

MARKERS OF EPITHELIAL-MESENCHYMAL TRANSITION AND COLORECTAL
CANCER MORTALITY: TIME-TO-EVENT AND LATENT-CLASS ANALYSES

Evan L. Busch

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
2015

Approved by:

Robert S. Sandler

Christy L. Avery

David A. Eberhard

Temitope O. Keku

David B. Richardson

©2015
Evan L. Busch
ALL RIGHTS RESERVED

ABSTRACT

Evan L. Busch: Markers of Epithelial-Mesenchymal Transition and Colorectal Cancer Mortality:
Time-to-Event and Latent-Class Analyses
(Under the direction of Robert S. Sandler)

Most cancer arises in epithelial cells and most cancer deaths are due to metastases. Many cancer patients diagnosed with local disease according to lymph-node evaluation and radiologic imaging later experience disease recurrence within a few years of surgery. An additional measure of cancer cell detachment from the primary tumor taken in the primary tumor itself at the time of surgery might improve our ability to identify which patients are at risk for recurrence and therefore should have their treatments adjusted accordingly.

Epithelial-mesenchymal transition (EMT) is a mechanism of cancer cell metastasis that connects epithelial cells to metastasis. It identifies candidate markers for the additional diagnostic test needed to stratify cancer patients by risk for recurrence.

I measured the EMT markers E-cadherin, Integrin beta-6, and Snail in primary tumors from subjects in a population-based, case-only prospective cohort of colorectal cancer patients. Using Cox proportional hazards models, I estimated the association between each marker and time from surgery to death. I found that E-cadherin expression measured as a weighted average of tumor cores was associated with time to death. No other marker expression variable was associated with outcomes.

Using latent class analysis, I estimated the sensitivity and specificity of E-cadherin expression as a weighted average of tumor cores to classify subjects as those likely to have cancer cells detaching or not detaching from the primary tumor at the time of surgery, in

conjunction with lymph node evaluation and radiologic imaging. Latent class analysis permitted estimation of sensitivity and specificity under the realistic assumption that none of the tests constituted a gold-standard measure of whether cancer cells had detached from the primary tumor by the time of diagnosis. Across the various latent class models that I explored, I found a peak E-cadherin sensitivity of 59% and peak specificity of 94%.

My results suggested that E-cadherin measurements in colorectal primary tumors at the time of surgery might improve the ability of clinicians to assess whether the patient is at risk for recurrence. Incorporating such measurements into standard colorectal cancer diagnostic procedures could thereby help to improve patient outcomes.

To my parents, the survivors who have inspired my work.

ACKNOWLEDGEMENTS

I have been especially fortunate to have completed my doctoral work at UNC-Chapel Hill. The outstanding resources and personnel at the university were used to the full to carry out this project. The Translational Pathology Laboratory performed the tissue staining and biomarker measurements. Throughout the process the entire lab was a model of helpfulness, efficiency, thoroughness, and high-quality work. Nana Feinberg helped to identify appropriate antibodies and advised on coordinating the entire process. Yongjuan Xia performed the staining. Stephanie Cohen segmented the images, developed the annotation and scoring algorithms, and carried out the application of the algorithms. Her generosity of advice and patience while I learned the process went far beyond the call of duty.

Nikki McCoy from the Keku Laboratory prepared the tissue microarray slides and maps. Christopher Martin provided database management assistance in working with the CanCORS datasets. Thanks go to the Funkhouser Lab for providing control tissue specimens. Annie Green Howard from the UNC Department of Biostatistics provided critical statistical advice late in the project. Special thanks go to Christopher Wiesen of the UNC Odum Institute for his spectacular statistical advice and support at several steps along the way. Kate McGraw of the UNC Health Sciences Library was our collaborator on the systematic review that guided the design of many aspects of the data collection and analysis. The entire administrative staff of the Epidemiology Department facilitated my time in graduate school in ways too numerous to specify, with particular thanks to Nancy Colvin, Carmen Woody, and Valerie Hudock from Epidemiology Student Services.

I literally could not have made it through graduate school without support from the various funding sources I have enjoyed during my time at UNC. The UNC Graduate School provided a scholarship that covered my first year in the program. Subsequently, Dora Il'yasova, Clara Lee, Victor Schoenbach, Lorraine Alexander, Christy Avery, Eric Whitsel, Yang Claire Yang, and Jose Zevallos gave me research and teaching assistant opportunities that supported me and made my dissertation work possible. I also learned a tremendous amount about research from them that indirectly influenced the dissertation in ways large and small.

David Richardson, Christy Avery, Temitope Keku, and David Eberhard reached the Platonic ideal for dissertation committee members. They generously provided every bit of time, effort, and insight that I needed. They were unfailingly flexible about making changes to the project as we discovered better approaches over the course of the work. I believe that, not far off in the future, cancer patients will experience real benefits that derive from the ideas they contributed to this project. I would like to highlight Dr. Keku's work years ago towards assembling the cohort, as well as her advice and training me in the collection and analysis of the marker expression data. Dr. Richardson provided analytic insights that fundamentally shaped both the time-to-event and latent class analyses.

I especially want to thank my advisor and chair Robert Sandler. Listing everything he has done for me would require another dissertation by itself. He gave me access to the data from NC-CanCORS, and all of the work depended on his efforts to put together the cohort before I arrived at UNC. He provided painstakingly-collected tissue specimens and, via the Sessions Professorship, funds to carry out the laboratory work. Above all, he made a tremendous investment of time and effort in my graduate training from which I will benefit for decades. His

attention to detail, rigor, and insights into both substantive and process matters have helped make me a better researcher than I ever could have expected to become.

Thanks to my family and friends for their love and support, especially to my mom Diane and my buddy Todd Bergstrom for shouldering most of the load with my whining and complaining over the past few years. Finally, perhaps the deepest gratitude of all goes to the American military, taxpayers, and police. They make everything in our society possible. Public health in general, and this dissertation in particular, are no exceptions.

TABLE OF CONTENTS

LIST OF TABLES	xiii
LIST OF FIGURES	xiv
LIST OF ABBREVIATIONS.....	xvi
CHAPTER 1. SPECIFIC AIMS	1
CHAPTER 2. BACKGROUND AND SYSTEMATIC LITERATURE REVIEW	3
2.1. Colorectal cancer	3
2.2. Role of metastasis in CRC outcomes.....	4
2.3. Epithelial-Mesenchymal Transition.....	6
2.4. Public health significance	8
2.5. Systematic review and marker selection.....	10
2.5.1. Systematic review of literature on EMT markers and CRC outcomes	10
2.5.2. Marker selection criteria	12
2.5.3. Summary of systematic review findings.....	13
2.5.4. Marker selection conclusions.....	17
CHAPTER 3. METHODS	20
3.1. Subject selection and enrollment	20
3.1.1. Study population	20
3.1.2. Subject identification, recruitment, and enrollment.....	21
3.1.3. Subject eligibility criteria.....	22

3.2. Specimen collection	22
3.2.1. Collection at clinics.....	22
3.2.2. Specimen preparation and storage at UNC.....	23
3.2.3. Control specimens.....	24
3.2.4. Sample size	24
3.3. Specimen staining	28
3.3.1. Acquisition of antibodies for immunohistochemistry.....	28
3.3.2. Immunohistochemistry staining procedures	28
3.4. Digital image annotation and analysis	31
3.4.1. Annotation rationale and goal	31
3.4.2. Image annotation and scoring procedures.....	32
3.4.3. Linkage of marker data with main CanCORS data	35
3.5. Assessment of reliability.....	35
3.6. Statistical analysis for time-to-event analysis.....	36
3.6.1. Overview	36
3.6.2. Definitions.....	37
3.6.3. Descriptive statistics	38
3.6.4. Identification of statistically-optimal cut points of marker expression	40
3.6.5. Bivariate time-to-event analysis	42
3.6.6. Covariate selection.....	44
3.6.7. Multivariate time-to-event analysis	46
3.6.8. Missing data and multiple imputation.....	47
3.7. Statistical analysis for latent class analysis.....	49

3.7.1. Overview	49
3.7.2. Latent class framework	51
3.7.3. Latent class model for diagnostics of cancer cell detachment from primary tumors.....	52
3.7.4. Data used for latent class analysis	53
3.7.5. Analysis.....	56
3.8. Ethical Considerations	57
3.8.1. IRB approval.....	57
3.8.2. Informed consent	57
3.8.3. Potential risks to subjects.....	58
3.8.4. Potential benefits to subjects.....	58
CHAPTER 4. MARKERS OF EPITHELIAL-MESENCHYMAL TRANSITION AND MORTALITY IN A POPULATION-BASED PROSPECTIVE COHORT OF COLORECTAL CANCER PATIENTS	59
4.1. Introduction.....	59
4.2. Methods.....	60
4.2.1. Study population	60
4.2.2. Immunohistochemistry	61
4.2.3. Automated analysis of digital IHC images	62
4.2.4. Outcome	64
4.2.5. Covariates	65
4.2.6. Marker expression cut point optimization	65
4.2.7. Statistical analysis	66
4.3. Results.....	67
4.4. Discussion	69

CHAPTER 5. ESTIMATING SENSITIVITY AND SPECIFICITY OF MARKERS OF EPITHELIAL-MESENCHYMAL TRANSITION TO EVALUATE CANCER CELL DETACHMENT FROM PRIMARY TUMORS: A LATENT CLASS ANALYSIS	95
5.1. Introduction.....	95
5.2. Methods.....	96
5.2.1. Study population	96
5.2.2. Latent class models	96
5.2.3. Diagnostic tests of cancer cell detachment from primary tumors	97
5.2.4. Analysis.....	99
5.3. Results.....	101
5.4. Discussion	102
CHAPTER 6. DISCUSSION.....	111
6.1. Summary of findings.....	111
6.1.1. Associations between EMT marker expression and time-to-death	111
6.1.2. Latent class estimates of EMT sensitivity and specificity to evaluate cancer cell detachment from primary tumors	113
6.2. Strengths and limitations.....	115
6.3. Public health implications and future directions.....	121
APPENDIX A. SEARCH TERMS FOR SYSTEMATIC LITERATURE REVIEW OF EMT MARKERS AND OUTCOMES	123
APPENDIX B. MARKER CUT POINT OPTIMIZATION MACRO: SAS CODE AND DETAILS	124
APPENDIX C. LATENT CLASS ANALYSIS: SAS CODE AND DETAILS.....	130
REFERENCES	132

LIST OF TABLES

Table 1. Sample size calculations for initial selection of subjects.....	25
Table 2. Rules for inferring lymph node and radiologic imaging test results from tumor stage	55
Table 3. Subject characteristics for overall North Carolina CanCORS and subset in whose primary tumors EMT markers were measured	78
Table 4. Average continuous EMT marker expression in tumor tissue compared to normal adjacent tissue	79
Table 5. Optimally-dichotomized marker expression status cross-tabulated with tumor stage and with risk of dying within 5 years of surgery	80
Table 6. Unadjusted and adjusted Cox proportional hazards models of the effect of optimally-dichotomized marker expression status on time-to-death censored at 5 years after surgery	81
Table 7. Trade-offs between strength of cut-point/time-to-death association and number of patients whose treatments would be reassigned, by stage distribution and 5-year risk of death for E-cadherin weighted average status	82
Table 8. Subject characteristics.....	107
Table 9. Estimated sensitivity and specificity of EMT markers, and prevalence of high risk for metastatic disease, at EMT cut point of 0.52 and varying fixed values of sensitivity or specificity for other tests	108
Table 10. Estimated sensitivity and specificity of EMT markers, and prevalence of high risk for metastatic disease, at EMT cut point of 0.60 and varying fixed values of sensitivity or specificity for other tests	109
Table 11. Estimated sensitivity and specificity of EMT markers, and prevalence of high risk for metastatic disease, at EMT cut point of 0.85 and varying fixed values of sensitivity or specificity for other tests	110

LIST OF FIGURES

Figure 1. Epithelial-mesenchymal transition and associated molecular markers	7
Figure 2. Conceptual model of the role of EMT markers in determining CRC patient outcomes	10
Figure 3. Flow of patient eligibility and inclusion.....	27
Figure 4. Immunohistochemistry examples for positive and negative staining for each of E-cadherin, Integrin beta-6, and Snail	29
Figure 5. Illustration of Tissue Studio analysis of colon tumor tissue.....	30
Figure 6. Illustration of Tissue Studio analysis of normal colon tissue.....	31
Figure 7. Reductionist and heterotypic representations of cancer tumors	32
Figure 8. Directed acyclic graph (causal diagram) depicting postulated relationships between primary tumor cancer cell EMT marker expression, time-to-death, and other variables.....	45
Figure 9. Generic latent class model with three indicator variables	52
Figure 10. Latent class model of diagnostic tests to evaluate cancer cell detachment from primary tumors.....	53
Figure 11. Kaplan-Meier survival stratified by levels of dichotomous E-cadherin weighted average expression status	83
Figure 12. Kaplan-Meier survival stratified by levels of dichotomous E-cadherin worst core expression status.....	84
Figure 13. Kaplan-Meier survival stratified by levels of dichotomous Integrin beta-6 weighted average expression status.....	85
Figure 14. Kaplan-Meier survival stratified by levels of dichotomous Integrin beta-6 worst core expression status	86
Figure 15. Kaplan-Meier survival stratified by levels of dichotomous Snail weighted average expression status	87
Figure 16. Kaplan-Meier survival stratified by levels of dichotomous Snail worst core expression status	88

Figure 17. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin weighted average expression status and dichotomous Integrin beta-6 weighted average expression status.....	89
Figure 18. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin worst core expression status and dichotomous Integrin beta-6 worst core expression status	90
Figure 19. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin weighted average expression status and dichotomous	91
Figure 20. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin worst core expression status and dichotomous Snail worst core expression status	92
Figure 21. Kaplan-Meier survival jointly stratified by levels of dichotomous Integrin beta-6 weighted average expression status and dichotomous Snail weighted average expression status	93
Figure 22. Kaplan-Meier survival jointly stratified by levels of dichotomous Integrin beta-6 worst core expression status and dichotomous Snail worst core expression status	94

LIST OF ABBREVIATIONS

BIC	Bayesian information criterion model fit statistic
CanCORS	Cancer Care Outcomes Research and Surveillance Consortium
CI	Confidence interval
CRC	Colorectal cancer
DAG	Directed acyclic graph
Df	Degrees of freedom
EMT	Epithelial-mesenchymal transition
HR	Hazard ratio
IHC	Immunohistochemistry
LCA	Latent class analysis
LN	Lymph node evaluation
MI	Multiple imputation
NC	North Carolina
RI	Radiologic imaging
ROC	Receiver operating characteristic curve
UNC	University of North Carolina at Chapel Hill

CHAPTER 1. SPECIFIC AIMS

Roughly 80% of cancer arises from epithelial cells and about 90% of cancer deaths are due to metastases (1). Epithelial-mesenchymal transition (EMT) is a mechanism of cancer cell metastasis that links epithelial biology to the detachment of cancer cells from primary tumors (2). EMT markers measured in primary tumor cancer cells could be important to identify patients at risk for metastatic disease, even among those with no evidence of metastatic disease according to lymph node evaluation or radiologic imaging. The impact of EMT markers could be especially great in colorectal cancer (CRC), which is 95% epithelial in origin (3) and in which about 25% of patients diagnosed with local disease eventually experience recurrence (4). This suggests that many patients have metastatic disease at the time of diagnosis that is not successfully captured by conventional diagnostics of lymph node evaluation and imaging.

Although several dozen studies of EMT markers in CRC primary tumors and patient outcomes have been conducted over the past decade, the methods and results have been inconsistent (5). In addition, data analyses have been crude and not informative in terms of whether the markers could be translated to the clinic, where they could impact public health by modifying tumor stage diagnosis procedures. I undertook this research to develop marker measurement and data analysis procedures for studies of EMT markers and patient outcomes that can standardize methods across studies and provide information that permits clear evaluation of whether and how a marker could be used clinically. The goal was to develop methods that will facilitate translation of EMT markers to clinical use so that cancer patients may benefit from the opportunity that our biological knowledge of EMT has presented. The procedures we developed

for this research were applied to subjects from a population-based, case-only prospective cohort of CRC patients. The specific aims of the project were as follows:

Specific Aim 1: Estimate associations between EMT marker expression levels in CRC primary tumor cancer cells and time from surgery to patient death.

- a) Identify the optimal scale on which to measure each marker.
- b) Determine the set of adjustment covariates to include in statistical models to produce valid estimates of associations between EMT marker expression levels and patient outcomes.
- c) Develop a method to identify clinically-informative cut points for observed values along the scale of expression chosen for a given marker.

Hypothesis: Low E-cadherin expression, and high Integrin beta-6 and Snail expression, will be associated with shorter times from surgery to death.

