to reduce the contributions of pre-analytical variables to assay variability, the committee recommended that pathologists minimize the time from specimen acquisition to fixation, section specimens at 5 mm intervals to promote penetration by the fixative solution, use 10% neutral buffered formalin as a fixative, and limit fixation time to a range between $6 - 72$ hours. Based on its review of studies linking patient outcomes to percentage of positive-staining cells, the committee lowered the threshold of positivity to a minimum of 1% positive-staining tumor cells from the previous recommendation of 10% for ER and PR. Lastly, in order to address variability in interpretation criteria, the ASCO/CAP committee made a variety of recommendations regarding the use of internal and external controls for immunohistochemical assays, voluntary participation in competency training for pathologists, and standardization of the reports of assay results.

An earlier committee, convened in 2007, reached similar conclusions about the sources of inaccuracy in HER2 testing algorithms: it cited pre-analytical variation, variability in assay reagents, and inadequate pathologist training [14]. The committee made recommendations for reducing sources of error and variability in the assays for HER2 that parallel those made by the committee on ER and PR. It also recommended standardized thresholds for positivity for both immunohistochemical assays ($> 30\%$ positive-staining cells) and *in situ* hybridization assays for HER2 (> 6 HER2 gene copies per nucleus or a fluorescent *in situ* hybridization ratio > 2.2).

However, an additional important possible cause of the high rate of immunohistochemical assay inaccuracy for all biomarkers, given little attention in the ASCO/CAP reports, is biomarker heterogeneity [32]. Biomarkers are known to be heterogeneously expressed in breast carcinoma. Biomarker heterogeneity likely stems from numerous etiologies, including both intrinsic biological causes and variations in specimen handling. Hypotheses for biological sources variation include the inherent DNA instability in malignant cells, which could generate genetic or epigenetic changes with successive cell divisions; differences in the tumor microenvironment, such as availability of local blood supply; and the “cancer stem cell” hypothesis, which holds that a subpopulation of stem cells produce a variety of tumor cells via a perturbed differentiation process [33]. Pre-analytical variables, such as slow formalin penetration of thick sections of tumor tissue, could also produce heterogeneity if some epitopes undergo proteolytic degradation prior to formalin fixation (Bai et al., in preparation).

Because of the importance of ER in determining therapy for breast carcinoma, many investigations have examined intra-tumor heterogeneity of ER expression, and three of the largest are discussed in detail below. Although each study employed different expression metrics of heterogeneity and methods of statistical analysis, all found statistically significant differences in ER expression between different regions of tumor from the same subject.

The work of Meyer et al., in 1991, represented one the earliest attempts to characterize ER heterogeneity and its potential contribution to the inaccurate assignment of hormone-receptor status in patients undergoing work-up for breast carcinoma [34]. A cohort of 65 tumors were sampled at 5mm intervals, with a maximum of 8 samples per tumor. Cytosolic preparations from each sample were processed with a LBA assay to determine its ER and PR status. The measured concentrations of ER and PR were divided into 4 ranges: 2, 10, 50, and 500 fmol/mg respectively (the ranges were selected arbitrarily, not based on their clinical significance). A tumor was considered to have heterogeneous expression of the marker if any two samples from the tumor had scores from non-contiguous ranges. The study found that 24% of tumors were heterogeneous for ER.

Chung et al. revisited the heterogeneity question using an immunohistochemical assay for ER, which became the standard of care for detecting ER and PR in the 1990s [6]. They measured ER expression in samples from 11 patients with breast carcinoma using quantitative immunofluorescence. For each patient, scores were measured in different "blocks" of tissue from the same tumor, with each block represented by a single 1 cm x 1 cm x 5 μm "whole tissue" slide [35]. The differences between scores of "blocks" from the same tumor was found to be statistically significant ($p<0.05$) in 78% of cases. Additionally, the study illustrated heterogeneity between regions of tumor from the same slide; many slides contained clusters of fields with either high-intensity or low-intensity staining.

Nassar et al. demonstrated heterogeneous intra-tumor expression of ER in a slightly larger cohort, using both subjective and objective metrics of ER expression. They constructed TMAs consisting of 44 cases of breast carcinoma, encompassing a variety of carcinoma subtypes, and five controls of normal breast tissue [36]. Each case was represented by three 1 mm cores from three distinct tumor foci (total of nine cores per case). ER expression was quantified both on an ordinal scale of staining intensity (0 to 3+) and as a percentage of positive-staining cells (0% to 100%) using subjective visual scoring by light microscopy. These scales were converted into a binary measure of "ER negative" and "ER positive": specimens were considered "negative" for ER if they had a staining intensity of 0 and no more than 10% positive-staining tumor cells. ER expression was also quantified by Automated Cellular Imaging System (ACIS; Dako). For the ACIS scoring, scores from the three TMA spots for each tumor focus were averaged, producing three scores per case.

Biomarker expression was quantified in this investigation using an "intraclass correlation coefficient." Briefly, the coefficient compares intra-tumor heterogeneity to the overall variance; a coefficient greater than 0.75 is considered to indicate low heterogeneity. The intraclass correlation coefficient for ER was less than 0.75 for all metrics – staining intensity, percentage of positive-staining cells, and binary score – both when measured visually and by ACIS.

Biomarkers other than ER have been shown to be heterogeneously expressed in breast carcinoma. Markers PR, HER2, p53, and MIB-1 have been shown to have statistically significant differences in intra-tumor expression. Nassar et al., using the same methods as those described for ER above, demonstrated statistically significant heterogeneity of expression for p53, MIB-1, and HER2 in breast carcinoma [36]. Kallioniemi et al. also reported heterogeneity in HER2 overexpression, having identified subpopulations with different degrees of gene amplification by fluorescent *in situ* hybridization [37]. In parallel with their analysis of ER, Meyer et al. demon-

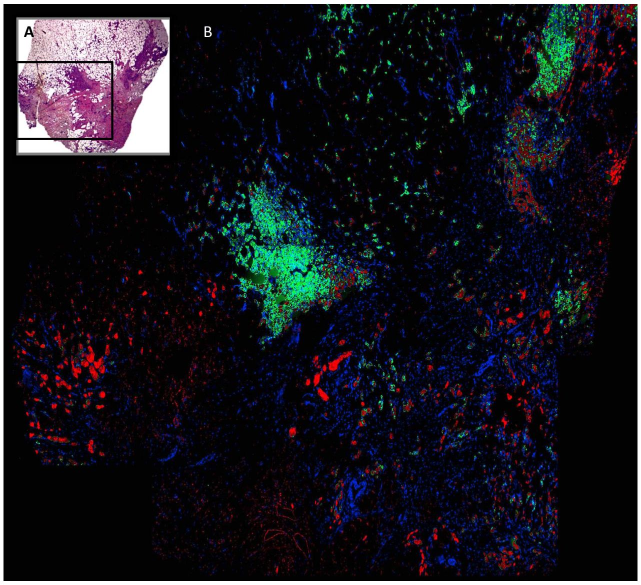

Figure 1: Heterogeneity of MAP-Tau expression in a whole tissue section of breast carcinoma. (A) H&E stain (B) Immunofluorescence. Nuclei are labeled with DAPI. Cytokeratin is labeled with Cy3. MAP-Tau is labeled with Cy5.

strated heterogeneity for PR in breast carcinoma, finding that 20% of tumors were heterogeneous for the marker [34]. It is likely that many other epitopes are heterogeneously expressed in tumors; the heterogeneity of MAP-Tau epitope can be visualized in immunostained whole tissue sections (Figure 1).