Specific Aim 2: Using latent class analysis, estimate the sensitivity and specificity of EMT markers to assess cancer cell detachment from primary tumors without assuming that lymph node evaluation or radiologic imaging are gold standard measures of such detachment.

- a) Conduct sensitivity analyses to evaluate how varying one's assumptions about the sensitivity and specificity of lymph node evaluation and imaging impact estimates of the sensitivity and specificity of EMT markers.
- b) Assess how varying the cut point used to determine dichotomous EMT marker expression status impacts estimates of the sensitivity and specificity of the EMT marker.

Hypothesis: EMT marker expression levels associated with time from surgery to death will have high sensitivity and specificity to assess cancer cell detachment from primary tumors.

CHAPTER 2. BACKGROUND AND SYSTEMATIC LITERATURE REVIEW

2.1. Colorectal cancer

Colorectal cancer (CRC) is a leading cause of morbidity and mortality both nationally and globally. In the United States alone there are over 140,000 new cases and 50,000 deaths per year attributable to the disease, giving CRC the fourth-largest incidence nationally of any tumor site (after prostate, breast, and lung cancer) and second-largest mortality (after lung cancer) (6). As is typical of cancer generally, a patient's prospects for a full recovery decline sharply the greater the spread of the tumor. The five-year relative survival proportion is approximately 90% for localized tumors, 69% for tumors that have spread to adjacent lymph nodes and organs, and 12% for tumors that have spread to distant organs at the time of diagnosis (7).

Despite substantial morbidity and mortality due to CRC, the United States has seen a declining disease burden. Overall age-adjusted CRC incidence has fallen from a peak of about 68 cases per 100,000 people per year in the mid-1980s to around 42 cases per 100,000 people per year in 2009 (8). Similarly, overall age-adjusted CRC mortality has decreased from about 28 deaths per 100,000 people per year in 1978 to roughly 16 deaths per 100,000 people per year in 2009 (8). The improved mortality rate has been attributed to a combination of screening via techniques such as colonoscopy; successful treatment via surgery, chemotherapy, and radiation therapy; and changes in population risk factor exposures (9, 10). However, the fact that the CRC morbidity and mortality burden remains high despite these trends suggests that additional clinical innovations will be needed to drive the overall disease burden still lower.

A number of risk factors are associated with CRC. Increasing risk is associated with increasing age, as over 90% of cases are diagnosed in people over the age of 50 (10). Smoking, diets high in fat and low in folate, and a history of colonic inflammatory disease also increase the risk of developing CRC (10). Blacks are at higher risk for CRC with an age-adjusted incidence of nearly 53 cases per 100,000 people per year in 2009 compared to 41 cases per 100,000 people per year among whites (8). Men are more likely to develop CRC with an age-adjusted incidence of about 48 cases per 100,000 people per year in 2009 versus 38 cases per 100,000 people per year among women (8).

Despite the vast increase in our biological understanding of cancer since the 1970s, our ability to capitalize on this knowledge to reduce cancer incidence and mortality has been relatively disappointing. Many of the anti-tumor treatments in clinical use were developed decades ago without the benefit of the molecular and genetic insights obtained since (1). However, the failure to date to match expectations does not mean that the biological insights of recent decades do not open up rich possibilities for new clinical interventions that can improve CRC outcomes. Translational research, which attempts to use basic biology knowledge to discover effective clinical interventions, holds tremendous promise to build on past advances in reducing CRC mortality.

2.2. Role of metastasis in CRC outcomes

To reduce the CRC mortality burden, effective interventions will target the ways in which CRC kills people. Of course, one approach is to prevent the disease from occurring altogether, as polypectomy during colonoscopy often does. When a tumor has already formed, an appropriate approach is to target the particular mechanisms that the tumor uses that ultimately lead to the patient's death. In CRC, as well as other cancers, this means controlling metastases.

The capacity to invade and metastasize is widely recognized as a distinguishing characteristic of cancer, where metastasis can be defined as the spread of cancer cells to parts of the body distant from the primary tumor (11). Such roving cancer cells are dangerous because they can found new tumors throughout the body. More importantly, metastases are responsible for approximately 90% of deaths due to cancer (1); it is surprisingly rare for the primary tumor to directly kill the patient.

Therefore, when any patient is diagnosed with cancer, two of the most critical questions that must be answered are whether cancer cells have already begun to detach from the primary tumor and, if so, whether healthcare providers can still successfully treat the illness. Our ability to answer these questions correctly depends on how accurately we can assess a host of additional factors. One important issue is determining whether any cancer cells in the primary tumor appear to have been capable of breaking off from it. A second consideration is whether metastases can be detected elsewhere in the body. The latter is particularly difficult to determine because most metastases break away from the primary tumor as micrometastases, defined as “[a] metastasis that is composed of a single cell or a small clump of cells and is only apparent through microscopy” (1). This definition suggests that there may be limits to how well radiologic imaging can detect every metastasis that may exist in a patient’s body at the time of tumor removal surgery. Micrometastases that break off from the primary just before the latter is removed would be too small to be seen by imaging at that time.

Determining whether cancer cells in the primary tumor at the time of surgery have acquired the ability to break off could suggest, by inference, whether other cancer cells might have already broken off. This is an appealing approach to cancer treatment. First, the primary tumor is relatively easy to locate. Second, molecular assessment of whether the primary tumor

has been producing cells capable of detaching may add crucial information beyond traditional prognostic factors such as tumor size or grade, which can be limited in their ability to predict CRC outcomes (12, 13). For example, while metastases are commonly thought of as being produced by large tumors, it has been shown that cancer cells can detach from small tumors (12). In short, whether a tumor has cancer cells detaching from it may be independent of its size, extent of locoregional spread, and therefore stage (except in the sense that any finding of distant metastases is by definition Stage IV). A tumor that appears to be Stage I or II may be generating undetected micrometastases. Thus, fully assessing the primary tumor after removal may require more measurements than the current diagnostics of lymph node evaluation and imaging.

2.3. Epithelial-Mesenchymal Transition

Given the importance of ascertaining whether a patient has metastatic disease, and the appeal of doing so in part in the primary tumor itself, what measurements of cancer cell ability to detach should be taken in the primary tumor? Basic research in recent decades suggests that an important mechanism of cancer invasion and metastasis is epithelial-mesenchymal transition (EMT). In EMT, epithelial cancer cells shed their epithelial characteristics and acquire a mesenchymal phenotype, which confers greater cell motility and ability to migrate out of the tissue of origin (2, 14). EMT could potentially play an enormous role in CRC metastasis because roughly 95% of CRC tumors are adenocarcinomas, which arise from epithelial cells (3). Indeed, about 80% of all cancer originates in epithelial cells (1). It has been suggested that patients develop metastatic disease after cancer cells in the primary tumor acquire the capacity to undergo EMT (2). Figure 1 illustrates the process:

Figure 1. Epithelial-mesenchymal transition and associated molecular markers (from Kalluri and Weinberg, 2009) (2)

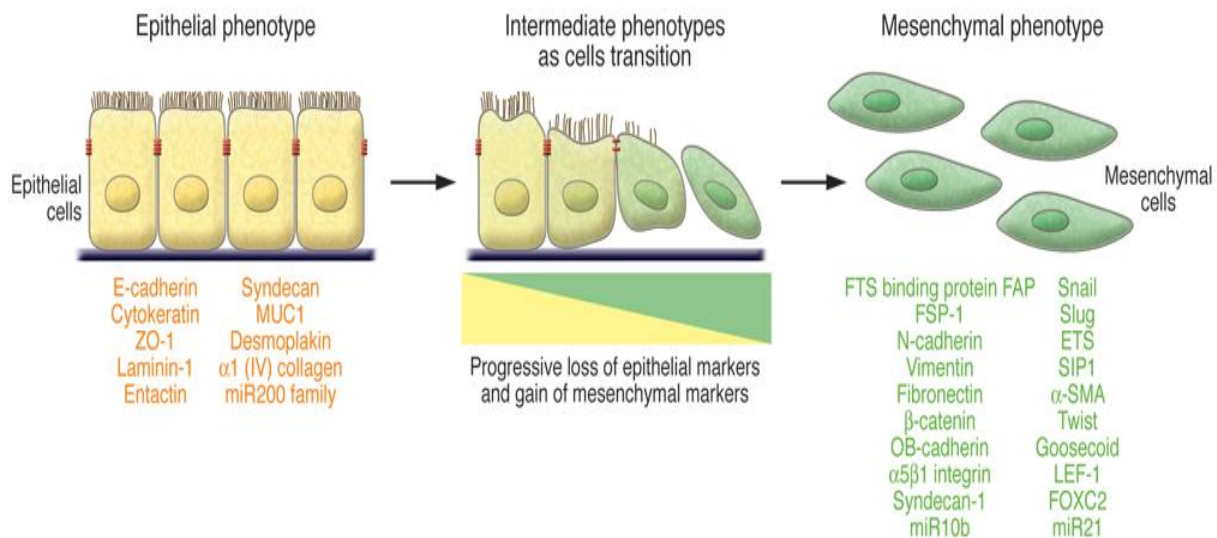


Figure 1 shows that a large number of molecular markers have been associated with EMT, and the lists are not exhaustive. This makes sense given that the model posits a global change in phenotype from one cell type to another. Therefore, expression levels of many markers should change as part of the process. Conceptually, EMT markers can be placed into three categories: 1) those that induce EMT (i.e. EMT inducers), such as the transcription factors Snail, Slug, and Twist; 2) markers of epithelial phenotype (i.e. epithelial markers), such as E-cadherin and cytokeratin; and 3) markers of mesenchymal phenotype (i.e. mesenchymal markers), such as N-cadherin and vimentin. For cells undergoing EMT, one would expect a decrease in expression of epithelial markers, as well as an increase in expression of both EMT inducers and mesenchymal markers (2, 14).

How can EMT markers be used to assess whether the primary tumor shows evidence of containing substantial numbers of cells capable of breaking off? Prior to undergoing EMT, the cancer cells in the tumor remain epithelial cells—albeit abnormal ones—that should exhibit typical epithelial-cell levels of EMT markers (high expression of epithelial markers, and low or

no expression of EMT inducers and mesenchymal markers). If the primary tumor, or at least part of it, begins generating metastases using the EMT mechanism, it is likely that a steady stream of cancer cells in the metastasis-generating portion will undergo the transition. Once the transition is complete for any one cell, the previously-epithelial cancer cell will have become a cancerous mesenchymal cell that could detach from the primary tumor and move away from it. At this point the cell would exhibit typical mesenchymal-cell levels of EMT markers (low expression of epithelial markers, and high expression of EMT inducers and mesenchymal markers).

While in transition, cancer cells will exhibit the “intermediate phenotypes” depicted in the middle of Figure 1. During this time they will be partially epithelial and partially mesenchymal, yet still attached to the primary tumor. Such cells may macroscopically appear to remain epithelial but show molecular expression of EMT markers that is abnormal for epithelial cells, such as low expression of E-cadherin or high expression of Snail. Finding cancer cells in the primary tumor with molecular expression levels indicative of an intermediate state between epithelial and mesenchymal phenotypes would suggest that the primary tumor is generating metastases. Crucially, such measurements could be taken in any primary tumor, regardless of its size, stage, or grade, thus satisfying the criterion set forth earlier that new diagnostics of metastatic disease measured in the primary tumor should not depend on the established measures of lymph node evaluation or imaging.

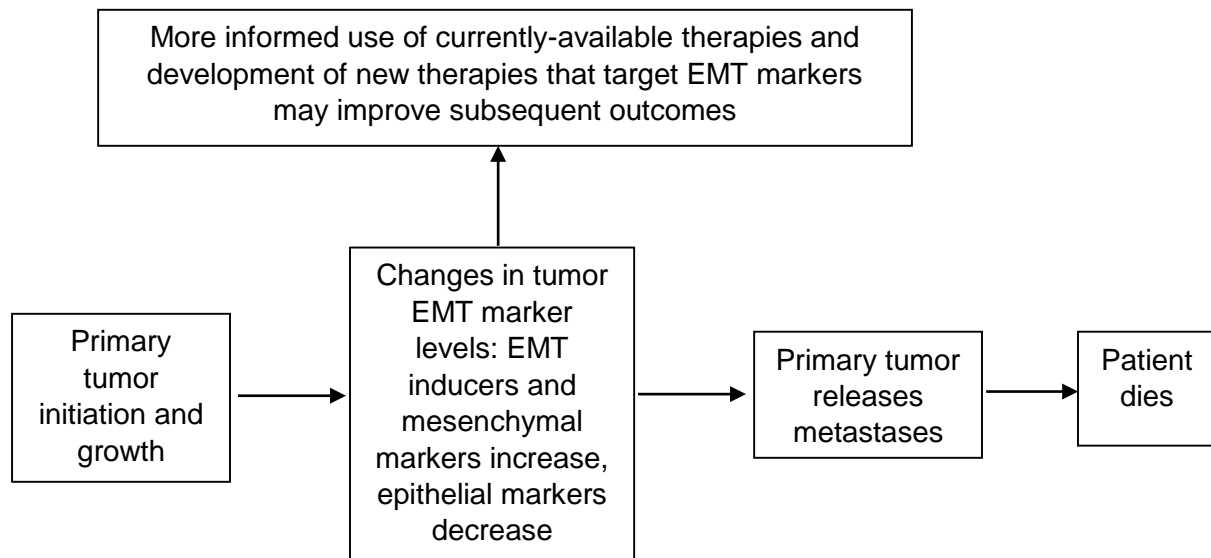
2.4. Public health significance

Given that EMT marker expression levels measured in primary tumor cancer cells could suggest whether the patient is at high risk for metastatic disease independent of any other measurements, there are two ways in which EMT markers could improve CRC outcomes. First, the markers can provide predictive information that can guide therapy. For example, suppose a

primary tumor is discovered that appears to be Stage II by current assessment procedures. The oncologist would likely decide based on current guidelines that surgery alone will suffice to treat the tumor and that no systemic chemotherapy is needed (15). However, if the tumor appears to be generating metastases via EMT marker measurements, the oncologist could instead prescribe chemotherapy in an attempt to destroy micrometastases that may be present but were not detected via imaging or lymph node assessment. Second, if EMT marker levels in primary tumor cancer cells are associated with patient outcomes, then new treatments could be developed that target the markers in an effort to hold metastasis in check. In this report, our concern is solely with the use of EMT markers to improve patient risk stratification.

Given how closely a CRC patient's outcome is tied to accurate clinical assessment of whether cancer cells have been detaching from the primary tumor, EMT markers hold tremendous promise as tools to reduce CRC mortality. The goal of the present research was to assess this promise by estimating associations between EMT marker levels in CRC primary tumors at the time of tumor removal and subsequent patient outcomes. More specifically, we sought to do this while designing measurements and analyses that could directly inform and facilitate translation of EMT markers to the clinic, where their public health impact stands to be realized.

Figure 2. Conceptual model of the role of EMT markers in determining CRC patient outcomes



2.5. Systematic review and marker selection

2.5.1. Systematic review of literature on EMT markers and CRC outcomes

There is no generally agreed upon histopathological definition of EMT (16).

Consequently, we developed our own approach to measure EMT for this dissertation. The first task was to select which EMT markers to measure in CRC specimens. To do so, we sought to identify original journal articles that measured EMT marker expression levels in clinical CRC tumor specimens and related those measurements to patient outcomes. On 26 October 2012, we searched PubMed, EMBASE, and BIOSIS using the same search terms for all databases (see Appendix A). No publication date or language limitations were applied to the search. The searches consisted of four groups of terms: EMT, tumor markers, outcomes, and colorectal cancer. To be captured, an item had to have at least one term from each group.

Our searches returned 545 abstracts. Removing 184 duplicates yielded 361 unique abstracts. I read the 361 abstracts and created a spreadsheet to record the following for each abstract: 1) which markers the study examined, 2) whether it looked at cell lines, animals, and/or

clinical tissue specimens, 3) whether the paper was a review, and 4) whether it looked at outcomes. Meeting abstracts, reviews, papers not written in English, and studies that only looked at animals and/or cell lines were excluded.

Based on the 361 abstracts, 122 appeared to have used clinical CRC specimens and therefore warranted closer examination. Inspection revealed that 46 were meeting abstracts and 1 was a case report. These 47 were excluded, leaving 75 original journal articles that examined EMT markers in CRC tissue specimens. I then created a second spreadsheet documenting the following information for each of the 75 papers: 1) sample size, 2) whether Kaplan-Meier survival curves were presented, 3) whether effect estimates such as hazard ratios were calculated, 4) whether correlations between EMT marker levels and other measurements were presented, 5) whether the percent of cases found to have positive expression of EMT markers was provided, 6) whether any measures of reliability were reported, and 7) which EMT markers were measured.