All of the above studies rely upon some form of immunohistochemistry for biomarker detection, but biomarker heterogeneity has also been demonstrated via the real-time polymerase chain reaction (RT-PCR) for a variety of biomarkers in a variety of cancers. Lassman et al. measured heterogeneity of CK20 expression in colorectal carcinoma using both immunohistochemistry and quantitative RT-PCR [38]. They first identified tissue with heterogeneous CK20 expression by immunohistochemistical staining, using visual scoring criteria. They then performed quantitative RT-PCR on the same tissue, finding an average 3.8-fold difference in CK20 mRNA expression between cells classified as “weak” staining and those classified as “strong” staining by immunohistochemistry. Sigalotti et al. demonstrated biomarker heterogeneity for cancer/testis antigens (MAGE, NYESO, and SSX gene families) in melanoma using qualitative RT-PCR [39]. They found different levels of expression for several biomarkers in single cell isolates cultured from the same melanoma lesion. These studies confirm that biomarker heterogeneity is demonstrable at the mRNA level and is not an artifact of immunohistochemical techniques.

A study of several biomarkers – HER2, epidermal growth factor receptor (EGFR), Bcl-2, p53, and proliferating cell nuclear antigen (PCNA) – conducted by Chhieng et al., produced results that appear to contradict the above studies [40]. The group examined 30 breast carcinoma tumors, each a minimum of 1 cm in diameter and 19 of which had a ductal carcinoma *in situ* component. It found that immunohistochemical assays for expression of this group of biomarkers in breast carcinoma exhibited only "minor regional variations" in EGFR and p53 expression, when those markers were measured in the invasive ductal carcinoma component, and in PCNA, when it was measured in ductal carcinoma *in situ* component.

The precise methods employed in this study are critical to the interpretation of its findings. For each case, serial 5 μm sections of tumor were immunohistochemically stained for one marker each. One section was also stained with hematoxylin and eosin (H&E). The (H&E) section was then divided into four "randomly oriented discrete regions," such that each region contained a portion of the invasive cancer. Each corresponding immunohistochemically stained serial section was divided along these identical lines. Expression of markers was quantified on by a value on an ordinal scale (0-4), calculated from the combination of a visual estimate of the percentage of positive-staining cells and a visual estimate of staining intensity on an ordinal scale.

For the statistical analysis, scores for each of the four regions and whole slides were then grouped over all 30 cases (although there was no relationship between region 1 case 1 and region 1 from other cases), with distinctions made between regions of invasive cancer and regions of ductal carcinoma *in situ*. Whole slide scores for a single slide were compared to scores from each discrete region by Wilcoxon Signed

Rank Test. Consistency between the four regional scores was assessed via Spearman Coefficient. The authors found significant differences between regional and whole slide scores of EGFR, p53 and PCNA.

Although, Chhieng et al. observed only a few significant differences between the immunohistochemical scores of tumor regions and whole slides, these results do not necessarily contradict the findings of previous studies. The authors do not report the size of the "discrete regions" analyzed, but, given the inclusion criteria minimum tumor diameter of 1 cm, it is likely that the regions were much larger than the 1 mm-diameter cores analyzed by Nassar et al. in their study of HER2 and p53. This suggests that there may be some definable minimum size of tumor sample required to represent biomarker expression in the whole tissue slide.

Furthermore, Chhieng et al. did not use the objective, continuous scoring for biomarker expression, such as the AQUA system employed by Chung et al.; an ordinal scoring scale has less statistical power to detect small differences in expression levels. Taken together with the previously described investigations of biomarker heterogeneity, this study suggests that there is a need to define a minimum size of representative tumor section using an objective, continuous scoring system for immunohistochemical assays.

Chhieng et al. also examined differences in the percentage of positive-staining cells between regions of invasive carcinoma on a case-by-case basis. Differences in the degree of heterogeneity were detected: up to 66% absolute difference in percentage staining for HER2 (membranous), 50% for HER2 (cytoplasmic), 50% for EGFR (membranous), 53% for EGFR (cytoplasmic), 113% for Bcl-2, 66% for p53, 62% and for PCNA. This suggests that different markers may have different degrees of heterogeneity and that any investigation seeking to define minimum sampling should do so on a marker-by-marker basis.

Given the overwhelming evidence of biomarker heterogeneity in breast carcinoma, it is plausible that insufficient tumor sampling in the clinical setting may lead to misclassification of biomarker status and inappropriate treatment of patients. Despite the extensive description of the phenomenon of biomarker heterogeneity in breast carcinoma, no evidence-based standards have been developed for the size of tissue sample necessary to correct for heterogeneity in assays of biomarker status. Core needle biopsies, which represent a very small percentage of the entire tumor tissue, are used for biomarker testing in many clinical pathology laboratories. It is possible that these small samples are inadequate in some fraction of cases. Although, the 2010 ASCO/CAP recommended that "large, preferably multiple core biopsies of tumor are preferred for testing if they are representative of the tumor (grade and type) at resection" [6], it gave no additional specific guidance regarding the minimum acceptable sample size. The committee could not offer a more precise recommendation because, to our knowledge, no prior investigations point to a precise standard for the minimum number of cores or sections of resection tissue required to account for biomarker heterogeneity in determining the biomarker expression of breast carcinoma tumors.

Statement of Purpose and Specific Aims

The goal of this study is to estimate the degree of sampling required to make an accurate assessment of biomarker status for the following 7 biomarkers in breast carcinoma:estrogen receptor (ER), human epidermal growth factor receptor 2 (HER2), AKT, extracellular signal-regulated kinase (ERK), ribosomal protein S6 kinase 1 (S6K1), glyceraldehyde 3-phosphate dehydrogenase (GAPDH), cytokeratin, and microtubule-associated protein-Tau (MAP-Tau). We expected these markers, based on knowledge of their biological roles, to represent a range from relatively homogeneous to relatively heterogeneous. GAPDH, a ubiquitously expressed "housekeeping" gene, and cytokeratin, a structural protein present in all epithelial cells, we predicted to be expressed relatively homogeneously within tumors. We expected ER and microtubule-associated protein-Tau (MAP-Tau), based on previous studies and visualization with immunofluorescence, to be more heterogeneous. The specific hypothesis is that markers with greater heterogeneity will require a larger number of sampled fields to produce a representative measurement.

Specific Aims:

1. Quantify the degree of heterogeneity for each marker using mixed-effects modeling.

2. Simulate sampling different amounts of tumor in order to determine the optimal number of 20X fields required to give a measurement of biomarker expression representative of the entire tissue sample.

Methods

Author Contributions

JT performed the statistical analyses detailed in **Methods: Statistical Methods**. YB and MB carried out the preparation of tissue, AQUA assays, and collection of data. LNH was responsible for tissue acquisition from TAX 307 cohort. DLM conceived of the study and participated in its design. AMM designed the statistical analyses and participated in the study design.