Thirty papers measured at least one EMT marker in clinical CRC primary tumors and evaluated the relationship between EMT marker expression and CRC patient outcomes via either Kaplan-Meier analyses or effect estimates. Between them, these papers measured dozens of markers or categories of markers (“categories of markers” meaning, for example, measurement of multiple micro-RNAs counts as simply “micro-RNA”). Because of the sheer diversity of EMT markers, we focused on 14 markers and marker categories prominently discussed in the EMT literature: E-cadherin, N-cadherin, Vimentin, Snail, Slug, Cytokeratins, Integrins, Fibronectin, Twist, ZEB1, ZEB2, Beta-Catenin, TGF-Beta, and Micro RNAs. We excluded 6 papers that did not measure at least one of these markers in clinical CRC specimens (17-22), leaving a final set of 24 papers that I evaluated for the markers of interest (23-46), though many of them measured other markers as well.

For each of the 24 papers, I recorded the following additional information for each EMT marker that the paper measured from our selection of 14 markers: whether marker expression was measured as protein and/or RNA, how the study defined positive expression for the marker, and whether the paper presented Kaplan-Meier analyses stratified by expression levels of that particular marker.

2.5.2. Marker selection criteria

In reviewing the literature, we considered seven criteria for judging whether a particular EMT marker might be useful as a clinical tool to assess whether a primary tumor has potentially been releasing cancer cells:

1. Biological Role: The role of the marker in the EMT mechanism should be well understood and critical.
2. Percent of subjects with positive expression: A marker that is positive for nearly 0% or 100% of patients is unlikely to provide much information about the prognosis for different patients, given how common metastasis is (47). Markers for which there are appreciable numbers of both marker-positive and marker-negative tumors are likeliest to be clinically useful.
3. Reliability: A useful EMT marker for clinical purposes will exhibit a high degree of reliability when measured in clinical CRC specimens in the same way. The reliability of EMT markers was difficult to assess because the literature provided little information about both inter-rater and intra-rater reliability. Such inconsistencies made comparisons between studies difficult, even between studies that measured the same marker using the same laboratory technique.
4. Validity: To determine the sensitivity or specificity of EMT markers, one would need to compare marker expression levels with a “gold standard” measure of whether the patient has

metastatic disease. It was not clear what could constitute such a gold standard in this context. Conventional diagnostics of metastatic disease—lymph node evaluation and imaging—are not satisfactory because the goal of using EMT markers is to identify those at risk of metastatic disease even when no metastases are detected by the other diagnostics. Correlation between EMT marker expression levels and results of the other diagnostics is a helpful, but not definitive, demonstration of the validity of an EMT marker for clinical purposes.

5. Association with patient outcomes: If the expression levels of a marker play a role in generating metastases and the marker is to serve as a clinical indicator of whether the patient is at high risk for metastatic disease, then the expression levels should be associated with patient time-to-death, that is, the length of time between primary tumor surgery and patient death. This can be assessed using Cox proportional hazards modeling and Kaplan-Meier curves stratified by marker expression levels.

6. Amount of prior data: Clinical utility ought to be supported by as many studies as possible, each of which includes as many subjects as possible. All else being equal, we had more confidence in markers the evidence for which was based on a greater number of subjects.

7. Ability to measure the marker: If a marker is difficult to measure accurately in a clinical setting, its utility is limited no matter how strong the evidence for it may be according to the other criteria described above.

2.5.3. Summary of systematic review findings

Because most studies measured markers only as protein, results refer to protein measurements unless noted otherwise.

Beta-Catenin: We found five studies that measured the mesenchymal marker beta-catenin in CRC tissue (26, 34, 37, 40, 42). The two that looked at survival by beta-catenin status

found no difference in outcomes between marker-positive and marker-negative subjects (34, 40). The only study that looked at effect estimates did not find any effect of beta-catenin measurements on outcomes (26). Percent of subjects with positive expression varied, particularly by location in the cell (nuclear, cytoplasmic, membranous), but was generally in the 40-50% range. Correlations between beta-catenin levels and location in tumor mass were inconsistent.

Cytokeratins: Of four studies that stained for cytokeratins, two (44, 46) used them only as a background stain. Of the other two studies, one (36) found that cytokeratin-8-positive subjects had better survival than cytokeratin-8-negative subjects. Cytokeratin-14-negative cases had better survival than cytokeratin-14-positive cases. The percent of subjects with positive expression was high for cytokeratin-8 (85%) and moderate for cytokeratin-14 (59%). The last study (31) found positive cytokeratin-7 expression in 9% of tumors and no effect on outcomes or difference in survival based on cytokeratin-7 expression.

E-cadherin: We found 15 papers that measured E-cadherin in clinical CRC tissue. Of these, three that measured protein (31, 34, 41) and one that measured RNA (39) failed to provide information on the number of tumors considered marker-positive. Among the other 11 studies, a wide range of definitions of marker-positive status were used (24-27, 29, 35, 37, 38, 42, 43, 45). Percent of subjects with positive expression varied but mainly fell between 30 and 70%. Three studies looked at survival by E-cadherin status, with all three finding poorer survival associated with reduced E-cadherin expression (24, 29, 43). Three studies looked at effect estimates, with two finding no effect of E-cadherin levels on outcomes (26, 27) and one finding that E-cadherin levels do affect patient outcomes (43).

Fibronectin: No paper measured fibronectin levels in clinical CRC specimens and related them to outcomes.

Integrins: Two papers looked at members of the integrin family of proteins, which are mesenchymal markers in the context of EMT. One study found 19% of subjects with positive expression for integrin alpha-5-beta-1 and 88% of subjects with positive expression for integrin alpha-3-beta-1 but did not look at the association of either protein with outcomes (45). The other, much larger study found 37% of subjects with positive expression for integrin alpha-v-beta-6 and clear differences in survival and effect estimates for the protein's relationship with outcomes (23).

Micro-RNAs: Two papers measured micro-RNAs (miR), but one (39) did not provide information on how many tumors were considered marker-positive. The other study reported that subjects with high expression of miR-19b and miR-194 had shorter survival than those with low expression (32).

N-cadherin: Two studies measured the mesenchymal marker N-cadherin in clinical CRC specimens. One of them that included 10 subjects did not find positive N-cadherin expression in any of their tumors, and did not look at the relationship of N-cadherin with outcomes (37). The other study found 44% of subjects with positive expression and, while it did not look at survival by N-cadherin status, calculated effect estimates and found no effect of N-cadherin on outcomes (26). However, in the latter study, the 193 subjects were divided into training and testing sets before effect estimates were calculated, thus reducing the power for each estimate.

Slug: Three studies measured the EMT inducer Slug, though one of them (30) that measured RNA failed to provide information on the number of tumors considered marker-positive. Of the other two studies, one with a sample size of 10 patients found 30% of subjects

with positive expression in primary tumors, and observed no difference in survival by Slug status (37). The other study found 37% of subjects with positive expression (43). Slug-positive patients in this last study had poorer survival than Slug-negative patients, and using hazard ratios, the authors concluded that Slug was an independent prognostic factor of outcomes.

Snail: Four studies measured the EMT inducer Snail. Three found 40-55% of subjects with positive expression (26, 27, 37) and one found 79% positive (28), though the studies used a variety of definitions of Snail-positive status. The two studies that looked at survival both found worse survival in Snail-positive subjects than Snail-negative subjects (28, 37). The two studies that did not look at survival did look at effect estimates, and each obtained mixed within-study results (26, 27).

TGF-beta: One study measured a member of the TGF-beta class of EMT inducers, namely TGF-beta-R2. It found almost 90% of subjects with positive expression and no difference in survival by TGF-beta status (40).

Twist: Four studies looked at the EMT inducing Twist family in clinical CRC specimens. One study (30) that measured mRNA of Twist1 found that 86% percent of subjects showed positive expression and that those with positive expression had worse survival than those with negative expression. Survival differences were especially large among early-stage subjects. Via effect estimates, Twist levels had an effect on outcomes. The other three studies measured protein. One with 10 subjects found 100% of subjects with positive expression but did not look at outcomes (37). The other two studies found roughly 50% of subjects with positive expression and did not look at survival (26, 27). Both calculated effect estimates, with mixed results.

Vimentin: Of the four studies that measured this mesenchymal marker, one (39) measured mRNA and did not provide information on how many specimens were considered

marker-positive. The other three studies measured protein. Of these, two (31, 37) found 0% of subjects with positive expression—despite using different definitions of marker-positive status—and another (38) found 9% of subjects with positive expression.

ZEB1: In the three studies that measured the EMT inducer ZEB1, two (41, 44) did not report percent of subjects with positive expression. The other found 29% of subjects with positive expression and that ZEB1-positive patients had much shorter average survival than ZEB1-negative patients (31 months vs. 67 months, respectively) (38).

ZEB2: Two studies measured the EMT inducer ZEB2. One with 10 subjects found 90% of subjects with positive expression and did not look at survival or effect estimates (37). The other study found that 48% of the tumors were ZEB2-positive at the tumor invasion front and 41% were ZEB2-positive at the tumor center (33). This study reported that 73% of the primary tumors had greater ZEB2 expression at the invasion front compared to the tumor center. ZEB2-positive patients had poorer survival than ZEB2-negative patients. In terms of effect estimates, ZEB2 levels at the invasion front were a predictor of outcomes, but ZEB2 expression at the tumor center was not.

2.5.4. Marker selection conclusions

Of the three kinds of EMT markers—epithelial markers, mesenchymal markers, and EMT inducers—it was not clear that one kind was likelier than the others to be a clinically useful predictive tool. Given budgetary constraints, we could only afford to measure a few markers. As a compromise between these considerations, we decided to measure three markers, namely, the most promising marker from each category based on prior literature.

Epithelial marker:

Options: cytokeratin, E-cadherin

Discussion: Effectively, only one study looked at cytokeratin. Although it provides some support for cytokeratin in terms of percent of subjects with positive expression and survival by cytokeratin status, this is swamped by the large number of studies on E-cadherin. The overall trend of the results suggests that E-cadherin is a promising marker in terms of percent of subjects with positive expression and the marker's association with outcomes. Furthermore, its biological role in EMT is clear and important, and it is by far the most commonly-studied EMT marker. Also, E-cadherin antibody had already been optimized by the UNC Translational Pathology Laboratory that performed the benchwork for the project, making E-cadherin more cost-effective to measure than cytokeratin.

Choice: E-cadherin

Mesenchymal marker:

Options: beta-catenin, fibronectin, integrins, N-cadherin, vimentin

Discussion: There was no relevant data about fibronectin. The data strongly suggest that beta-catenin is not a good marker based on lack of association with outcomes and an inconsistent percent of subjects with positive expression. Likewise, the data do not support vimentin based on its percent of subjects with positive expression of 0% or close to it, as well as the lack of information about whether it is associated with outcomes.

The two best candidates are N-cadherin and integrins, specifically integrin alpha-v-beta-6. N-cadherin is appealing because it plays an important biological role in the EMT mechanism (part of "cadherin switch" with E-cadherin) and in mesenchymal cell motility. In the one meaningful study that measured N-cadherin (the other study being tiny), the percent of subjects with positive expression was right where one would want it to be (44%). The effect estimates of

its association found nothing, but the sample sizes involved in the calculations were not large (roughly n=100).

Two of the three integrins on which we have any data are not serious candidates; only integrin alpha-v-beta-6 is. The percent of subjects with positive expression and clear association with outcomes in the one large study provide stronger support than N-cadherin received from its one relevant study. Integrin alpha-v-beta-6 and N-cadherin have similar biological roles as membrane-bound proteins that help mesenchymal cells move through their environment. All told, integrin alpha-v-beta-6 is the stronger option. However, the prior study that measured this integrin in CRC primary tumors did so by measuring only the beta-6 sub-unit and then making an inference from that to the alpha-v-beta-6 complex (23). To make our study comparable, we decided to measure only the beta-6 sub-unit as well.

Choice: Integrin beta-6

EMT inducer:

Options: Slug, Snail, TGF-beta, Twist, ZEB1, ZEB2

Discussion: TGF-beta was dismissed because the one study that measured it gave clear evidence that it is not a good marker. The evidence was similar and moderately supportive for the other five options. Each was measured in 2-4 studies, with a mixture of results suggesting that the marker is and is not associated with outcomes. Considering all of our selection criteria, Snail seemed the best choice because it was measured in the greatest number of studies, generally had ideal percentages of subjects with positive expression, and had as much evidence showing an association with outcomes as any of the other candidates.

Choice: Snail

CHAPTER 3. METHODS

3.1. Subject selection and enrollment

3.1.1. Study population

The Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) study was undertaken by a consortium of seven teams of investigators across the United States (48). Their aim was to study the impact of the characteristics of healthcare delivery systems, patient characteristics, and patient beliefs on cancer outcomes. The study was a population-based, case-only prospective cohort of lung and colorectal cancer patients. Subjects were enrolled between 2003 and 2006. Every site collected patient surveys, physician surveys, and medical records data. Subjects completed surveys at baseline, 12 months after baseline, and 5 years after baseline (49). Upon enrollment, each subject identified a proxy respondent who completed the next follow-up survey in the event of the subject's death or inability to complete the survey.

The patient surveys asked questions on a range of topics, including demographics, treatments received (surgery, radiation, chemotherapy), quality of life, and health history and behaviors. Tumor specimens were collected only among North Carolina subjects. Since the main exposures of interest in this dissertation were primary tumor expression levels of EMT markers, our study sample was limited to NC subjects for whom tumor tissue was available.

The NC site was based at the University of North Carolina at Chapel Hill (UNC). It enrolled 990 incident colorectal cancer (CRC) cases during the CanCORS enrollment period (50) that were drawn from 33 counties in eastern and central NC (51). These subjects constitute a

population-based, case-only prospective cohort of incident CRC cases from the 33 counties in 2003-06.

3.1.2. Subject identification, recruitment, and enrollment

During the enrollment period, incident CRC cases were identified using a rapid-case ascertainment protocol through the NC Central Cancer Registry, to which all new cases of CRC in the study region must be reported (52). Of 1,899 nominally eligible case reports, 350 were found to be ineligible and 43 could not be contacted due to physician refusal. Of 1,506 eligible, contactable cases, 326 refused to participate, 77 could not be reached, and 85 were not capable of giving consent and had no proxy to participate in their stead. This left 1,018 cases who agreed to be interviewed, provide medical record data, or both, giving a response proportion of $1,018/1,506 = 67.6\%$. Of the 1,018 eligible subjects who agreed to participate, 28 did not end up enrolling for unspecified reasons, giving the final UNC study sample of 990.

Whenever possible, the patient survey was administered to the patient or to a proxy 4 months from the date of diagnosis. Information from all patients selected for contact was included in the data from the cohort, regardless of the survival status of the patient. For patients who died by the time of initial contact or before the time of a scheduled interview, attempts were made to interview an eligible proxy or surrogate.

If the patient was not able to complete an interview, a surrogate had to be interviewed instead. Eligibility criteria for surrogates were: at least 18 years old; patient gave consent for investigators to speak with the surrogate or the patient is dead; the patient indicated that the surrogate knows best how the patient has done since diagnosis; and the surrogate had to reside in one of the 33 counties in the study region.

Of the 990 subjects enrolled in NC, 506 subjects (51%) provided tumor specimens. The major reasons for not obtaining tumor tissue from all study subjects were that subjects did not consent to donate tumor tissue, and among those who did consent to give tissue, adequate tumor blocks for research were not always available after surgery. Our study sample for the dissertation was a subset of the 506 NC subjects who provided tumor specimens.

3.1.3. Subject eligibility criteria

Inclusion criteria were as follows (52):

- a) Sex: males and females eligible
- b) Age: 21 years or older at time of diagnosis
- c) Race: no exclusion based on race/ethnicity
- d) Residence: At the time of diagnosis and initial contact, patient had to be a resident of one of the 33 counties constituting the study region.
- e) Language: Able to complete study interviews in English, Spanish, or Chinese
- f) Cancer Diagnosis: A histologically-confirmed diagnosis of colorectal cancer on 1 January 2003 or later. All stages of disease except in situ were eligible.

3.2. Specimen collection

3.2.1. Collection at clinics

Tumor specimens were collected at the clinics across NC that performed cancer surgery on enrolled subjects. Portions of these tumors were transported to UNC. Tumor samples were fixed in formalin and embedded in paraffin after surgical removal (53). Since this initial tissue preparation was performed at different hospitals and clinics throughout the study region, there

may have been variation in terms of fixation protocols, time from tumor removal to fixation, storage conditions at the clinic, and technician skill.

During the CanCORS data-collection period, investigators were informed via the rapid-case ascertainment system of new, nominally-eligible incident cases of CRC within the study region. Investigators checked the pathology report of each case to confirm eligibility (54). For example, they ensured that the diagnosis was for invasive carcinoma since in situ tumors were not eligible. After confirming patient eligibility, investigators requested tumor blocks from the hospital or clinic that performed the patient's cancer surgery. The goal was to obtain two blocks of tumor tissue and two blocks from the tumor margins per patient, the latter including adjacent normal tissue (55). Biopsies and metastases were not requested. When the primary tumor had spread to an adjacent organ beyond the colon or rectum (but without metastasizing), investigators obtained normal tissue from the adjacent organ.

3.2.2. Specimen preparation and storage at UNC

Upon receiving a patient's tumor blocks at UNC, investigators cut the tissue to prepare histology slides, which were then stained with hematoxylin and eosin (53). A trained pathologist verified which parts of a stained slide were normal or tumor. The tissue was then incorporated into a tissue microarray (TMA) along with tissue samples from other subjects. Using the pathologist's indications on the stained slides as to which parts of tissue samples were normal and which were tumor, a TMA technician punched cores of normal and tumor samples from a given patient's blocks. Most patients had six cores on a TMA: 3 cores of normal tissue and 3 cores of tumor tissue. Some patients had more or less than 3 cores of a given tissue type. Tissue slides for the dissertation were prepared from the TMAs.

Once prepared, TMAs were stored at 4 degree Celsius (54). Slides cut from the TMA but left unstained for future use were stored at -80 degrees Celsius. Any unused tissue was returned to the hospital or clinic from which it was obtained.

3.2.3. Control specimens

Tissue specimens for positive and negative controls were obtained from the Funkhouser Laboratory in the UNC Department of Pathology and Laboratory Medicine. Positive control specimens were colorectal tumor tissue and negative control specimens were normal (tumor-free) colorectal tissue. Controls came from UNC patients who were not enrolled in CanCORS.

3.2.4. Sample size

The specimens of the 506 CanCORS participants who provided CRC tissue samples were distributed across 56 TMAs stored at the Keku Laboratory. To minimize costs and use of specimens, the dissertation committee recommended staining slides from a subset of subjects large enough to provide a reasonable chance of observing a substantial hazard ratio estimate. If little variation was found for a particular marker in a sufficiently large subset, this could suggest that the expression levels of the marker likely are not able to distinguish between tumors that are behaving differently from each other (47), assuming the sample includes a mixture of tumors with high and low levels of cancer cell detachment. Logically, this would imply that it was not worthwhile to measure a marker with a narrow expression distribution in more subjects.

Many of the TMAs containing CanCORS subject specimens also contain specimens collected for another study (the Rectal Study). Some subjects were enrolled in both CanCORS and the Rectal Study; such individuals were treated as CanCORS subjects. Random selection of TMAs for the subset might have resulted in selecting TMAs containing only one or two CanCORS subjects, which would have been an inefficient use of resources. For greater

efficiency, we selected TMAs from among those containing the greatest number of CanCORS subjects so that a minimal number of TMAs had to be stained to provide the desired sample size.

Table 1 shows the results of sample size calculations using PROC POWER in SAS. The parameters used in the calculations were based on the E-cadherin, Integrin, and Snail results from the systematic literature review (see Section I.E). From the literature overall, a 30% proportion of patient deaths within 5 years of surgery in the unexposed (those with non-EMT-like marker expression levels) seemed reasonable and the analysis of the subset aimed to achieve 80% power. As Table 1 indicates, this meant the analysis should have a sample size of at least 165 subjects.

Table 1. Sample size calculations for initial selection of subjects*

% outcomes (deaths) In unexposed (those with Non-EMT marker levels)	Statistical Power		
	70%	80%	90%
20%	190	240	325
30%	130	165	220
40%	100	130	170

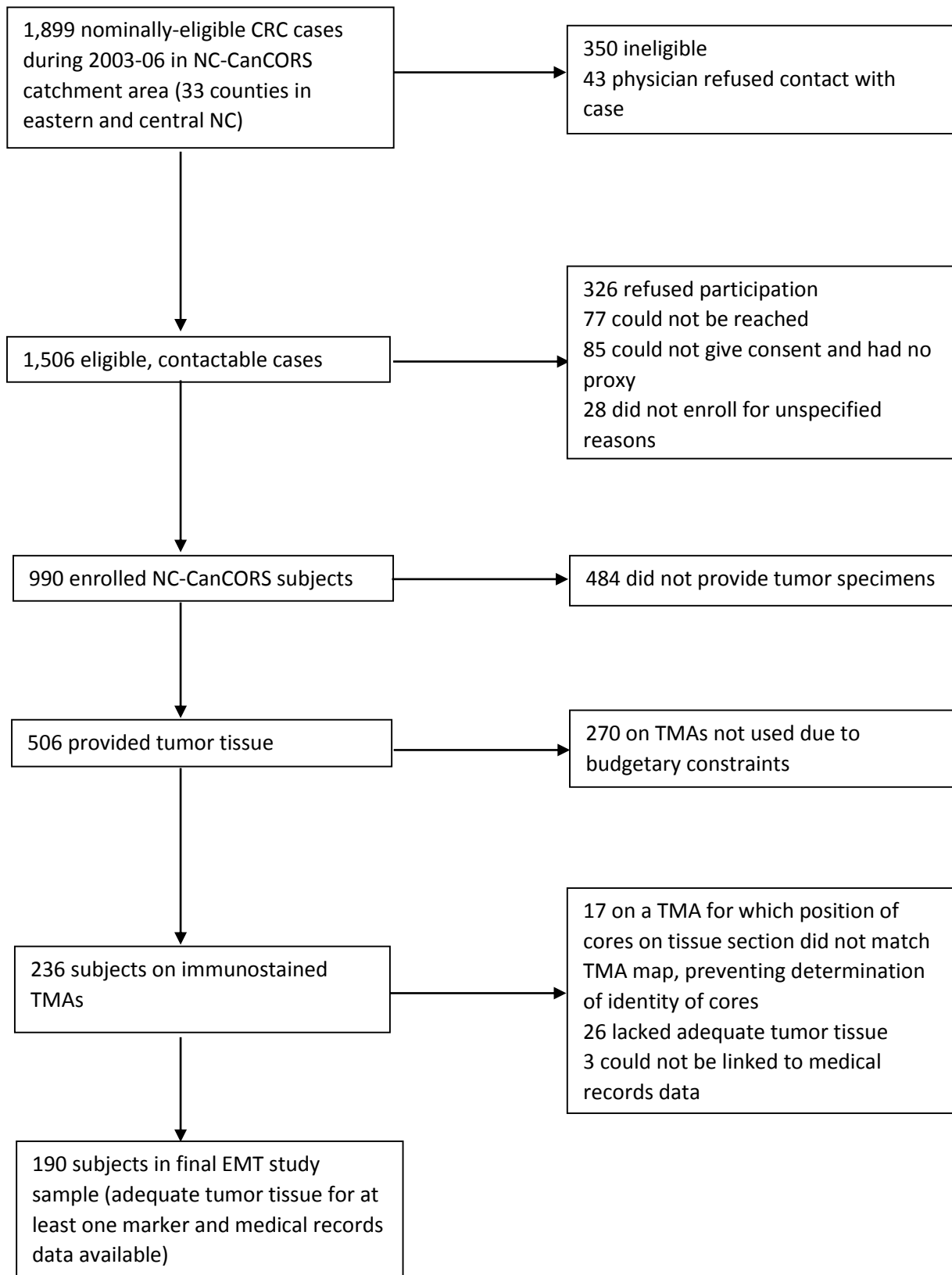
*Total sample size for 70%, 80%, or 90% power to detect a hazard ratio of 2.00 across 5 years of follow-up, for 40% exposed (i.e. EMT-like marker levels) and with varying percentages of outcomes (deaths) in unexposed (i.e. non-EMT-like marker levels) over the 5-year window, with a 2-sided alpha of 5%.

In anticipation of possible laboratory error when handling the tissue samples, specimens from 236 subjects across 13 TMAs with large numbers of CanCORS subjects were prepared for staining. One TMA with specimens from 17 subjects could not be used because the position of cores on the tissue section did not match the map for the TMA, making it impossible to determine the identity of the cores. This reduced the available sample size to 219 subjects. To be included in the study sample, subjects had to have at least one core of tumor tissue successfully stained for one marker, with that core having at least 50 epithelial cells and

unambiguous histology. From the 12 TMAs, we excluded 26 subjects lacking adequate tumor tissue and an additional 3 subjects who could not be linked to medical records data, yielding a final study sample of 190 subjects.

Figure 3 shows the overall flow of subject inclusion from nominally-eligible CRC cases to the final study sample.

Figure 3. Flow of patient eligibility and inclusion



3.3. Specimen staining

3.3.1. Acquisition of antibodies for immunohistochemistry

The following antibodies were purchased:

- a) E-cadherin: Mouse monoclonal ready to use (RTU), clone 36B5
(cat #PA0387) from Leica Microsystems Inc. (Norwell, MA)
- b) Integrin beta-6: Goat polyclonal (sc-6632) from Santa Cruz
Biotechnology (Dallas, Texas)
- c) SNAIL1: Goat polyclonal (ab53519) from Abcam (Cambridge, MA)

3.3.2. Immunohistochemistry staining procedures

Immunohistochemistry (IHC) was performed at the UNC Translational Pathology Laboratory (TPL) using the Bond fully-automated slide staining system (Leica Microsystems Inc., Norwell, MA). Slides were deparaffinized in Bond Dewax solution (AR9222) and hydrated in Bond Wash solution (AR9590). Antigen retrieval was performed at 100°C for Integrin beta-6 and Snail (for 20 minutes or 30 minutes, respectively) in Bond-epitope retrieval solution 1 at pH 6.0 (AR9961) and for E-cadherin for 20 minutes at 100°C in solution 2 at pH 9.0 (AR9640). After pretreatment, anti-E-cadherin was applied for 15 minutes, anti-Snail (1:200) for 30 minutes and anti-Integrin beta-6 (1:100) was applied for 1 hour.

Detection of Snail and Integrin beta-6 was performed using the Bond Intense R Detection System (DS9263) supplemented with the LSAB+ kit (DAKO, Carpinteria, CA). E-cadherin detection used the Bond Polymer Refine Detection System (DS9800). Stained slides were dehydrated and cover-slipped. Positive and negative controls (no primary antibody) were included for each antibody. All assays were single-marker (i.e. no multiplex assays).

Stained slides were digitally imaged at 20× magnification using the Aperio ScanScope XT (Aperio Technologies, Vista, CA). Digital images were stored in the Aperio Spectrum Database. Example images of staining and annotations are provided below in Figures 4-6.

Figure 4. Immunohistochemistry examples for positive and negative staining for each of E-cadherin, Integrin beta-6, and Snail

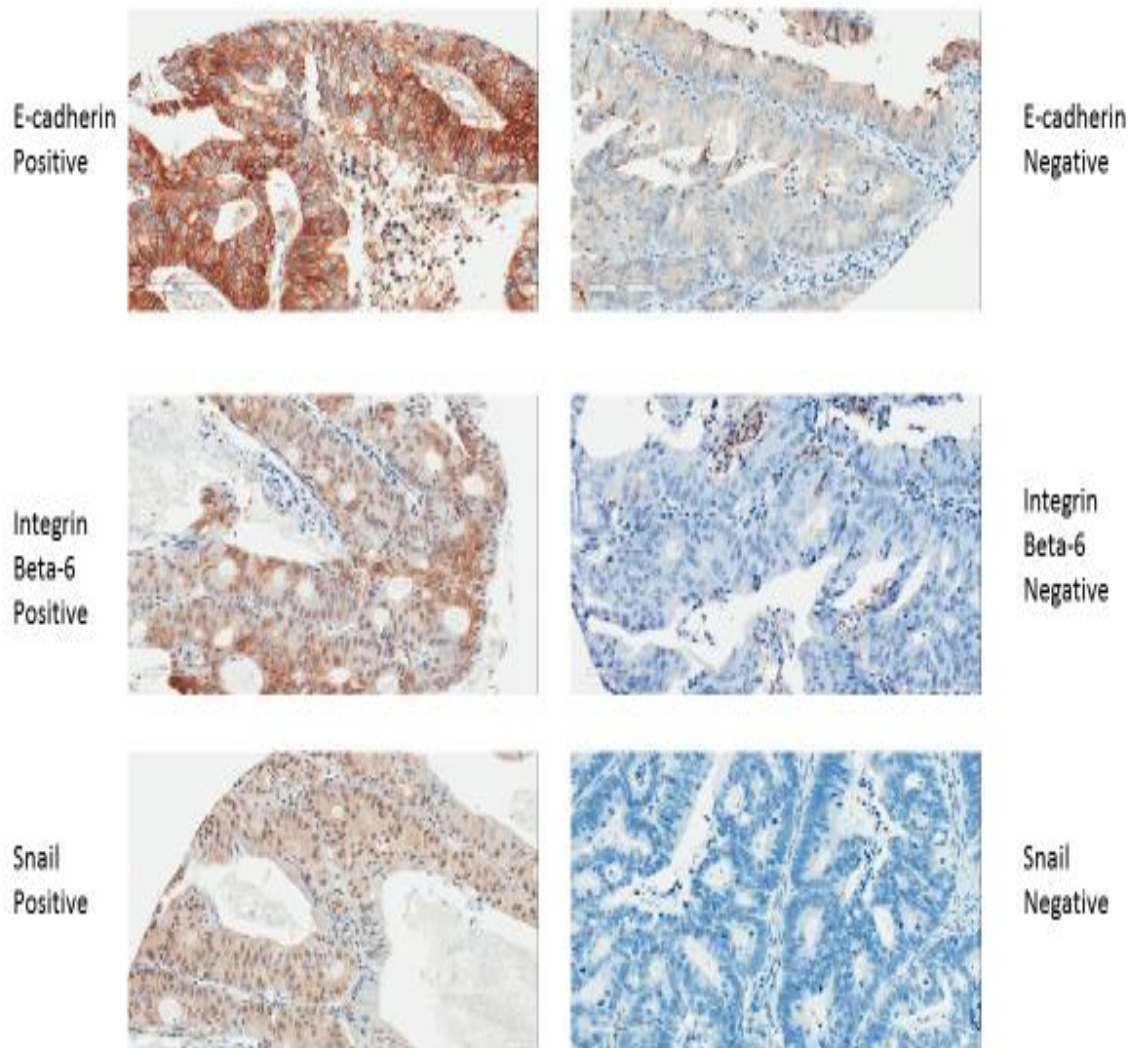
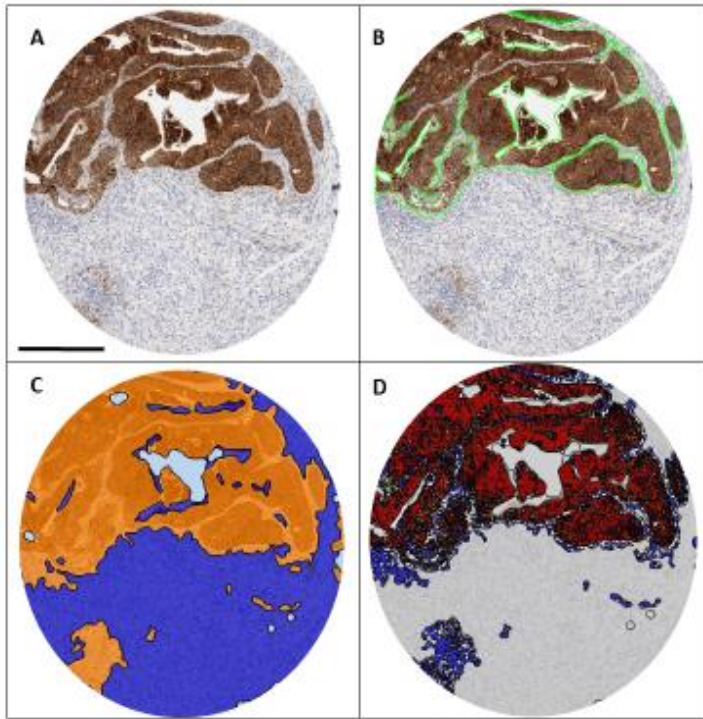
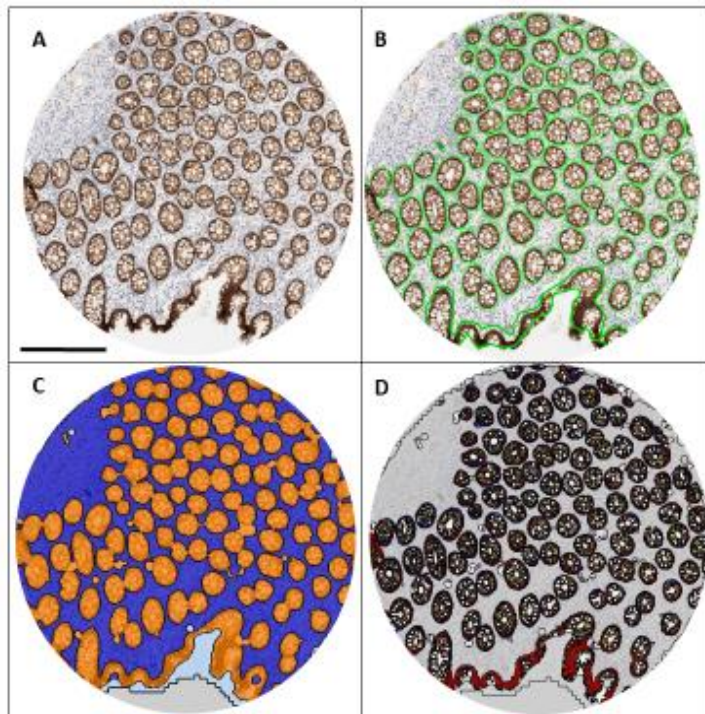


Figure 5. Illustration of Tissue Studio analysis of colon tumor tissue



(A) Original image of TMA core stained for E-Cadherin. Bar = 300um. (B) Green lines show manual annotation of tumor areas on image (C) Mark-up of image by Tissue Studio Composer. The algorithm was trained to differentiate between epithelial and stromal regions. Orange highlighted areas are enriched in epithelial cells and closely match the regions that were manually annotated in (B). Blue highlighted areas are enriched for stromal cells. Only the epithelial enriched regions were analyzed. (D) Mark-up of analysis results for tumor tissue. Blue= IHC negative, yellow = 1+, orange = 2+, red = 3+ (staining intensity). This tumor tissue core has high concentrations of E-Cadherin, with most of the staining classified as 3+ in intensity.