Cohorts

The first collection of subjects consisted of 14 tumor resection specimens from patients who underwent surgery at Yale University/New Haven Hospital between 2001 to 2005. Whole tissue sections of formalin-fixed, paraffin-embedded primary invasive breast cancer tumors were obtained from the archives of the Pathology Department of Yale

University. All the patients were diagnosed with infiltrating ductal carcinoma of the breast. None received chemotherapy or radiation prior to resection. The study was approved by the institutional review board for Yale University.

The second collection of subjects was a cohort (n = 122) from TAX 307, a prospectively collected, independent phase III clinical trial comparing docetaxel-doxorubicin-cyclophosphamide (TAC) versus 5-fluorouracil- doxorubicin-cyclophosphamide (FAC). Patients were enrolled between January 1, 1998 and December 31, 1999, with a total of 484 patients randomized to receive either FAC (75/50/500 mg/m2) or TAC (500/50/500 mg/m2) as first line chemotherapy for metastatic breast cancer. All patients provided clinical consent prior to enrollment. Specimens and associated clinical information were collected under the guidelines and approval of the Dana Farber Human Investigation Committee under protocol #8219 to L.H.

Antibodies and Immunohistochemistry

The TAX 307 clinical trial cohort consisted of 122 whole section slides. Five μm tissue sections from formalin-fixed paraffin-embedded tumor blocks were mounted on aminosilane glass slides (plus slides) and heated. Slides were immunostained using MAP-Tau monoclonal antibody which recognizes all human MAP-Tau isoforms independent of phosphorylation status (1:750; mouse monoclonal, clone 2B2.100/T1029, US Biological, Swampscott, MA). Slides were divided into six individual batches, each including one Breast Cancer Cell Line Control TMA slide. TAX 307 slides were incubated for 24 hours at 60°C. Slides were deparaffinized by oven incubation at 60°C for 20 minutes, followed by two 20 minute incubations in xylene. After slides were washed twice in 100% ethanol, once in 70% ethanol, and rehydrated with tap water, antigen retrieval by pressure cooking was performed in 6.5 mM sodium citrate buffer (pH 6.0) for 10 minutes. Endogenous peroxidase activity was quenched in methanol and 3% hydrogen peroxide for 30 minutes followed by rinsing in tap water and placement in 1X trisethanolamine-buffered saline (TBS; pH 8.0). Non-specific binding was reduced using a 30 minute preincubation in 0.3% bovine serum albumin (BSA) in 0.1M tris-buffered saline (TBS, pH=8) with 0.05% Tween (TBS-T).

Slides were prepared for 4°C overnight incubation (12 hours) by adding a cocktail of MAP-Tau primary antibody (1:750) plus a wide-spectrum rabbit anti-cow cytokeratin antibody (Z0622; Dako, Carpinteria, CA) diluted 1:100 in BSA/1X TBS-T. Following overnight incubation, slides were washed twice in 1X TBS with 0.05% Tween for 10 minutes and once in 1X TBS. Secondary antibody was then applied for 1 hour at room temperature. Goat antirabbit Alexa 488 (Molecular Probes, Eugene OR) was diluted 1:100 in horseradish peroxidase-conjugated EnVision antimouse secondary antibody (Dako). Following incubation with secondary antibodies, slides were washed twice (10minutes, then 5minutes) in 1xTBS-T and once (5 minutes) in 1xTBS. Cyanine-5 (Cy5) directly conjugated to tyramide (FP1117, Perkin-Elmer, Boston MA), diluted 1:50 in amplification diluent (Perkin-Elmer) was used as the fluorescent chromagen for target detection and was added to all slides for 10 minutes at room temperature. Two final washes (10minutes, then 5minutes) in 1X TBS-T and one 5 minute wash in 1X TBS were performed. Slides were stained for double-stranded DNA using Prolong Gold mounting medium with anti-fade reagent 4',6-diamidino-2-phenylindole ("DAPI", Molecular Probes, Eugene OR). Normal breast epithelium served as internal positive controls while omission of the primary antibody served as the negative control for each immunostaining event.

For all epitopes other than MAP-Tau, immunostaining was performed on sets of serial slides from the first collection of subjects (n=14) and the following protocol was used for MAP-Tau. Whole tissue sections were incubated at 60°C for 20 minutes before being deparaffinized with xylene, rehydrated, endogenous peroxidase blocked, and antigen-retrieved by pressure cooking for 15 min in citrate buffer (pH = 6). Slides were pre-incubated with 0.3% bovine serum albumin in 0.1 mol/L TBS (pH =

Protein	Species	Clone	Dilutions	Supplier
ER	Mouse mAb	1D5	1:50	Dako
HER2	Rabbit pAb	A0485	1:2000	Dako
AKT	Rabbit mAb	11E7	1:1000	CST
ERK1/2	Mouse mAb	L34F12	1:1000	CST
S6K1	Rabbit mAb	49D7	1:450	CST
GAPDH	Rabbit mAb	14C10	1:500	CST

Table 1: Antibodies, epitopes, sources and dilutions

8) for 30 min at room temperature. The procedure for pAKT staining was a follows: slides were incubated with a cocktail of ERK1/2 antibody diluted at 1:1000 (Mouse monoclonal, clone L34F12; Cell Signaling Technology, Danvers, MA) and a wide-spectrum rabbit anti-cow cytokeratin antibody (Z0622; Dako Corp, Carpinteria, CA), diluted 1:100 in bovine serum albumin/TBS overnight at 4°C. This was followed by a 1-hour incubation at room temperature with Alexa 546-conjugated goat anti-rabbit secondary antibody (A11010; Molecular Probes, Eugene, OR) diluted 1:100 in mouse EnVision reagent (K4001, Dako Corp, Carpinteria, CA). Cyanine 5 (Cy5) directly conjugated to tyramide (FP1117; Perkin-Elmer, Boston, MA) at a 1:50 dilution was used as the fluorescent chromogen for pAKT detection. Prolong mounting medium (Prolong Gold, P36931; Molecular Probes, Eugene, OR) containing 4',6-diamidino-2-phenylindole was used to identify tissue nuclei. Immunostaining for all remaining epitopes was done in a similar manner with antibodies as follows outlined in Table 1.

Image Capture and Analysis

Automated Quantitative Analysis (AQUA) allows exact measurement of protein concentration within subcellular compartments, as described in detail elsewhere [41]. In brief, a series of high-resolution monochromatic images were captured by the PM-2000 microscope (HistoRx). For whole tissue sections, multiple regions of interest (ROIs) containing invasive tumor were circled on the AQUA system screen based on the low-resolution cytokeratin (cytoplasm) image of the immunohistochemically stained slide taken with the AQUA system. The selected ROIs were automatically overlaid with a grid by the image capturing program and each 20X field of view (FOV) was defined automatically.

For each FOV, in-focus and out-of-focus images were obtained using the signal from the 4',6-diamidino-2-phenylindole, cytokeratin-Alexa 546 and target protein-Cy5 channel. Target protein antigenicity was measured using a channel with emission maxima above 620 nm, in order to minimize tissue autofluorescence. Tumor was distinguished from stromal and non-stromal elements by creating an epithelial tumor "mask" from the cytokeratin signal. The binary mask – in which each pixel is either "on" or "off" – is created on the basis of an intensity threshold set by visual inspection of FOVs.