Figure 6. Illustration of Tissue Studio analysis of normal colon tissue



(A) Original image of TMA core stained for E-Cadherin. Bar = 300um. (B) Green lines show manual annotation of colon crypt areas on image (C) Mark-up of image by Tissue Studio Composer. The algorithm was trained to differentiate between epithelial and stromal regions. Orange highlighted areas are enriched in epithelial cells and closely match the regions that were manually annotated in (B). Blue highlighted areas are enriched for stromal cells. Only the epithelial enriched regions were analyzed. (D) Mark-up of analysis results for normal tissue. Blue= IHC negative, yellow = 1+, orange = 2+, red = 3+ (staining intensity).

3.4. Digital image annotation and analysis

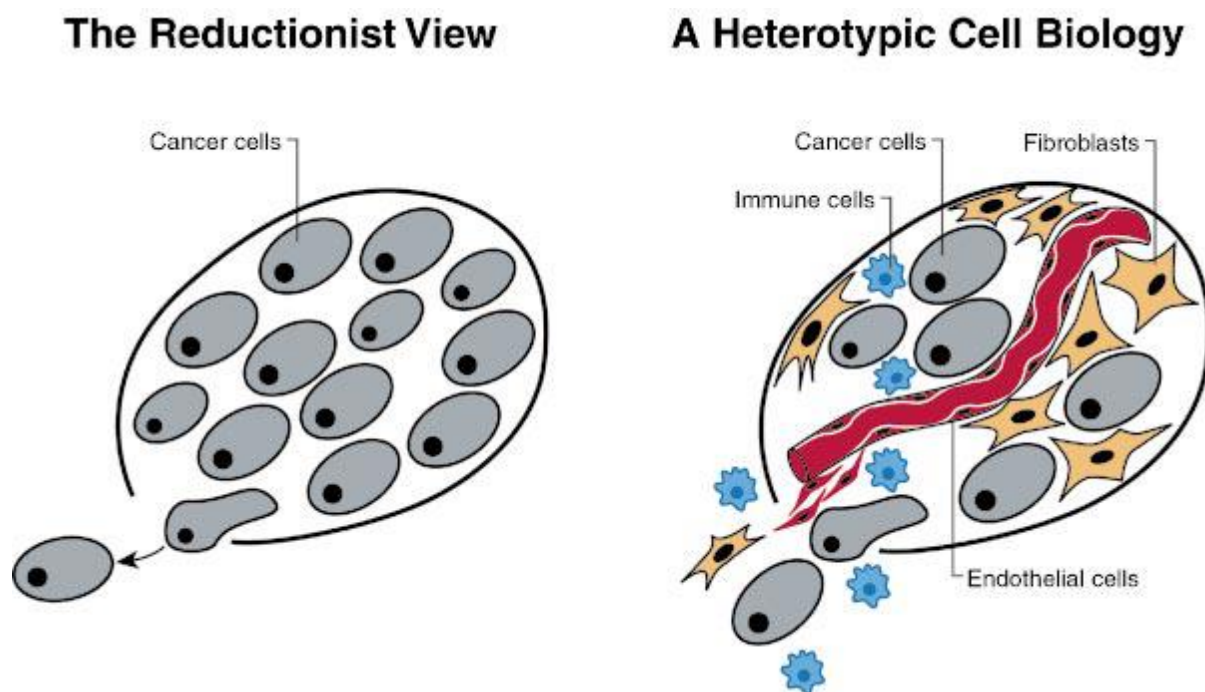
3.4.1. Annotation rationale and goal

Tumors contain not only cancer cells but numerous other kinds of cells, including fibroblasts, endothelial cells, cancer stem cells, and immune cells (Figure 7) (11). Metastases form from a subset of the cancer cells within a primary tumor and not from the other cell types. However, EMT markers may be expressed—or not expressed—by all types of cells found in a tumor. Thus, when digital slide images are scored by computer, results could be misleading if the entire image is scanned when not all of the cells are cancer cells. Since only EMT marker

expression of cancer cells is relevant to determining whether the tumor has been producing cells capable of acting as metastases, images must be marked (i.e. annotated) prior to scanning to restrict the computer's analysis of marker expression to cancer cells.

The goal of image annotation was to mark images prior to scoring such that all cancer cells in the image would be included in the analysis of marker expression while all other cells would be excluded.

Figure 7. Reductionist and heterotypic representations of cancer tumors (from Hanahan and Weinberg, 2000) (11)



3.4.2. Image annotation and scoring procedures

Computer algorithms annotated and scored every eligible tissue core to obtain continuous marker expression data. Continuous expression data were sought because marker expression is inherently continuous. In addition, continuous data provided maximum flexibility in identifying a clinically-useful cut point to define dichotomous marker expression status, the latter reflecting

the binary nature of treatment decisions. We used approximately 65 cores originating from two TMAs for algorithm training and automated-analysis validation.

Definiens Composer Technology (Tissue Studio version 2.1.1 with Tissue Studio Library version 3.6.1; Definiens Inc., Carlsbad CA) was used to identify regions enriched in epithelial cells in IHC-stained TMA cores. To detect differences in cell shape and tissue structure, we developed two Composer algorithms per marker—one for cores containing normal adjacent tissue and a second for tumor cores—as both types of tissue were present on each TMA.

After Composer training, we developed two Tissue Studio scoring algorithms (“solutions”) per marker. Different Composers were used to identify epithelial cell regions but identical settings were used to determine relative DAB staining intensity. The Composer_MarkerArea Solution was used to detect E-cadherin staining. This algorithm gave average intensity readings for each core on a continuous scale of 0-3 and included both membrane and cytoplasmic E-cadherin staining. Integrin beta-6 membrane staining was detected using the Composer_Nuclei_Membranes&Cells Solution. We used the Composer_Nuclei(Positive_vs_Negative) Solution to detect nuclear Snail expression. Integrin beta-6 and Snail were measured as core percent positive cells (Integrin) or nuclei (Snail) on a continuous scale of 0-100.

To evaluate the reliability of computer annotations, I used Aperio ImageScope (version 11.2; Leica Biosystem, Buffalo Grove, IL) to manually annotate the same 65 cores per marker that were used to optimize Tissue Studio solutions. Included regions were marked with the positive pen and excluded regions were marked with the negative pen. I remained blind to patient and tumor characteristics while annotating. This was accomplished by keeping TMA data and identification numbers separate from other CanCORS data and identification numbers

until after annotation was finished. As part of the original CanCORS study design, TMAs and the rest of the subject data were assigned different identification numbers and could only be brought together by use of a linking file. I did not have access to a linking file until after annotating the specimen images, which kept me from connecting any specimen images or data with other subject data.

The manually-annotated cores—considered the gold standard for digital separation of tissue types—were then analyzed using appropriate Aperio scoring algorithms (Membrane v9 algorithm for E-cadherin and Integrin beta-6, Nuclear v9 algorithm for Snail). Automated scores obtained via manual and automated annotation produced Pearson correlations of 0.91 for E-cadherin, 0.88 for Integrin beta-6, and 0.94 for Snail. Having verified the accuracy of the annotation algorithms for the three markers, all 12 TMAs stained for E-cadherin, Integrin beta-6 and Snail were analyzed (36 slides in total).

Subjects typically had multiple cores available of a given tissue type (tumor or normal). To assign an expression value for each subject by marker and tissue type, we handled replicate cores in two ways: first, as a weighted average of cores, and second, by assigning the expression value of the subject's "worst" core as the marker expression value. For weighted averages, the weights were area analyzed for E-cadherin, number of cells for Integrin beta-6, and number of nuclei for Snail. The worst score by tissue type was assigned as the lowest core average intensity for E-cadherin, as the highest core percent positive cells for Integrin beta-6, and as the highest core percent positive nuclei for Snail.

After observing unexpected results when comparing mean Snail expression in tumor tissue to normal tissue, we trained two additional Snail scoring algorithms on a different TMA than the one used to develop the original scoring algorithms: a second nuclear algorithm and a

whole-cell algorithm that scored Snail expression in any part of a cell. These additional Snail scoring algorithms measured expression on a continuous percent positive scale and were applied to all 12 TMAs. The results of the two additional Snail algorithms were qualitatively similar to those from the original nuclear algorithm. Thus, for analysis we used only Snail expression values based on the original algorithm.

Examples of manual and automated annotations are given for tumor tissue (Figure 5) and normal tissue (Figure 6).

3.4.3. Linkage of marker data with main CanCORS data

After image analysis, the CanCORS database manager (Christopher Martin) provided a file linking TMA identification numbers with CanCORS identification numbers. Marker data and general CanCORS data (surveys, medical records) were merged via the linking file to provide a complete set of variables. CanCORS subjects for whom EMT markers were not measured were dropped, leaving a final working dataset with all variables and containing only the subset of CanCORS subjects in whose tumor specimens EMT markers were measured.

3.5. Assessment of reliability

The public health significance of measuring EMT markers in primary tumors lies in their potential clinical use as tools to stratify patients according to risk for worse outcomes. Clinical usefulness depends in part on marker reliability, which is how closely repeated measures of the marker in the same specimen match each other. Two kinds of reliability could be considered. First is inter-rater reliability, the consistency of measurements when different people score the same specimens. Second is intra-rater reliability, the consistency of measurements when the same person scores a specimen multiple times.

In this research we evaluated a form of inter-rater reliability. In theory, this could be applied at each of two steps: annotation and scoring. The only scoring that was performed was the automated scoring described above. Thus, computer programs were the only scoring rater and no test of scoring reliability was possible given the lack of manual scoring.

We did, however, assess the inter-rater reliability of annotation. The primary rater in this case was the set of computer algorithms developed to perform automated annotation as described above. Such annotation was applied to all eligible cores in the EMT dataset. I served as secondary rater with my manual annotations of a subset of eligible cores. The roughly 65 cores that were annotated by computer and also manually constituted our reliability sample.

After both sets of annotations were complete, the same automated scoring algorithms were applied to each set to generate image scores. These scores—the automated annotation of a given image paired with the manual annotation of the same image—could then be compared by Pearson correlations to measure the reliability of the automated annotations.

As stated earlier, these correlations were 0.91 for E-cadherin, 0.88 for Integrin beta-6, and 0.94 for Snail. These high correlations suggest that the automated annotations were excellent approximations of the gold-standard manual annotations.

3.6. Statistical analysis for time-to-event analysis

3.6.1. Overview

The clinical and public health benefit of an EMT marker would be as a diagnostic of metastatic disease in addition to lymph node evaluation and radiologic imaging. Introducing a marker into clinical practice would require first determining whether it provides useful information beyond what is provided by established prognostic and predictive factors. The most important aspect of this is whether marker expression levels are associated with patient mortality

in the form of length of time from surgery to death, independent of relevant confounders.

Addressing this issue for our selected EMT markers was the goal of our time-to-event analysis.

While marker expression is inherently continuous, an EMT marker would be used to make a clinical decision, which is inherently binary. Therefore, the association of interest is between dichotomous marker expression and time-to-death, not the association between continuous marker expression and time-to-death. However, the best way to define dichotomous marker expression is to initially measure continuous marker expression for all subjects, then identify a clinically-informative cut point along the continuum to classify each individual as marker-positive or marker-negative.

3.6.2. Definitions

Primary Exposures of Interest: EMT marker protein expression levels in the biologically-relevant portions of cancer cells in CRC primary tumor tissue at the time of tumor removal surgery. “Biologically-relevant portions of cancer cells” meant membrane expression for E-cadherin and Integrin beta-6, as well as nuclear expression for Snail. Marker expression is inherently continuous and was initially measured as such. For translational purposes, the relevant form of marker expression is dichotomous (marker-positive versus marker-negative) based on a clinically-informative cut point along the original continuum. This is because the markers have to be used to make clinical decisions, which are inherently binary. Thus, the primary exposures of interest were dichotomous tumor expression variables for each marker.

Outcome of Interest: Time from tumor removal surgery until all-cause mortality. Vital status for all subjects was verified using the Social Security Death Index on 4 May 2010, providing at least 42 months of follow-up observation per individual. The Index records all

deaths in the United States, though typically individuals are not entered until two years after their deaths.

Censoring: Subject follow-up time was administratively censored at 5 years after surgery.

Truncation/Immortal Person-Time: An individual could not participate in this analysis without providing tumor tissue. Furthermore, specimens could not be provided to the study until tumor removal surgery. Therefore, all time from a subject's birth until the date of tumor removal was immortal time and was not included in the analysis. This meant we excluded individuals who died before tumor removal. The "origin" time point for each subject was the date of tumor removal.

Loss to Follow-Up and Withdrawals: A subject was lost to follow-up if vital status as of 4 May 2010 could not be assigned. A subject could have died without being recorded in the Death Index if the Index had not yet recorded the date of death, which might have happened to any subjects who died after 2008. For withdrawals, consulting the Death Index allowed verification of vital status even if the subject had withdrawn from the study. In practice, all subjects who were not verified as dead as of 4 May 2010 were last observed at a time more than 5 years after surgery. This meant that no outcomes were missing due to loss to follow-up or withdrawals.

3.6.3. Descriptive statistics

For categorical variables, distributions of demographic, tumor, and treatment characteristics were determined using PROC FREQ and presented as frequencies and percentages. For continuous variables, subject characteristics were calculated using PROC TTEST and presented as means and standard deviations.

The EMT study sample represents a 10% sample of all nominally-eligible CRC cases in the catchment area during the enrollment period (190/1,899), a 19% sample of all NC-CanCORS

subjects (190/990), and a 38% sample of NC-CanCORS subjects who provided tumor tissue (190/506) (Figure 3). This raises the question of whether the EMT study sample remains representative of the underlying source population of CRC cases. The question cannot be answered definitively because doing so would require comparing the EMT subjects to all nominally-eligible CRC cases across an effectively-infinite number of personal characteristics, such as exhaustive considerations of genetics, environmental exposures, life experiences, and socioeconomic status. The difficulty in addressing this is especially great for nominally-eligible cases who did not enroll in CanCORS because detailed information on them is not available.

Since a great deal of information is available for all CanCORS subjects—not just those in the EMT sample—it is possible to compare subject characteristics for overall CanCORS subjects to the EMT sample. This allows some assessment of whether the EMT sample remains representative of all enrolled CanCORS subjects. Consequently, descriptive statistics were obtained for both the EMT sample and overall CanCORS. To formally test whether the distributions of the two samples for a given characteristic were different to a statistically-significant degree, chi-square tests were used for categorical variables and t-tests for continuous variables. P-values less than 0.05 were considered evidence that the EMT sample differed from overall CanCORS for the variable in question; p-values of 0.05 or above suggested that the two samples were reasonably similar to each other.

I also examined the distributions of the expression of each EMT marker in both normal and tumor tissue. This was done for the original continuous expression variables that were obtained directly from the data collection and therefore PROC TTEST was used to obtain means and standard deviations. For each marker expression variable, I used unpaired two-sample t-tests to compare expression in normal tissue to tumor tissue. The EMT mechanism sets up clear

expectations of how expression levels in the two tissue types should relate to each other: compared to normal, tumor tissue should have, on average, lower E-cadherin expression and higher Snail and Integrin beta-6 expression. Comparing expression in normal tissue versus tumor allowed me to assess whether these expectations were met in the observed data.

3.6.4. Identification of statistically-optimal cut points of marker expression

The data collection yielded six continuous marker expression variables: weighted average and worst score for each of Snail, E-cadherin, and Integrin beta-6. For each variable, I wished to identify the cut point distinguishing marker-positive status from marker-negative that was most strongly associated with time-to-death.

For every possible cut point along any marker expression continuum, I defined marker-positive status as expression at or above the cut point and marker-negative status as expression below the cut point. Thus, marker-positive always meant high expression and marker-negative always meant low expression. Whether marker-positive status is clinically desirable depends on the particular marker. E-cadherin-positive status would be expected to correlate with better outcomes (i.e. longer time-to-death) (2, 5, 56), whereas Snail-positive (2, 5) or Integrin beta-6-positive (23) status would be expected to correlate with worse outcomes.

To identify the statistically-optimal cut point for each continuous marker expression variable, I used a SAS macro to assess the model goodness of fit for a series of bivariate associations between dichotomous marker expression status and time-to-death. The macro iteratively dichotomized marker expression at every possible cut point in the observed tumor tissue data, with each cut point corresponding to a different subject's expression value. Each dichotomization of marker expression status was fit as the only independent variable in a Cox model with time-to-death as the outcome, producing a Bayesian Information Criterion (BIC)

model fit statistic. The expression value with the lowest BIC statistic was considered the statistically-optimal cut point for that marker expression variable.

This approach is an alternative to receiver operating characteristic curves (ROC). Both methods can be used to select a cut point along a continuum of marker expression values based on a criterion that relates marker expression to patient outcomes. The methods differ in two important respects.

First, the form of subject outcomes is different. ROC curves use a binary outcome of whether a subject died (yes/no) while our approach uses continuous time-to-death. It matters whether a patient died 5 months or 50 months after surgery. ROC curves do not account for such distinctions whereas our approach does.

Second, the criterion used to identify a cut point differs between the methods. In ROC curves, the cut point selected is typically the one corresponding to the most upper-left-hand point on a plot of sensitivity versus $(1 - \text{specificity})$ (i.e. true positive rate versus false positive rate). In our approach, the statistically-optimal cut point is the one yielding the best model fit in a bivariate proportional hazards model of dichotomous marker expression and continuous time-to-death.

This difference between cut point selection criteria implies a difference in interpretation between the two methods. Being based on measures of sensitivity and specificity, the cut point selected by an ROC curve is usually interpreted as the one that should be implemented clinically. When a cut point is selected in the upper-left-hand corner of the ROC curve, the presumption is that false positives and false negatives have clinical consequences of roughly equal importance, which is not always true. In contrast, the optimal cut point in our approach is a statistical measure of the largest difference in the observed data between hazard functions for marker-

positive and marker-negative subjects. Thus, the direct application of our method is to determine whether an association exists between marker expression and time-to-death. The statistically-optimal cut point might or might not be judged to be best for clinical use, but that determination requires consideration of additional information besides model fit alone.

For macro SAS code and further details about our approach, see Appendix B.

3.6.5. Bivariate time-to-event analysis

The optimization macro was used to identify the statistically-optimal cut point in the observed data for each of the six continuous marker expression variables. On the average intensity scale of 0-3 for E-cadherin, we found that the statistically-optimal cut point was about 0.52 for weighted averages and 0.42 for worst cores. On a percent positive cells scale for Integrin beta-6, the optimal cut point was about 9.4% for weighted averages and 7.7% for worst cores. On a percent positive nuclei scale for Snail using the first nuclear scoring algorithm, the optimal cut point was about 25.2% for weighted averages and 63.6% for worst cores. This information was used to create dichotomous expression variables defined by these cut points. I performed bivariate analyses of the relationship between dichotomous marker expression and subject survival by generating marker expression-stratified Kaplan-Meier survival curves.

The Kaplan-Meier or product limit estimator of survival is given by

$$S(t) = \text{Product}_{t_i < t} [1 - (d_i / n_i)]$$

where $S(t)$ is the probability of a member of a given population having a lifetime exceeding time t , n_i is the number of subjects at risk for the event of interest (here: death) just prior to time t_i , and d_i is the number of deaths at time t_i (57). I used PROC LIFETEST to generate Kaplan-Meier survival curves among subjects in our dataset stratified by dichotomous marker expression status. This involved the use of each subject's recorded time from surgery to death or last

observation, as well as information as to whether the subject's observation time was censored at 5 years after surgery. Differences in survival between subjects in different strata of marker expression status were tested for statistical significance using the logrank test, using a significance threshold of 0.05.

Stratified survival curves were generated by stratifying on the following sets of variables:

- a) E-cadherin weighted average
- b) E-cadherin worst core
- c) Integrin beta-6 weighted average
- d) Integrin beta-6 worst core
- e) Snail weighted average
- f) Snail worst core
- g) Jointly by E-cadherin weighted average and Integrin beta-6 weighted average
- h) Jointly by E-cadherin worst core and Integrin beta-6 worst core
- i) Jointly by E-cadherin weighted average and Snail weighted average
- j) Jointly by E-cadherin worst core and Snail worst core
- k) Jointly by Integrin beta-6 weighted average and Snail weighted average
- l) Jointly by Integrin beta-6 worst core and Snail worst core

In addition to Kaplan-Meier estimation, I also performed bivariate time-to-event analysis using Cox proportional hazards modeling. The survivor and hazard functions are equivalent and are related by the formula

$$h(t) = -d/dt [\log S(t)]$$

where $h(t)$ is the hazard function and $S(t)$ is the survivor function (57). Thus, one would expect bivariate Cox models to produce the same results, in different form, as Kaplan-Meier estimates.

The bivariate Cox models were run using PROC PHREG and included dichotomous marker expression as the only independent variable in a model with outcome of time-to-death. Subjects were administratively censored at 5 years after surgery in Cox models as they were in Kaplan-Meier estimates.

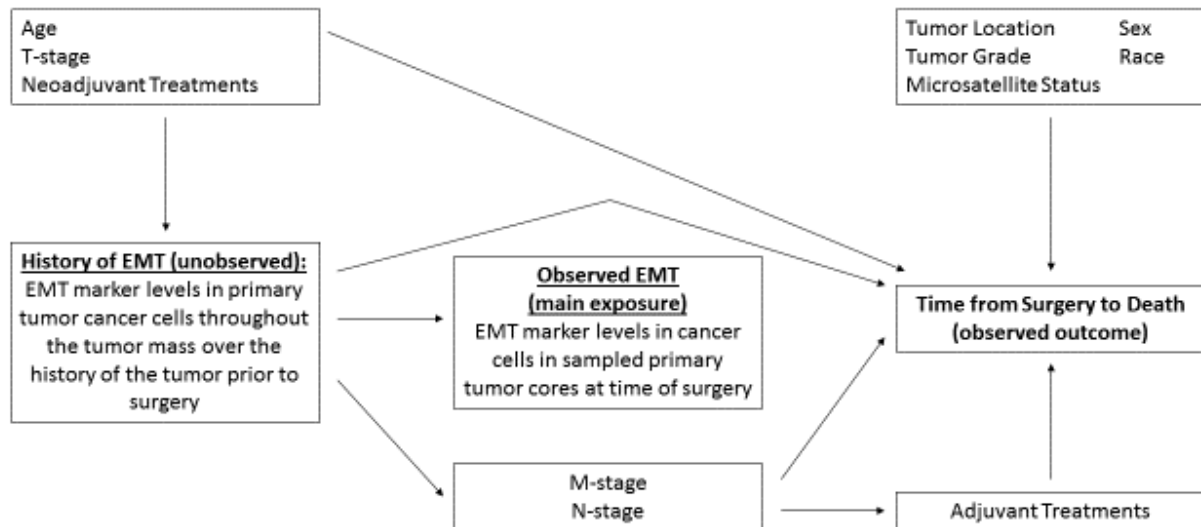
3.6.6. Covariate selection

The ultimate goal of our time-to-event modeling was to produce valid estimates of the association between each dichotomous marker expression variable and time-to-death. Achieving that validity requires, among other things, that confounding of the association of interest by other variables be controlled so as not to bias the association of interest (58). Thus, our most valid estimates of the association between dichotomous marker expression and death are those adjusted for sources of confounding, insofar as those sources can be discerned.

Past studies of associations between EMT marker expression in primary tumors and patient outcomes were not consistent in terms of what covariates they adjusted for in multivariate models (5). None of them provided a justification for their decisions. To identify the most valid adjustment set possible supported by a reasonable rationale, we selected covariates for our multivariate Cox models based on the results of past studies, considerations of biological plausibility, and directed acyclic graph (DAG) theory (59).

Figure 8 (below) presents our DAG showing postulated causal relationships between the exposure (observed primary tumor cancer cell EMT marker expression), outcome (time from surgery to death), and other relevant variables.

Figure 8. Directed acyclic graph (causal diagram) depicting postulated relationships between primary tumor cancer cell EMT marker expression, time-to-death, and other variables



Before describing the consequences of the postulated relationships among the variables, the presentation of the DAG should be clarified. Several nodes contain multiple variables. This is done solely to streamline the presentation. For each node with multiple variables—say, the one containing both M-stage and N-stage—putting the variables in the same node merely indicates that each of these variables is thought to have the same relationships with ancestor and descendant nodes as every other variable in the node. Variables “sharing” a node should be considered independent of each other apart from sharing the same ancestors and descendants.

The measured association of interest is that between Observed EMT and Time from Surgery to Death. Elimination of confounding requires adjusting for a set of covariates that will “block” any open paths of association between the main exposure and outcome other than the path leading directly from the main exposure to the outcome. These alternate paths of association that must be blocked by adjustment are known as “backdoor paths,” which are open paths connecting the main exposure to the outcome while including an arrow head pointing into

the main exposure (59). In this case, there are several backdoor paths, all of them leading into the main exposure strictly through the unobserved variable History of EMT, which is the expression of EMT markers in cancer cells throughout the tumor mass over the entire history of the tumor prior to surgery. All backdoor paths could in theory be blocked by adjusting for History of EMT. Since the variable is unobserved, each backdoor path must be blocked by adjusting for other variables along the path.

According to the postulated relationships among variables in Figure 8, the minimal set of covariates that would need to be adjusted to block all backdoor paths would be age, T-stage (tumor size), neoadjuvant treatments (chemotherapy and radiation), N-stage (lymph node metastasis), and M-stage (distant metastasis). Since overall TNM tumor stage is a composite of the component T-, N-, and M-stages, tumor stage can be adjusted for as a single variable for overall stage rather than as three separate variables for the components. Confounding due to adjuvant treatments is addressed by adjusting for stage. The variables in the node including “Tumor Location,” while associated with patient outcomes, do not appear to influence EMT marker expression in cancer cells and therefore need not be included in models.

3.6.7. Multivariate time-to-event analysis

Adjusted Cox proportional hazards models to control for identified sources of confounding included the following independent variables: dichotomous marker expression status (marker-positive/marker-negative), age (continuous), overall tumor stage (local/regional/distant), neoadjuvant chemotherapy (yes/no), and neoadjuvant radiation therapy (yes/no). Each model contained only one marker expression variable. I did not include more than one marker expression variable simultaneously because markers were not considered

mutual sources of confounding for each other based on our DAG. As with bivariate models, the outcome was time from surgery to death, with administrative censoring at 5 years after surgery.

3.6.8. Missing data and multiple imputation

Several variables for multivariate modeling had large proportions of missing and/or non-informative responses. “Non-informative responses” included responses such as “Don’t Know,” “Unknown,” and “No Answer.” Such non-informative responses were effectively missing data. Only subjects with informative responses or values for all model variables could be included in models. Since our sample size of 190 subjects was small, adequate precision was an important concern. Retaining all subjects in multivariate models was essential to preserve the maximum possible precision.

In the presence of missing data, retaining all subjects required imputation of reasonable “guesses” for missing values on variables to be included in multivariate Cox models (60). Many imputation methods based on conventional analytic approaches exist—marginal mean imputation, multiple regression, and weighted least squares, among others—but these methods tend to underestimate standard errors and overestimate test statistics (60). I used multiple imputation (MI), a method that provides consistent, asymptotically efficient, and asymptotically normal estimates, assuming the data are missing at random (60).

Briefly, MI involves assigning an imputed value for each missing value for a given variable conditional on relevant covariates (the imputation model) and by introducing a random component into the imputation process. The imputation model should be at least as rich as the analysis model used subsequently. Using this approach, a statistical package performing MI will assign imputed values for all missing values for every variable being imputed, creating a complete dataset. This was done 100 times to generate 100 complete datasets. Each of these

datasets was then analyzed separately using the desired analysis model—in our case, Cox proportional hazards models of time-to-death regressed on dichotomous marker expression status, age, stage, and neoadjuvant treatments—to produce 100 point estimates and standard errors. The 100 analysis results were finally combined into a single, stabilized estimate and confidence interval. To perform MI, I used the following procedure for every multivariate Cox model that I ran after recoding non-informative responses as missing values:

First, PROC MI was used to implement Markov Chain Monte Carlo imputation for 100 imputed datasets. The imputation model included the following variables: dichotomous marker expression status, age, sex, race, tumor stage, tumor location, tumor grade, whether received neoadjuvant chemotherapy, whether received neoadjuvant radiation therapy, time from diagnosis to death, time from diagnosis to surgery, and time from surgery to death.

Therefore, any missing values for all independent variables and *also* the dependent variable of time from surgery to death were substituted for an imputed value. Multiple time-to-event variables were included in the imputation model because, while time from surgery to death was the outcome of interest, it was missing for some subjects in the final EMT sample. In fact, time from surgery to death was not a variable collected by CanCORS. Instead, I derived it from two variables that were collected: time from diagnosis to death (or last observation), and time from diagnosis to surgery. The formula for time from surgery to death was (time from diagnosis to death or last observation) – (time from diagnosis to surgery), keeping in mind that surgery virtually always comes after diagnosis.

The dataset had no missing data for time from diagnosis to death (or last observation), but had about 14% missing data for time from diagnosis to surgery. Applying the formula above produced about 14% missing data for time from surgery to death. Therefore, to retain all

subjects, imputation had to be performed for the dependent variable as well as independent variables. Of note, there was no association between having a missing value for time from surgery to death and whether a subject died during the observation period (chi square p-value=0.3).

Second, after creating 100 imputed datasets, I used PROC PHREG to carry out the Cox analysis model on each imputed dataset, producing 100 point estimates and standard errors. The analysis model used time from surgery to death as the dependent variable and, for the independent variables, included dichotomous marker expression status, age, tumor stage, receipt of neoadjuvant chemotherapy, and receipt of neoadjuvant radiation therapy.

Third and finally, PROC MIANALYZE was used to combine the 100 point estimates and standard errors into a single, stabilized point estimate and 95% confidence interval that retained all 190 subjects.

3.7. Statistical analysis for latent class analysis

3.7.1. Overview

The time-to-event analysis addressed the basic question of whether expression levels of a given marker were associated with patient outcomes, especially independent of tumor stage. The answer to that question does not, however, provide all of the information we would want to know about the marker in terms of its diagnostic usefulness. We would also want a measurement of the diagnostic validity or accuracy of the marker (58). The basic measures of such accuracy are sensitivity and specificity, and estimating these quantities in an appropriate fashion for EMT markers was the aim of my latent class analysis.

Sensitivity and specificity are defined in terms of both a dichotomous diagnostic test and a dichotomous outcome or health state (58). The clinical use of EMT markers fits comfortably

within this framework. Binary EMT marker expression status—marker-positive versus marker-negative—is easily defined, even if EMT marker expression measurements are initially continuous. The decision that an oncologist faces and that an EMT marker would inform is inherently binary: at the time of diagnosis, and given that a primary tumor has been found, does the patient likely have or not have metastatic disease?

Traditional, straightforward calculations of sensitivity and specificity are based on a cross-tabulation of patient classification according to the new or “test” diagnostic and patient classification according to a gold standard measure of the health state of interest (58). A gold standard measure has perfect (or nearly perfect) accuracy in correctly classifying *both* those with and without the health state of interest, which in this case is presence or absence of metastatic disease. The standard measures of presence or absence of metastatic disease are lymph node evaluation (LN) and radiologic imaging (RI).

Empirically, it is clear that these measures, either individually or taken together, are not sufficiently accurate to be considered a gold standard. Specifically, about 25% of CRC patients found to have local disease according to these diagnostics ultimately experience disease recurrence (4). Such patients likely had metastatic disease at the time of diagnosis that was not successfully detected by LN or RI. Since about 40% of CRC patients are diagnosed as having local disease, this suggests that approximately 10% of all CRC patients—about 14,000 people nationwide each year—have metastatic disease at the time of diagnosis that the two conventional diagnostics jointly fail to detect (8, 10).