The AQUA score of the target protein in each subcellular compartment was calculated by dividing the target protein compartment pixel intensities by the area of the compartment within which they were measured. AQUA scores were normalized to the exposure time and bit depth at which the images were captured; thus, scores collected at different exposure times are directly comparable.

Statistical Methods

Normalization

Similar to other methods for quantifying fluorescent signals, AQUA scores are subject to some variation between analyses performed at different times. Potential sources of variation, such as buffer lot and microscope bulb hours, are numerous and impossible to completely eliminate. We therefore normalized AQUA scores between analyses performed at different times.

All epitopes with the exception of MAP-Tau and ER were processed in a single AQUA run and therefore did not require normalization. MAP-Tau and ER were processed in combination with standardized index TMAs, which consisted of a sample of tissue from breast carcinoma cases and cell lines. To normalize scores of the experimental subjects for MAP-Tau and ER, quantile normalization was first performed on the index TMA [42]. The normalization was performed separately for each epitope. The algorithm for quantile normalization is as follows:

1. Build a p x n matrix $\mathbf{X}$ with observations p in rows and different AQUA processing runs of the index TMA n in columns.
2. Sort values in descending order within array $\mathbf{X}$ columns to create $\mathbf{X}_{sort}$.
3. Replace each value in rows of $\mathbf{X}_{sort}$ with the mean value of $\mathbf{X}_{sort}$.
4. Get $\mathbf{X}_{Normalized}$ by rearranging each column in $\mathbf{X}_{sort}$ to have the same ordering as the original $\mathbf{X}$.

The quantile normalization algorithm assumes that the two sets of data to be normalized are identically distributed [42]. Given that the un-normalized data in this case consists of two repeated measurements of an identical index TMA, processed under the same protocol, the assumption holds true for this data set.

Next, smoothing splines S_j were fit to describe the transformation between each column j in the original matrix of index TMA scores and the corresponding column in the normalized matrix:

$$S_j(\mathbf{X}_j) = \mathbf{X}_{Normalized_j}$$

Smoothing splines are functions defined by locally fit third-degree polynomials, which are constrained by a smoothing parameter to produce a continuous function over the range of the whole data set [43].

A single "baseline" column i was selected from the matrix $\mathbf{X}$. In the last step of normalization, the spline transformation from each index matrix column S_j $(j \neq i)$ was applied to the scores of cases processed with that array, followed by the application of the inverse of the spline function for the baseline array. This transformed the scores to the scale of the "baseline" run. Thus, for the matrix of cases $\mathbf{X}_c$, the final transformation applied was:

$$S_i^{-1}(S_j(\mathbf{X}_{c_j})) \text{ for } j \neq i$$

This normalization method has been validated on several independent cohorts for breast carcinoma (Tolles et al., in preparation). It is one of many possible normalization algorithms that could be employed for our data.

Linear Mixed-Effects Modeling

A linear mixed-effects model is a type of linear model that incorporates both *fixed effects*, which are associated with a population or predictable levels of experimental factors, and *random effects*, which are associated random variation among individuals within the population [44]. Mixed effects models are used to characterize relationships between a response variable (AQUA score) and covariates in the data grouped according to one or more classification factors. In this study the classification factors are the subject and ROI within a given subject. Parameter coefficients are calculated by restricted maximum likelihood estimation.

Linear mixed-effects modeling makes several assumptions about the underlying structure of the data. First, it assumes that within-group errors are independent, identically normally distributed and independent of the random effects. Second, it assumes that the random effects are normally distributed and independent for different groups. These assumptions were verified for our data by the use of quantile-quantile plots of both the residuals and the random effects. A quantile-quantile plot plots the quantiles of the observed data against the predicted quantiles of a normal distribution. If the resulting plot is linear, the observed data are judged to be normally distributed.

Mixed Effects models were fit for each epitope of interest. The form of the model was:

$$y_{ijk} = \beta + b_i + b_j + \epsilon$$

where y_{ijk} is the AQUA Score of the i^{th} subject, in the j^{th} ROI, at the k^{th} FOV. β is the intercept term and ϵ is the residual. The model assumes $b_i \sim N(0, \sigma_1^2)$ and $b_j \sim N(0, \sigma_2^2)$. The interpretation of the model is that σ_1^2 represents the variance between AQUA scores of individual subjects and that σ_2^2 represents variance between AQUA scores of regions within a sample from a subject.

In order to quantify the degree of heterogeneity with a metric that would be comparable across epitopes, we calculated the coefficient of variation for each epitope. Generally the coefficient of variation is defined to be the ratio of the standard deviation of a distribution to the mean of that distribution. Thus, the coefficient of variation for the study was calculated as $\frac{\widehat{\sigma_2}}{\beta_0}$.

The `R` Language and Environment for Statistical Computing and `NLME` package were used for all computations [44].

Sampling Simulation: Model Selection and Cross-Validation

Due to the inherent differences in the two cohorts, the analyses of the biomarkers differed slightly. However, in both, to choose the optimal number of fields (i.e. model selection) and estimate the corresponding prediction error we used two layers of resampling [45, 46]. The first, or outer, layer was for estimating prediction error and the second, or inner, layer for model selection (see Figure 2).

For the MAP-Tau cohort, we employed 10-fold cross-validation for the first layer and Monte Carlo cross-validation for the second [47]. In the first layer the cohort was divided equally into ten groups. For each iteration, one of the groups served as an independent test set for calculation of prediction error while the other nine groups (i.e. 90% of the subjects) constituted the training set. In the second layer, this training set was subdivided into a learning set (90% of training set) and an evaluation set (10% of training set), for the purposes of selecting the optimal number of 20X FOVs. For each of the total 10 training sets, the learning and evaluation sets were both reconstituted 1000 times. A linear regression model was fit to the subjects in the learning set. The corresponding independent variable was the average AQUA Score of a subset of 20X FOVs sampled from each whole tissue slide, and the dependent variable was the overall average score for all FOVs on that slide. A separate regression was calculated for each potential number of FOVs $(1-35)$. Using the coefficients estimated from the regression model developed on the learning set, a predicted score was calculated for each subject in the evaluation set for every number of FOVs. The prediction error (PE) was calculated as follows for each number of FOVs and then averaged over the 1000 evaluation sets:

$$\mathrm{PE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - \bar{x}_i)^2, \tag{1}$$

where $N = \#$ of subjects, $\bar{x}_i = \frac{1}{K}\sum_{j=1}^{K} x_j$, and $K = \#$ of fields in subject i. The first local minimum of the average prediction error was recorded.

Lastly, the mean PE for the independent test sets was calculated by averaging the PE over the 10 independent test sets for each potential number of FOVs $(1-35)$. The average first local minimum and standard error for the test set PE was recorded. In accordance with rules of parsimonious model selection [48], if there existed a model (here, a model is the number of FOVs) with mean PE within one standard error of that of the minimum model, the smaller model was selected as optimal. The entire process was repeated 100 times and the result averaged to produce a stabile estimate of the optimal number of FOVs. The standard deviation over the 100 repetitions was also calculated.