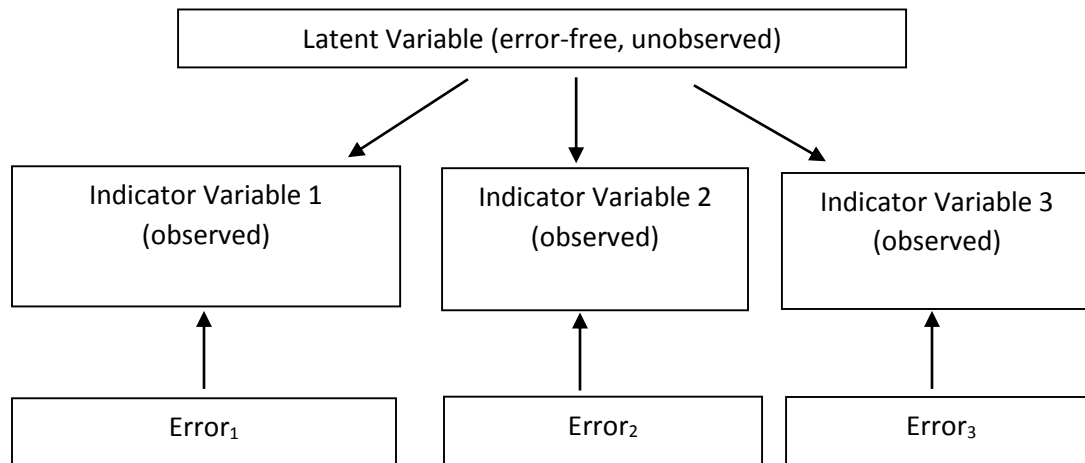
The failure of LN and RI to constitute a gold standard measure implies that EMT marker sensitivity and specificity cannot be estimated by simple calculations based on a cross-tabulation. However, it is possible to estimate the sensitivity and specificity from statistical models that do

not require the assumption that any member of a panel of diagnostic tests counts as a gold standard. Latent class models are precisely suited for this purpose (61). We used latent class analysis (LCA) to estimate the sensitivity and specificity of EMT markers to assess cancer cell detachment from primary tumors while accounting for the sensitivity and specificity of LN and RI to do the same, but without assuming that any of the diagnostics constituted a gold standard.

3.7.2. Latent class framework

Latent class models are a subset of latent variable models, which attempt to identify subgroups within a population by postulating that the subgroups represent different levels of an unobserved, error-free latent variable (62). The latent variable is measured indirectly by multiple observed indicator variables (also called manifest variables), each of which is conceived as being determined by its own error term and the latent variable. Individuals' response patterns to the indicators can be used to estimate two kinds of parameters: the prevalence of each latent subgroup (gamma parameters or class-membership probabilities) and, within each latent subgroup, the probability of a given response to each indicator (rho parameters or item-response probabilities). Latent class models are latent variable models that use categorical indicators and postulate a categorical latent variable. Figure 9 depicts a generic latent class model with three manifest variables.

Figure 9. Generic latent class model with three indicator variables (based on Collins and Lanza, 2010) (62)

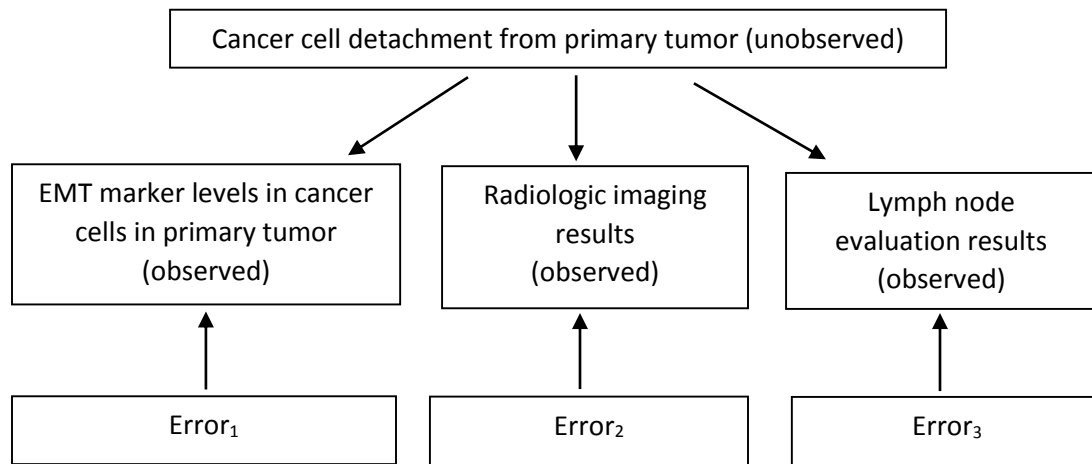


In a valid latent class model, the indicators are related to each other through the latent variable but are otherwise independent of one another (62). This is known as the local independence assumption. It implies that the error terms of the indicators are not associated or correlated with each other. A further implication is that, within any single (unobserved) latent class, the indicator variables are independent of each other. In fact, this conditioning on latent class is what the “local” in “local independence assumption” refers to.

3.7.3. Latent class model for diagnostics of cancer cell detachment from primary tumors

Figure 10 depicts my conceptual model of the relationship between cancer cell detachment from the primary tumor and the diagnostic tests to assess it. Detachment is the latent variable, that is, the phenomenon of interest that is not observed directly. EMT markers, LN, and RI are the indicators.

Figure 10. Latent class model of diagnostic tests to evaluate cancer cell detachment from primary tumors



3.7.4. Data used for latent class analysis

Based on the results of the time-to-event analysis, only E-cadherin expression as a weighted average of tumor cores was associated with time-to-death. Therefore, EMT status in the LCA was defined solely in terms of E-cadherin weighted average expression. The LCA was limited to the 188 subjects in the EMT study sample with E-cadherin weighted average measurements, as 2 of the 190 subjects did not have E-cadherin measurements.

The only data input into LCA models were the three diagnostic tests of cancer cell detachment. The sensitivity and specificity of each diagnostic could be estimated freely or set to fixed values. Free estimation requires observed data for that particular diagnostic, while a fixed value does not. All diagnostics were treated as binary test-positive versus test-negative. Test-positive meant evidence supporting cancer cell detachment from the primary tumor while test-negative meant evidence supporting lack of cancer cell detachment.

EMT

In the time-to-event analysis, we identified three different cut points along the continuum of observed E-cadherin weighted average expression that provided informative relationships between dichotomous marker expression and patient outcomes. On the average intensity scale of 0-3, these were 0.52, 0.60, and 0.85. In the LCA we examined how sensitivity and specificity varied with cut point. For any given cut point, EMT-positive status meant low E-cadherin expression (below the cut point) and EMT-negative status meant high E-cadherin expression (at or above the cut point). EMT sensitivity and specificity were both freely estimated in all latent class models.

Lymph Node Evaluation and Radiologic Imaging

Freely estimating all possible sensitivity and specificity parameters in a setting with only three binary diagnostics leads to poor model specification (see Analysis section below). To cope with this, in each model I fixed (i.e. set or restricted) some of the values of LN and RI sensitivity and specificity while freely estimating the others as well as EMT sensitivity and specificity. This approach required both observed LN and RI test results for freely estimated parameters and also determination of fixed values for restricted parameters.

Observed LN and RI results were not available in CanCORS. As an approximation, I inferred them from subject tumor stage using the rules presented in Table 2. These inferred test results were used for freely estimated LN and RI sensitivity and specificity parameters.

For LN and RI parameters assigned fixed values in a given model, a value of 60% or 80% was assigned for each restricted sensitivity parameter and a value of 90% or 100% was assigned for each restricted specificity parameter. These values were chosen to represent bounds to assess the influence on model results of different combinations of assumptions about high or low

diagnostic accuracy for RI and LN. We varied the combinations of fixed values and of which LN and RI parameters were restricted across models to observe how doing so impacted the EMT estimates.

The choice of 60% to represent poor sensitivity and 80% to represent good sensitivity was based in part on the fact that about 25% of colorectal cancer patients diagnosed with local disease experience recurrence after surgery (4). In addition, roughly 40% of colorectal cancer patients are currently diagnosed with local disease (63). Together, these facts suggest that the “double false negative” proportion for LN and RI is about 10% of all colorectal cancer patients.

Table 2. Rules for inferring lymph node and radiologic imaging test results from tumor stage

Tumor Stage Diagnosis	Lymph Node Assignment	Imaging Assignment
Local	Negative	Negative
Regional	Positive	Negative
Distant	Based on diagnosed N-stage ^a	Positive

^aFor 23 subjects diagnosed with distant disease, lymph node status was assigned as lymph node-negative if N0 (n=3) and as lymph node-positive if N1 or N2 (n=11). Nine subjects with unknown N-stage were assigned as lymph node-positive.

For metastatic disease, false negative results are more clinically serious than false positive results. However, the problem is not testing falsely negative for either LN or RI, but testing falsely negative for both. The false negative proportion for a single diagnostic is equal to $(100 - \text{sensitivity})\%$ (58). Thus, a test with sensitivity of 60% or 80% will have a false negative proportion of 40% or 20%, respectively.

Assuming that LN and RI test results are conditionally-independent events, the double false negative proportion is the product of the false negative proportions for each of the two tests. Both tests having sensitivity of 80% yields a double false negative proportion of 4%. If both tests have sensitivity of 60%, then the double false negative proportion is 16%. If one test has

sensitivity of 60% and the other has sensitivity of 80%, then the double false negative proportion is 8%.

I chose 60% and 80% as values representing low and high test sensitivity, respectively, because when either LN or RI has low sensitivity and the other has high sensitivity, the assumed double false negative proportion of 8% is close to the observed estimate of 10%. In addition, when both tests are set to high sensitivity or both set to low sensitivity, the resulting double false negative proportions of 4% and 16% set up reasonable extreme bounds around the observed estimate.

Specificity bounds of 90% and 100% were chosen because I assume that the specificities of both LN and RI are close to perfect. To get a false positive result for either of these tests, the radiologist or pathologist would have to believe that he is looking at cancer cells in a scan or under the microscope when in fact he is not. This might occur occasionally but I assume it is relatively rare.

Whether freely estimated or restricted, every LN and RI sensitivity and specificity parameter represented a binary diagnostic test. LN-positive would mean that cancer cells were found in lymph nodes near the primary tumor and LN-negative that no cancer cells were found there. RI-positive would indicate detection of a distant metastasis via imaging and RI-negative would indicate no such detection of distant metastasis.

3.7.5. Analysis

All models included all three diagnostic tests. In every case, I sought a 2-class solution distinguishing high risk from low risk for metastatic disease.

A properly specified latent class model has a positive number of degrees of freedom (df), given by $df = W - P - 1$, where W is the number of possible response patterns and P is the

number of parameters estimated (62). Three binary indicators yield $2^3 = 8$ possible response patterns. When all class-membership and item-response parameters are freely estimated, a 2-class model with 3 indicators estimates 7 parameters (1 class-membership and 6 item-response), leaving $8 - 7 - 1 = 0$ df. Using fixed sensitivity or specificity values for LN and RI reduced the number of parameters estimated in a given model, thereby giving positive degrees of freedom.

I ran separate models using each of the different E-cadherin cut point values mentioned earlier—about 0.52, 0.60, and 0.85—to create the EMT indicator to assess the impact of E-cadherin cut point on the sensitivity and specificity of the resulting indicator. The fixed values assigned for LN and RI sensitivity or specificity were also varied across models. I used PROC LCA in SAS for the latent class analyses (64). The SAS procedure used an iterative expectation-maximization algorithm to obtain maximum likelihood estimates for freely-estimated parameters.

3.8. Ethical considerations

3.8.1. IRB approval

The dissertation proposal was defended on 16 April 2013. On 23 April 2013, an application for approval of the project was submitted to the UNC Institutional Review Board (IRB). The project was assigned IRB Study #13-1897. On 25 April 2013, the IRB notified us that the project was approved. No implementation activities were undertaken until IRB approval was received. The IRB granted one-year extensions on 24 February 2014 and 14 January 2015, covering the entire period from just after the proposal through graduation from the PhD program.

3.8.2. Informed consent

Subjects consented to participate in the study and provide tumor specimens at the time of enrollment in 2003-06. While they did not specifically consent to have their tumor specimens

used to study EMT markers, they did consent to the tissue samples being used for biomarker research generally. Thus, the original consent covered all dissertation project activities, and therefore subjects did not need to be sought out and re-consented to carry out the project.

3.8.3. Potential risks to subjects

The only risk this research posed to subjects was possible breach of confidentiality in the event that data security was compromised. Subject data used in the project were deidentified and never linked to the CanCORS key of subject identifiers. Further, all analyses were carried out on a password-protected computer used only by the candidate. To the best of our knowledge, data security was never breached. Even if it was breached without our knowledge, it appeared unlikely that subject confidentiality could be compromised even then since the data were deidentified.

3.8.4. Potential benefits to subjects

There were no direct benefits of this research to participating subjects. While the knowledge gained could benefit future CRC patients, the work was performed long after the knowledge could have been beneficial to those enrolled.

CHAPTER 4. MARKERS OF EPITHELIAL-MESENCHYMAL TRANSITION AND MORTALITY IN A POPULATION-BASED PROSPECTIVE COHORT OF COLORECTAL CANCER PATIENTS

4.1. Introduction

Epithelial-mesenchymal transition (EMT) is widely considered an important mechanism of cancer cell metastasis (2, 65, 66). It connects epithelial cells to metastasis, which account for roughly 80% of cancer and 90% of cancer deaths, respectively (1). EMT markers measured in primary-tumor cancer cells are potentially useful diagnostic tools to assess patient risk of metastatic disease and guide treatment decisions, even when metastases are not detected by lymph-node evaluation or radiologic imaging.

The EMT mechanism involves epithelial cells temporarily becoming mesenchymal cells (2). This occurs when cellular expression levels of EMT inducers increase, leading to decreased expression of epithelial markers and increased expression of mesenchymal markers. These changes are manifest as loss of adhesion to adjacent cells that enables detachment of the transitioning cell from the primary tumor, as well as cytoskeletal and other modifications that enhance motility and invasiveness.

Past studies of associations between EMT markers in primary tumors and patient outcomes have measured marker expression in different ways and often performed statistical analyses that were not as informative as possible (5). For example, previous studies have used single-hospital samples that typically are not representative of a well-defined population. Different studies of the same marker measured using the same laboratory technique often used different scoring scales and defined marker-positive versus marker-negative status in different

ways. The lack of uniform methods across studies could contribute to inconsistent findings and hamper translation of EMT markers to clinical use.

Selecting markers based on our previous literature review (5), we measured the EMT inducer Snail, epithelial marker E-cadherin, and mesenchymal marker Integrin beta-6 in primary tumors from a population-based prospective cohort study of colorectal cancer (CRC) mortality and estimated their associations with time-to-death. We hypothesized that low expression of E-cadherin, and high expression of Snail and Integrin beta-6, would be associated with shorter times from surgery to death compared to opposite expression levels. We also introduce methods that might help to improve standardization of measurements and analyses across studies.

4.2. Methods

4.2.1. Study population

Subjects were enrolled in the Cancer Care Outcomes Research and Surveillance Consortium (CanCORS), a population-based, prospective, case-only, multi-site observational study of colorectal and lung cancer patients (48). Briefly, the study assessed the impact of health-system, provider, and patient factors on cancer outcomes. Patients were at least 21 years of age at diagnosis and were enrolled within 3 months of diagnosis during 2003-06. The study collected patient surveys, surrogate surveys for patients who were deceased or too ill to participate, and medical records data. Vital status for all subjects was verified using the Social Security Death Index on 4 May 2010, providing at least 42 months of follow-up observation per individual.

Medical records abstractors at each site collected information on tumor characteristics and cancer treatments. Patient and, when necessary, surrogate surveys were completed using computer-assisted telephone interviews that queried demographic and socioeconomic factors

(age, insurance coverage, income), as well as treatment preferences and interactions with providers (67).

The North Carolina CanCORS site was the only one to collect tumor specimens. It enrolled 990 CRC patients but no lung cancer patients. Subjects in the present biomarker study came from a catchment area of 33 counties in eastern and central North Carolina at the time of diagnosis. Investigators obtained primary tumor and normal adjacent colorectal tissue samples from 506 subjects.

Formalin-fixed, paraffin-embedded tissue specimens were sent from hospitals across the catchment area to the University of North Carolina at Chapel Hill (UNC), where they were used to construct tissue microarrays (TMAs) as described previously (50). Most subjects had multiple cores from both primary tumor and normal margin. For this analysis, we measured EMT markers in 12 representative TMAs that included specimens from 219 subjects. To be included in the study sample, subjects had to have at least one core of tumor tissue successfully stained for one marker, with that core having at least 50 epithelial cells and unambiguous histology. From the 12 TMAs, we excluded 26 subjects lacking adequate tumor tissue and an additional 3 subjects who could not be linked to medical records data, yielding a final study sample of 190 subjects.