For all epitopes of interest other than MAP-Tau, the small number of FOVs measured for each subject required an alternative to the method of direct sampling used

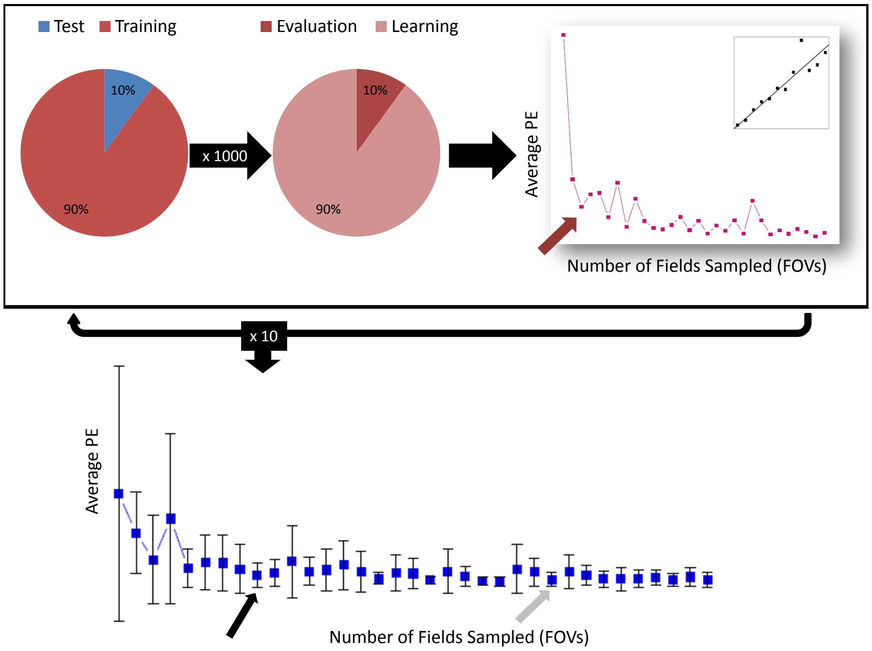

Figure 2: (1) Division of Cohort into Test Set and Training Set. Repeated 10 times. (2) Division of training set into learning set and evaluation set. Repeated 1000 times. (3) Fitting of linear regression over learning set. Performed for sample sizes of 1 – 35 FOVs. Calculation of average prediction error over evaluation set. Red arrow indicates first local minimum. (4) Calculation of average prediction error over the test set. Gray arrow indicates over local minimum over 10 training sets. Black arrow indicates smallest value within one standard error of average first local minimum.

for MAP-Tau. Direct sampling would have introduced bias into the analysis, because of the relatively small number of FOVs available for each subject. For example, given a subject with only 10 FOVs, a sample of size of 10 would have consisted of all available FOVs from that subject's whole tissue section. Therefore, the average and standard deviation from each subject was used to describe a normal distribution. Then, randomly generated observations from that normal distribution were sampled as above.

For epitopes other than MAP-Tau, in the first layer, leave-one-out cross-validation was used in place of 10-fold cross-validation. That is, in each iteration of the cross-validation, the test set consisted of one subject and the remaining subjects constituted the training set. Again, in the second layer, the training set was subdivided into learning and evaluation sets. However due to the small sample sizes, instead of Monte Carlo Cross-Validation, we employed bootstrap sampling, in which a training set of size n was sampled with replacement to create a learning set of size n. Subjects not selected for the learning set made up the evaluation set. A linear model was used in a similar manner as for MAP-Tau and an optimal number of FOVs was selected by averaging the prediction error in the evaluation set over 1000 iterations of the training set splitting procedure. Test set error was calculated in the same manner as for MAP-Tau and the one-standard-error parsimony rule again applied to select the final "optimal" number of FOVs. As in the MAP-Tau cohort, the entire process was repeated 100 times and the average and standard deviation calculated.

In order to test the validity of the simulated sampling method used for these epitopes, an additional analysis was performed on the MAP-Tau data. For each of the 122 subjects, a subset of 20 FOVs was randomly sampled from all FOVs available. Randomly generated values from a normal distribution described by the mean and variance of the 20 FOV subset was then used for selection of optimal number of FOVs and calculation of prediction error was then performed.

For all epitopes, to assess how close the predicted value was to the overall average AQUA score, we computed the absolute distance of the two values divided by the standard deviation of AQUA scores for each person as:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{x}_i - \bar{x}_i|}{s_{x_i}}, \tag{2}$$

where N, $\bar{x}_i$, and K are defined in Equation 1 and $s_{x_i} = \frac{1}{K}\sum_{j=1}^{K}(x_j - \bar{x}_i)^2$. This value was then averaged over the layers of cross-validation resulting in an average absolute standardized score. The `R` Language and Environment for Statistical Computing was used for all computations.

Results

Mixed-Effects Analysis of Intra-tumor Heterogeneity

We calculated an average intra-tumor coefficient of variation by epitope via a mixed-effects model fit to the AQUA scores from the 20X FOVs. Results appear in Figure 3 and are expressed as percentages with 95% confidence intervals. Overlapping intervals indicate that there is no significant difference between the coefficients of variation. Information about the location of FOVs in ROIs on the whole tissue slide was not collected for MAP-Tau and cytokeratin proteins; it therefore was not possible to calculate a coefficient of variation for these epitopes. The only significant differences detected were between the coefficients for ERK and ER. Of note, the "housekeeping" protein GAPDH, which we expected to show relatively homogeneous expression, has a coefficient of variation that is not statistically significantly different from that of ER or HER2.

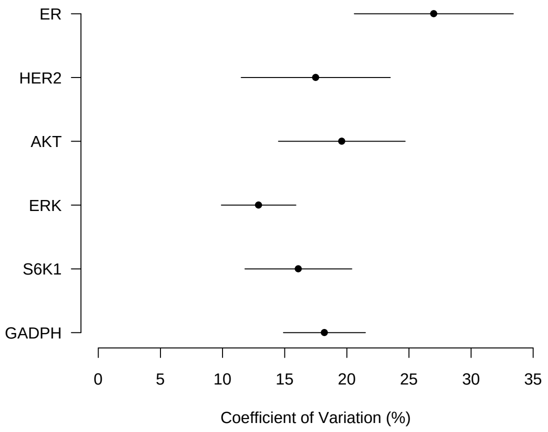

Figure 3: Coefficient of Variation (%) by epitope with 95% confidence intervals.

Cross-Validated Optimal Number of FOVs

For each epitope of interest, we simulated taking $1-35$ FOVs for a subset of subjects (the learning set). We then used the average AQUA Score of the sampled FOVs to develop a linear model. The model was used to predict scores for a distinct group of subjects, the test set, from which the same number of FOVs were sampled. Next, we calculated the PE, which is the average squared error from each set of predictions over the test set. We repeated this simulation with different learning and test sets, as described in the methods. Lastly, we located the average first local minimum of the PE and recorded the smallest number of FOVs within one standard error of this minimum. The result appears in the first column of Table 2. Also shown are the standard error of the estimate and the corresponding average absolute standardized score (Equation 2).

The optimal number of fields for epitopes ranged from $3-14$. Standard error of the estimate ranged from $1.1-4.2$, demonstrating that the estimates generated were stable. There are significant differences in the optimal number of FOVs between some of the epitopes. These differences roughly correlate with the results of the mixed-effects analysis of heterogeneity: the coefficients of variation for ER, HER2, AKT, S6K1 were not found to be significantly different and, correspondingly, the optimal FOV results for these epitopes are similar. Cytokeratin and MAP-Tau, for which it was not possible to calculate coefficients of variation, have optimal numbers of FOVs of 3 and 14 respectively. Given the qualitative heterogeneity of MAP-Tau on visual analysis and contrastingly ubiquitous expression of cytokeratin in breast carcinoma, these results support the hypothesis that markers with greater heterogeneity have a larger optimal number of FOVs. However, the correspondence between biomarker heterogeneity and optimal number of FOVs was not perfect: ER and ERK had significantly different coefficients of variation and yet had optimal number of FOVs of 8 and 6 respectively. The average absolute standardized score at the optimal number of

Marker	Optimal Number of 20X FOVs	SE of Optimal Number (FOVs)	Average Absolute Standardized Score (Equation 2)
ER	8	3.4	.31
HER2	5	3.0	.56
AKT	4	1.5	.65
ERK	6	2.5	.31
S6K1	6	3.4	.21
GAPDH	12	4.1	.24
Cytokeratin	3	4.3	.41
MAP-Tau	14	4.2	.60
MAP-Tau (direct sampling)	14	4.2	.55

Table 2: Optimal Number of Fields by Epitope with PE

fields is reported as an average distance in terms of a subjects' AQUA score standard deviation. For example, for ER, a subject's predicted score, as calculated from the optimal number of FOVs, will, on average, differ from the subject's "true" score by .31 standard deviations. The average absolute distance at the optimal number of FOVs varies slightly between epitopes but remains below one standard deviation for all but one epitope.

As described in the methods, due to the small sample size and number of FOVs, the biomarkers besides MAP-Tau were imputed by simulating from a normal distribution based on the observed mean and standard deviation of the each individual biomarkers. To test the validity of this imputation, we performed the simulation with MAP-Tau and the results were almost identical to the results when we employed direct sampling of observed data (Table 2).

Discussion

We investigated biomarker heterogeneity and the optimal number of 20X FOVs required for accurate immunostaining assessment of biomarker expression in breast carcinoma. Our mixed-effects analysis showed that, of the 7 biomarkers we examined, there were significant differences in heterogeneity, as quantified by the intra-tumor coefficient of variation. The optimal number of 20X FOVs, as determined by the cross-validated average prediction error, varied by epitope from 3 (for cytokeratin) to 14 (for MAP-Tau).

The clinical significance of our findings is two-fold. First, they demonstrate that very small core needle biopsies may be inadequate for use in diagnostic immunostains because they may not contain enough 20X FOVs to account for biomarker heterogeneity. Second, they suggest that the optimal tissue sampling algorithm required to account for biomarker heterogeneity must be determined individually for each biomarker introduced into clinical use.

The results for the optimal number of FOVs by biomarker trended with the results of the mixed-effects analysis of heterogeneity. The markers S6K1, ERK, AKT, and HER2 had similar optimal FOV sample sizes and a correspondingly large overlap in the 95% confidence intervals for their coefficients of variation. ER, which had the highest measured coefficient of heterogeneity, had a relatively large optimal sample size. Although it was not possible to calculate a coefficient of variation for MAP-Tau, its large optimal FOV sample size is consistent with the qualitative heterogeneity observed in immunostains. The similarity of the optimal number of FOVs between ER and ERK, despite significant differences in their coefficients of correlation, demonstrates imperfect correspondence between mixed-effects modeling of heterogeneity and the optimal number of FOVs. This suggests that optimal sampling must be empirically calculated for each marker rather than predicted from statistical models of marker heterogeneity.

The differences between the optimal number of FOVs for the biomarkers we tested suggests that there exists no single, optimal sampling algorithm for all biomarkers in breast carcinoma. Instead, the optimal number must be determined on a marker-by-marker basis. Biomarkers that are known to be more heterogeneous, such as

MAP-Tau, are likely to require more FOVs; however, for the reasons stated above, precise sampling algorithms must be empirically determined.

This study has several limitations. The the most important limitation is that we used the average AQUA score over all FOVs in a whole tissue slide to model the "true" representative score for each tumor when calculating prediction error. Similarly, we used FOVs from one whole tissue slide per subject to calculate the coefficient of variation for each biomarker. The variation within a single whole tissue slide may be far less than the variation between histologic "blocks" (1 cm^3 sections) from different regions of tumor. As described above, Chung et al. found statistically significant differences between AQUA scores from different blocks of tumor [35]. Different blocks are more likely to encompass various tumor micro-environments and different tumor cell subpopulations. Consequently both the coefficient of variation and the optimal number of FOVs reported in this study may underestimate variation in the tumor as whole.

Our results may be conservatively interpreted as the minimum number of FOVs required for clinical use. In clinical practice, immunostaining for biomarkers is sometimes performed on core needle biopsies, which are much smaller than whole tissue sections. Our findings offer guidance regarding the size of core needle biopsy sample required to represent biomarker expression in a single whole tissue section. Additional studies, using multiple blocks from the same tumor, will be required to draw inferences about the optimal number of FOVs required to represent the tumor as a whole.

A second limitation of this study is the relatively small size of the cohort on which most of the biomarkers were measured. The mixed-effects analysis only detected a significant difference between coefficients of variation for two of the epitopes examined: ER and ERK. It is possible that the study was underpowered to detect small differences in coefficients of variation between epitopes; a larger cohort size may have produced more stable estimates of the coefficients of variation, allowing for detection of small, but statistically significant differences. In addition – based on what is known about distinct biological roles of GAPDH versus ER and HER2 – we would not have predicted that we would find no significant difference between the coefficients of variation for these markers. It is possible that, in this small cohort, the cases selected had relatively homogeneous expression of HER2 and ER that was not representative of the typical level of intra-tumor heterogeneity in breast carcinomas.

Another consequence of the small cohort size was that we were required to use imputed values in the cross-validation analysis. For all biomarkers other than MAP-Tau, in order to avoid introducing bias, we simulated sampling FOVs from a normal distribution described by the measured mean and variation of observed FOVs. However, the validity of this method is supported by our dual analysis of MAP-Tau, which was measured on a large cohort (n=122), with a large number of FOVs measured per subject. When the MAP-Tau data was analyzed by both direct sampling and simulation, the results for the optimal number of fields and standard error of the estimate were identical. This is strong evidence that neither the point estimate for optimal number of FOVs for epitopes other than MAP-Tau nor the stability of this estimate were affected by the small cohort size.

The third limitation is that AQUA is not currently used in many clinical laboratories. AQUA employs fluorescence for visualization and optimal quantification rather than the diaminobenzidine (DAB) stain used in most conventional labs. However, the underlying immunohistochemistry technique and biology are the same, so the results should be generalizable to any method of visualization. For several reasons, AQUA is superior to DAB for the purposes of this study. In validation studies, AQUA has demonstrated superior reproducibility and predictive power (of clinical outcomes) when compared to pathologist-based scoring systems of DAB stains [41]. AQUA measures a much greater dynamic range of scores than DAB, allowing it differentiate between levels of biomarker expression that might be indistinguishable with DAB staining [49]. AQUA also allows for more powerful statistical analysis: it measures biomarker expression on a continuous scale, which has greater statistical power to detect differences than the ordinal scale employed in pathologist-based scoring systems of DAB stains.

This pilot study offers guidance regarding the size of tissue sample that is required to account for heterogeneity in the specific biomarkers studied. More broadly, it suggests that further investigations are necessary in order to describe optimal sampling for other biomarkers in pre-clinical or clinical use. The implication for clinical practice is that number of fields assessed is a critical parameter for companion diagnostic tests and should be optimized prior to introduction of new biomarker assays. While this is a study of breast carcinoma tumors, the implications of these findings extend to biomarkers used in other types of tissue.

References

[1] W. McGuire, “Current status of estrogen receptors in human breast cancer,” *Cancer*, vol. 36, no. S2, pp. 638–644, 1975.

[2] W. Knight, R. Livingston, E. Gregory, and W. McGuire, “Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer,” *Cancer research*, vol. 37, no. 12, p. 4669, 1977.

[3] B. Fisher, C. Redmond, A. Brown, N. Wolmark, J. Wittliff, E. Fisher, D. Plotkin, D. Bowman, S. Sachs, J. Wolter, *et al.*, “Treatment of primary breast cancer with chemotherapy and tamoxifen,” *New England Journal of Medicine*, vol. 305, no. 1, pp. 1–6, 1981.

[4] B. Fisher, C. Redmond, A. Brown, D. Wickerham, N. Wolmark, J. Allegra, G. Escher, M. Lippman, E. Savlov, and J. Wittliff, “Influence of tumor estrogen and progesterone receptor levels on the response to tamoxifen and chemotherapy in primary breast cancer,” *Journal of Clinical Oncology*, vol. 1, no. 4, p. 227, 1983.

[5] B. Fisher, J. Costantino, C. Redmond, R. Poisson, D. Bowman, J. Couture, N. Dimitrov, N. Wolmark, D. Wickerham, E. Fisher, *et al.*, “A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor–positive tumors,” *New England Journal of Medicine*, vol. 320, no. 8, pp. 479–484, 1989.

[6] M. Hammond, D. Hayes, M. Dowsett, D. Allred, K. Hagerty, S. Badve, P. Fitzgibbons, G. Francis, N. Goldstein, M. Hayes, *et al.*, “American society of clinical oncology/college of American pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer,” *Journal of Clinical Oncology*, vol. 28, no. 16, p. 2784, 2010.

[7] M. Clarke, R. Collins, S. Darby, C. Davies, P. Elphinstone, E. Evans, J. Godwin, R. Gray, C. Hicks, S. James, *et al.*, “Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials.,” *Lancet*, vol. 366, no. 9503, p. 2087, 2005.

[8] J. Harvey, G. Clark, C. Osborne, and D. Allred, “Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer,” *Journal of Clinical Oncology*, vol. 17, no. 5, p. 1474, 1999.

[9] G. Clark, W. McGuire, C. Hubay, O. Pearson, and J. Marshall, “Progesterone receptors as a prognostic factor in stage II breast cancer,” *New England Journal of Medicine*, vol. 309, no. 22, p. 1343, 1983.

[10] M. Press, M. Pike, V. Chazin, G. Hung, J. Udove, M. Markowicz, J. Danyluk, W. Godolphin, M. Sliwkowski, R. Akita, *et al.*, “Her-2/neu expression in node-negative breast cancer: direct tissue quantitation by computerized image analysis and association of overexpression with increased risk of recurrent disease,” *Cancer research*, vol. 53, no. 20, p. 4960, 1993.

[11] M. Press, L. Bernstein, P. Thomas, L. Meisner, J. Zhou, Y. Ma, G. Hung, R. Robinson, C. Harris, A. El-Naggar, *et al.*, “HER-2/neu gene amplification characterized by fluorescence in situ hybridization: poor prognosis in node-negative breast carcinomas,” *Journal of Clinical Oncology*, vol. 15, no. 8, p. 2894, 1997.

[12] D. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, *et al.*, “Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2,” *New England Journal of Medicine*, vol. 344, no. 11, p. 783, 2001.

[13] D. Slamon, W. Eiermann, N. Robert, T. Pienkowski, M. Martin, M. Pawlicki, *et al.*, “Phase III randomized trial comparing doxorubicin and cyclophosphamide followed by docetaxel (ACT) with doxorubicin and cyclophosphamide followed by docetaxel and trastuzumab (ACTH) with docetaxel, carboplatin and trastuzumab (TCH) in HER2 positive early breast cancer patients: BCIRG 006 study,” *Breast Cancer Res Treat*, vol. 94, no. suppl 1, p. S5, 2005.

[14] A. Wolff, M. Hammond, J. Schwartz, K. Hagerty, D. Allred, R. Cote, M. Dowsett, P. Fitzgibbons, W. Hanna, A. Langer, *et al.*, "American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer," *Journal of Clinical Oncology*, vol. 25, no. 1, p. 118, 2007.

[15] R. Bast, P. Ravdin, D. Hayes, S. Bates, H. Fritsche, J. Jessup, N. Kemeny, G. Locker, R. Mennel, and M. Somerfield, "2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology," *Journal of Clinical Oncology*, vol. 19, no. 6, p. 1865, 2001.

[16] J. Ross, W. Symmans, L. Pusztai, and G. Hortobagyi, "Breast cancer biomarkers," *Advances in Clinical Chemistry*, vol. 40, pp. 99–125, 2005.

[17] K. Kerlikowske, A. Molinaro, M. Gauthier, H. Berman, F. Waldman, J. Bennington, H. Sanchez, C. Jimenez, K. Stewart, K. Chew, B.-M. Ljung, and T. Tlsty, "Biomarker expression and risk of subsequent tumors after initial ductal carcinoma in situ diagnosis," *Journal of the National Cancer Institute*, vol. 102, no. 9, pp. 627–637, 2010.

[18] S. Wolman, R. Pauley, A. Mohamed, P. Dawson, D. Visscher, and F. Sarkar, "Genetic markers as prognostic indicators in breast cancer," *Cancer*, vol. 70, no. S4, pp. 1765–1774, 1992.

[19] D. Weinstat-Saslow, M. Merino, R. Manrow, J. Lawrence, R. Bluth, K. Wittenbel, J. Simpson, D. Page, and P. Steeg, "Overexpression of cyclin D mRNA distinguishes invasive and in situ breast carcinomas from non-malignant lesions," *Nature Medicine*, vol. 1, no. 12, pp. 1257–1260, 1995.

[20] K. Keyomarsi, S. Tucker, T. Buchholz, M. Callister, Y. Ding, G. Hortobagyi, I. Bedrosian, C. Knickerbocker, W. Toyofuku, M. Lowe, *et al.*, "Cyclin E and survival in patients with breast cancer," *New England Journal of Medicine*, vol. 347, no. 20, p. 1566, 2002.

[21] C. Rochlitz, G. Scott, J. Dodson, E. Liu, C. Dollbaum, H. Smith, and C. Benz, "Incidence of activating ras oncogene mutations associated with primary and metastatic human breast cancer," *Cancer research*, vol. 49, no. 2, p. 357, 1989.

[22] J. Bridge, M. Nelson, E. McComb, M. McGuire, H. Rosenthal, G. Vergara, G. Maale, S. Spanier, and J. Neff, "Cytogenetic findings in 73 osteosarcoma specimens and a review of the literature* 1," *Cancer genetics and cytogenetics*, vol. 95, no. 1, pp. 74–87, 1997.

[23] P. Bertheau, F. Plassa, M. Espie, E. Turpin, A. De Roquancourt, M. Marty, F. Lerebours, Y. Beuzard, and A. Janin, "Effect of mutated TP53 on response of advanced breast cancers to high-dose chemotherapy," *The Lancet*, vol. 360, no. 9336, pp. 852–854, 2002.

[24] M. Daidone, S. Veneroni, E. Benini, G. Tomasic, D. Coradini, M. Mastore, C. Brambilla, L. Ferrari, and R. Silvestrini, "Biological markers as indicators of response to primary and adjuvant chemotherapy in breast cancer," *International Journal of Cancer*, vol. 84, no. 6, pp. 580–586, 1999.

[25] C. Brinckerhoff and L. Matrisian, "Matrix metalloproteinases: a tail of a frog that became a prince," *Nature Reviews Molecular Cell Biology*, vol. 3, no. 3, pp. 207–214, 2002.

[26] L. McCawley and L. Matrisian, "Matrix metalloproteinases: multifunctional contributors to tumor progression," *Molecular medicine today*, vol. 6, no. 4, pp. 149–156, 2000.

[27] C. Benaud, R. Dickson, and E. Thompson, “Roles of the matrix metalloproteinases in mammary gland development and cancer,” *Breast cancer research and treatment*, vol. 50, no. 2, pp. 97–116, 1998.

[28] L. Harris, H. Fritsche, R. Mennel, L. Norton, P. Ravdin, S. Taube, M. Somerfield, D. Hayes, and R. Bast, “American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer,” *Journal of Clinical Oncology*, vol. 25, no. 33, p. 5287, 2007.

[29] G. Francis, M. Dimech, L. Giles, and A. Hopkins, “Frequency and reliability of oestrogen receptor, progesterone receptor and HER2 in breast carcinoma determined by immunohistochemistry in Australasia: Results of the RCPA Quality Assurance Program,” *Journal of clinical pathology*, vol. 60, no. 11, p. 1277, 2007.

[30] A. Rhodes, B. Jasani, D. Barnes, L. Bobrow, and K. Miller, “Reliability of immunohistochemical demonstration of oestrogen receptors in routine practice: interlaboratory variance in the sensitivity of detection and evaluation of scoring systems,” *Journal of clinical pathology*, vol. 53, no. 2, p. 125, 2000.

[31] G. Uy, A. Laudico, A. Fernandez, F. Lim, J. Carnate, R. Rivera, *et al.*, “Immunohistochemical assay of hormone receptors in breast cancer at the Philippine General Hospital: Importance of early fixation of specimens,” *Philipp J Surg Spec*, vol. 62, no. 3, pp. 123–127, 2007.

[32] G. Vance, T. Barry, K. Bloom, P. Fitzgibbons, D. Hicks, R. Jenkins, D. Persons, R. Tubbs, and M. Hammond, “Genetic heterogeneity in HER2 testing in breast cancer: panel summary and guidelines,” *Archives of pathology & laboratory medicine*, vol. 133, no. 4, pp. 611–612, 2009.

[33] M. Clarke, J. Dick, P. Dirks, C. Eaves, C. Jamieson, D. Jones, J. Visvader, I. Weissman, and G. Wahl, “Cancer stem cellsperspectives on current status

and future directions: AACR workshop on cancer stem cells," *Cancer research*, vol. 66, no. 19, p. 9339, 2006.

[34] J. Meyer and J. Wittliff, "Regional heterogeneity in breast carcinoma: thymidine labelling index, steroid hormone receptors, DNA ploidy," *International Journal of Cancer*, vol. 47, no. 2, pp. 213–220, 1991.

[35] G. Chung, M. Zerkowski, S. Ghosh, R. Camp, and D. Rimm, "Quantitative analysis of estrogen receptor heterogeneity in breast cancer," *Laboratory investigation*, vol. 87, no. 7, pp. 662–669, 2007.

[36] A. Nassar, A. Radhakrishnan, I. Cabrero, G. Cotsonis, and C. Cohen, "Intratumoral Heterogeneity of Immunohistochemical Marker Expression in Breast Carcinoma: A Tissue Microarray-based Study," *Applied Immunohistochemistry & Molecular Morphology*, 2010.

[37] O. Kallioniemi, A. Kallioniemi, W. Kurisu, A. Thor, L. Chen, H. Smith, F. Waldman, D. Pinkel, and J. Gray, "ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization," *Proceedings of the National Academy of Sciences*, vol. 89, no. 12, p. 5321, 1992.

[38] S. Lassmann, M. Bauer, R. Soong, J. Schreglmann, K. Tabiti, J. Nährig, R. Rüger, H. Höfler, and M. Werner, "Quantification of CK20 gene and protein expression in colorectal cancer by RT-PCR and immunohistochemistry reveals inter-and intratumour heterogeneity," *The Journal of Pathology*, vol. 198, no. 2, pp. 198–206, 2002.

[39] L. Sigalotti, E. Fratta, S. Coral, S. Tanzarella, R. Danielli, F. Colizzi, E. Fonsatti, C. Traversari, M. Altomonte, and M. Maio, "Intratumor heterogeneity of cancer/testis antigens expression in human cutaneous melanoma is methylation-

regulated and functionally reverted by 5-aza-2′-deoxycytidine," *Cancer research*, vol. 64, no. 24, p. 9167, 2004.

[40] D. Chhieng, A. Frost, S. Niwas, H. Weiss, W. Grizzle, and S. Beeken, "Intratumor heterogeneity of biomarker expression in breast carcinomas," *Biotechnic and Histochemistry*, vol. 79, no. 1, pp. 25–36, 2004.

[41] R. Camp, G. Chung, and D. Rimm, "Automated subcellular localization and quantification of protein expression in tissue microarrays," *Nature medicine*, vol. 8, no. 11, pp. 1323–1328, 2002.

[42] B. Bolstad, R. Irizarry, M. Astrand, and T. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, p. 185, 2003.

[43] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[44] J. Pinheiro and D. Bates, *Mixed-effects models in S and S-PLUS.* Springer Verlag, 2009.

[45] A. Molinaro and K. Lostritto, "Statistical resampling for large screening data analysis such as classical resampling bootstrapping, markov chain monte carlo, and statistical simulation and validation strategies.," in *STATISTICAL BIOINFORMATICS: A GUIDE FOR LIFE AND BIOMEDICAL SCIENCE RESEARCHERS* (J. K. Lee, ed.), pp. 219–248, John Wiley & Sons, Inc., 2010.

[46] A. Molinaro, R. Simon, and R. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005.

[47] A. Molinaro and K. Lostritto, “STATISTICAL RESAMPLING TECHNIQUES FOR LARGE BIOLOGICAL DATA ANALYSIS,” *Statistical Bioinformatics: For Biomedical and Life Science Researchers*, p. 219, 2010.

[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Springer Verlag, 2001.

[49] D. Rimm, “What brown cannot do for you,” *Nature*, vol. 200, p. 6, 2006.