4.2.2. Immunohistochemistry

We selected EMT markers for evaluation based on the results of previous studies and criteria discussed in our prior literature review (5). Marker protein expression was measured using the following antibodies: E-cadherin (mouse monoclonal ready to use [RTU], clone 36B5 [cat #PA0387] from Leica Microsystems Inc. [Norwell, MA]), Integrin beta-6 (goat polyclonal

[sc-6632] from Santa Cruz Biotechnology [Dallas, Texas]), and Snail (goat polyclonal [ab53519] from Abcam [Cambridge, MA]).

Immunohistochemistry (IHC) was performed at the UNC Translational Pathology Laboratory (TPL) using the Bond fully-automated slide staining system (Leica Microsystems Inc., Norwell, MA). Slides were deparaffinized in Bond Dewax solution (AR9222) and hydrated in Bond Wash solution (AR9590). Antigen retrieval was performed at 100°C for Integrin beta-6 and Snail (for 20 minutes or 30 minutes, respectively) in Bond-epitope retrieval solution 1 at pH 6.0 (AR9961) and for E-cadherin for 20 minutes at 100°C in solution 2 at pH 9.0 (AR9640). After pretreatment, anti-E-cadherin was applied for 15 minutes, anti-Snail (1:200) for 30 minutes and anti-Integrin beta-6 (1:100) was applied for 1 hour.

Detection of Snail and Integrin beta-6 was performed using the Bond Intense R Detection System (DS9263) supplemented with the LSAB+ kit (DAKO, Carpinteria, CA). E-cadherin detection used the Bond Polymer Refine Detection System (DS9800). Stained slides were dehydrated and cover-slipped. Positive and negative controls (no primary antibody) were included for each antibody. All assays were single-marker (i.e. no multiplex assays).

Stained slides were digitally imaged at 20× magnification using the Aperio ScanScope XT (Aperio Technologies, Vista, CA). Digital images were stored in the Aperio Spectrum Database. Example images of staining and annotations are provided in Figures 4-6.

4.2.3. Automated analysis of digital IHC images

Computer algorithms annotated and scored every eligible tissue core to obtain continuous marker expression data. We used approximately 65 cores originating from two TMAs for algorithm training and automated-analysis validation.

Definiens Composer Technology (Tissue Studio version 2.1.1 with Tissue Studio Library version 3.6.1; Definiens Inc., Carlsbad CA) was used to identify regions enriched in epithelial cells in IHC-stained TMA cores. To detect differences in cell shape and tissue structure, we developed two Composer algorithms per marker—one for cores containing normal adjacent tissue and a second for tumor cores—as both types of tissue were present on each TMA.

After Composer training, we developed two Tissue Studio scoring algorithms (“solutions”) per marker. Different Composers were used to identify epithelial cell regions but identical settings were used to determine relative DAB staining intensity. The Composer_MarkerArea Solution was used to detect E-cadherin staining. This algorithm gave average intensity readings for each core on a continuous scale of 0-3 and included both membrane and cytoplasmic E-cadherin staining. Integrin beta-6 membrane staining was detected using the Composer_Nuclei_Membranes&Cells Solution. We used the Composer_Nuclei(Positive_vs_Negative) Solution to detect nuclear Snail expression. Integrin beta-6 and Snail were measured as core percent positive cells (Integrin) or nuclei (Snail) on a continuous scale of 0-100.

To evaluate the reliability of computer annotations, one of us (ELB) used Aperio ImageScope (version 11.2; Leica Biosystem, Buffalo Grove, IL) to manually annotate the same 65 cores per marker that were used to optimize Tissue Studio solutions. He remained blind to patient and tumor characteristics while annotating. The manually-annotated cores—considered the gold standard for digital separation of tissue types—were then analyzed using appropriate Aperio scoring algorithms (Membrane v9 algorithm for E-cadherin and Integrin beta-6, Nuclear v9 algorithm for Snail). Automated scores obtained via manual and automated annotation produced Pearson correlations of 0.91 for E-cadherin, 0.88 for Integrin beta-6, and 0.94 for Snail.

Having verified the accuracy of the annotation algorithms for the three markers, all 12 TMAs stained for E-cadherin, Integrin beta-6 and Snail were analyzed (36 slides in total).

Subjects typically had multiple cores available of a given tissue type (tumor or normal). To assign an expression value for each subject by marker and tissue type, we handled replicate cores in two ways: first, as a weighted average of cores, and second, by assigning the expression value of the subject's "worst" core as the marker expression value. For weighted averages, the weights were area analyzed for E-cadherin, number of cells for Integrin beta-6, and number of nuclei for Snail. The worst score by tissue type was assigned as the lowest core average intensity for E-cadherin, as the highest core percent positive cells for Integrin beta-6, and as the highest core percent positive nuclei for Snail.

After observing unexpected results when comparing mean Snail expression in tumor tissue to normal tissue, we trained two additional Snail scoring algorithms on a different TMA than the one used to develop the original scoring algorithms: a second nuclear algorithm and a whole-cell algorithm that scored Snail expression in any part of a cell. These additional Snail scoring algorithms measured expression on a continuous percent positive scale and were applied to all 12 TMAs. The results of the two additional Snail algorithms were qualitatively similar to those from the original nuclear algorithm. Thus, for analysis we used only Snail expression values based on the original algorithm.

4.2.4. Outcome

In statistical models, the dependent variable was length of time in days from primary tumor surgery until all-cause mortality, with administrative censoring at 5 years after surgery.

4.2.5. Covariates

Covariates for multivariate statistical models were selected based on prior studies (5), considerations of biological plausibility, and directed acyclic graph theory (59).

We adjusted for age, neoadjuvant chemotherapy, neoadjuvant radiation therapy, tumor size (T-stage), lymph-node metastasis diagnosis (N-stage), and distant metastasis diagnosis (M-stage). We adjusted for overall TNM stage as a single variable instead of adjusting for the component stages as three separate variables since including both overall stage and any of the component stages would constitute inappropriate overadjustment. Because we adjusted for stage, cancer treatments prior to surgery were included as covariates but we excluded cancer treatments occurring after surgery.

Age was modeled as a continuous variable. We categorized stage (local/regional/distant), neoadjuvant chemotherapy (yes/no), and neoadjuvant radiation (yes/no).

4.2.6. Marker expression cut point optimization

Our data collection yielded six continuous marker expression variables: weighted average and worst score for each of Snail, E-cadherin, and Integrin beta-6. For each variable, we wished to identify the cut point distinguishing marker-positive status from marker-negative that was most strongly associated with time-to-death.

For every possible cut point along any marker expression continuum, we defined marker-positive status as expression at or above the cut point and marker-negative status as expression below the cut point. Thus, marker-positive always meant high expression and marker-negative always meant low expression. Whether marker-positive status is clinically desirable depends on the particular marker. E-cadherin-positive status would be expected to correlate with better

outcomes (i.e. longer time-to-death) (2, 5, 56), whereas Snail-positive (2, 5) or Integrin beta-6-positive (23) status would be expected to correlate with worse outcomes.

To identify the statistically-optimal cut point for each continuous marker expression variable, we used a SAS macro to assess the model goodness of fit for a series of bivariate associations between dichotomous marker expression status and time-to-death. The macro iteratively dichotomized marker expression at every possible cut point in the observed tumor tissue data, with each cut point corresponding to a different subject's expression value. Each dichotomization of marker expression status was fit as the only independent variable in a Cox model with time-to-death as the outcome, producing a model fit statistic. The expression value with the lowest model fit statistic was considered the statistically-optimal cut point.

For macro SAS code and further details, including comparison of this approach to receiver operating characteristic curves, see Appendix B.

4.2.7. Statistical analysis

We first used unpaired two-sample t-tests to assess whether average continuous marker expression differed between tumor and normal tissue. All subsequent analyses used tumor tissue only. We applied the macro to the tumor tissue data for the six marker expression variables to identify the statistically-optimal cut point for each. Every optimal cut point was used to create a dichotomous marker expression variable (positive/negative).

We generated Kaplan-Meier survival curves stratified by dichotomous marker expression status for one marker or two markers jointly, assessing differences between strata using the logrank test. Next, for each optimally-dichotomous marker expression variable, we fit unadjusted and adjusted Cox proportional hazards models of time-to-death. Prior to modeling, non-informative observations (e.g. "No Answer," "Don't Know," "Unknown") were recoded as

missing. Missing data for all model variables were evaluated using multiple imputation. P-values of 0.05 or below were considered statistically significant. All analyses were performed using SAS 9.3 (SAS Institute, Cary, NC).

The Institutional Review Board at UNC approved the protocol. All subjects provided informed consent.

4.3. Results

Comparison of subject characteristics for overall North Carolina CanCORS and the subset for whom EMT markers were measured in primary tumors suggests that the EMT study sample remains representative of the source population (Table 3).

On average, tumor tissue had lower E-cadherin expression and greater Integrin beta-6 expression than normal adjacent tissue regardless of whether expression values were assigned as a weighted average of cores or as the worst core (Table 4). However, average Snail expression was higher in normal tissue than in tumor tissue. While the difference was not large, this relationship was consistent across both ways of assigning expression values and all three Snail scoring algorithms.

On the average intensity scale of 0-3 for E-cadherin, we found that the statistically-optimal cut point was about 0.52 for weighted averages and 0.42 for worst cores. On a percent positive cells scale for Integrin beta-6, the optimal cut point was about 9.4% for weighted averages and 7.7% for worst cores. On a percent positive nuclei scale for Snail using the first nuclear scoring algorithm, the optimal cut point was about 25.2% for weighted averages and 63.6% for worst cores. For weighted average and worst core expression for each marker, Table 5 presents the cross-tabulation of dichotomous marker expression status defined by the

statistically-optimal cut point, first, with tumor stage, and second, with the risk of dying within 5 years of surgery.

For E-cadherin weighted averages, subjects with low tumor expression had worse survival than those with high tumor expression (Figure 11). None of the Kaplan-Meier curves stratified by the other five optimally-dichotomous marker expression variables revealed a statistically-significant difference in survival (Figures 12-16). We also generated survival curves jointly stratified by two dichotomous marker expression variables (Figures 17-22). While several of these produced statistically-significant logrank test results, graphical examination did not suggest any clear patterns based on biological expectations or clinical usefulness.

Bivariate proportional hazards model results paralleled the single-variable stratified survival curves: low E-cadherin weighted average expression was associated with greater hazards of dying than high expression (Hazard Ratio [HR] =2.84, 95% Confidence Interval [CI] 1.29, 6.28), and no associations were found for any of the other optimally-dichotomous expression variables (Table 6). These relationships held in adjusted models: low E-cadherin weighted average expression remained strongly “harmful” relative to high expression (HR=2.57, 95% CI 1.10, 6.03), while no other optimally-dichotomous marker expression variables had an effect on time-to-death.

We explored several trade-offs between strength of cut-point/time-to-death association and the number of subjects whose treatments might change due to clinical use of EMT markers. Specifically, we considered three different E-cadherin weighted average cut points that were either statistically significant or nearly so: about 0.52 (statistically-optimal value), 0.60, and 0.85 (Table 7). Setting the cut point to a value other than the statistically-optimal value led to hazard ratio point estimates that were weaker than the one at the optimal cut point, but marker

expression status was still effectively associated with outcomes at each of these cut points. Notably, the precision of hazard ratio estimates was better at cut points other than the statistically-optimal value.

The number of subjects whose treatments might change based on E-cadherin measurements—those diagnosed with local disease who are E-cadherin-negative—varied substantially with cut point. Of 99 subjects with E-cadherin measurements and diagnosed with local disease, 6 were E-cadherin-negative at the optimal cut point, 16 at a cut point of 0.60, and 56 at a cut point of 0.85.

4.4. Discussion

EMT markers can link epithelial cancer cells in primary tumors to risk for metastatic disease. We found that, when measured as a weighted average of tumor cores, E-cadherin expression in colorectal cancer cases was associated with mortality independent of stage. By identifying possible metastatic disease that might be undetected by radiologic imaging or lymph-node evaluation, measurement of E-cadherin or other EMT markers in primary tumor cancer cells has the potential to alter the treatment and outcomes of cancer patients.

An EMT marker must overcome at least two hurdles to become a useful diagnostic tool. First, expression levels measured in cancer cells from resected primary tumors must be associated with the length of time from surgery to death, independently of other relevant clinical factors. Second, the most clinically-useful definition of marker expression status must be determined. Though related, these issues are not identical: the first is statistical and the second practical. They differ because marker expression is inherently continuous, while treatment decisions are binary and therefore imply cut points. Many cut points along a continuum could

yield significant associations with time-to-death. Deciding which one to use clinically is not a matter of best model fit alone.

E-cadherin has been the most-studied EMT marker in prior reports (5, 56). Among hospital-based samples that examined survival stratified by E-cadherin expression status, four found that reduced E-cadherin was associated with worse outcomes (24, 29, 43, 68). Four studies found that E-cadherin expression by itself was not associated with survival (16, 69-71). In terms of multivariate modeling of overall survival, one study found that E-cadherin was an independent prognostic marker (43) while three concluded that it was not (29, 68, 70).

To our knowledge, there has been only one previous population-based study that measured E-cadherin in CRC primary tumor cancer cells and estimated its association with patient outcomes (72). The study assigned each specimen a separate score for membrane and cytosolic staining. They found no evidence that membrane staining was associated with either overall survival or time-to-recurrence. Cytosolic staining was associated with time-to-recurrence but not overall survival.

A meta-analysis concluded that reduced E-cadherin expression was associated with worse CRC outcomes (56). This summary must be interpreted with caution because it combined results from studies with very different definitions of E-cadherin positive/negative status. We have argued elsewhere that combining studies in this fashion is not valid (5). Nevertheless, the summary suggests that, given conflicting findings when comparing individual studies of the prognostic value of E-cadherin, overall the evidence supports its utility. Past studies do not, however, provide much clarity in deciding the best way to measure E-cadherin or what cut point to use for clinical purposes.

Our study supports the use of E-cadherin as an independent prognostic marker of CRC patient outcomes, at least when measured as a weighted average of tumor cores. While the marker had no effect on time-to-death when measured as worst (i.e. lowest) core expression, the weighted average is more representative of the tumor's overall E-cadherin status because it incorporates measurements from at least as many cells as the worst core. Notably, the association between dichotomous E-cadherin weighted average status and outcomes was independent of TNM stage. This suggests that the marker provides prognostic information beyond what is captured by tumor size, lymph-node evaluation, and radiologic imaging.

Our results can only be directly compared to other studies using the same data collection and analysis procedures in terms of type of data collected (continuous or ordinal), marker expression scale (for continuous data: average intensity, percent positive, or H scores), and cut point and covariate selection. At present, no such direct comparison with another study is possible.

Most prior studies that measured continuous E-cadherin data did so on the percent positive scale. We encourage future investigators to adopt our use of the average intensity scale for E-cadherin, for two reasons. First, the average intensity scale imposes fewer assumptions on the data than the percent-positive and H-score scales. Second, at the level of the individual cell, the percent-positive and H-score scales assign coarser expression measurements than average intensity. Altogether, these arguments suggest that the average intensity scale provides the richest, most-informative continuous scale on which to measure E-cadherin.

On the average intensity scale, we found that Integrin beta-6 and Snail had much lower expression than E-cadherin, with a high proportion of cores having expression values below the threshold for background staining (Figure 4). Therefore, we suggest that the percent positive

scale is appropriate for EMT inducers and mesenchymal markers, and would reserve the average intensity scale for epithelial markers.

Clinical implementation of an E-cadherin assay based on a continuous average intensity scale would require automated analysis of IHC slide images. Standardized controls with known E-cadherin staining intensities would have to be included with each run to ensure proper staining and analysis calibration. However, it would be possible to establish such an assay since whole-slide imaging for diagnostic purposes is widely used for the evaluation of estrogen receptor, progesterone receptor and HER2/Neu IHC stains. The College of American Pathologists recently released guidelines for validating new digital analysis assays for diagnostic use (73).

For Integrin beta-6, to our knowledge only one previous study has examined its prognostic role in CRC (23). The study reported that high expression was associated with worse survival, especially among Stages I and II patients. It found that Integrin beta-6 was an independent prognostic marker in a multivariate Cox model. These results matched what one would expect for this mesenchymal marker.

We found no evidence that Integrin beta-6 expression was associated with survival or could serve as an independent prognostic marker. Numerous differences in design and analysis could have contributed to the discrepancy between the first study and ours. The earlier study used manual IHC scoring, a different antibody, had a larger sample size (n=488), and used a different outcome for analysis (disease-related deaths compared to our use of all-cause mortality).

Our Integrin beta-6 estimates have poor precision, particularly for worst cores. It is possible that the particular antibody we used had nonspecific staining, which would be a problem

with the antibody rather than the marker. We encourage further work on Integrin beta-6 to resolve the inconsistent findings between studies.

For Snail, previous studies had highly conflicting results. Two studies found that elevated expression was associated with worse survival than low Snail expression (37, 70) while two reported no difference in survival by Snail expression status (71, 74). In multivariate time-to-death modeling, one paper reported an effect of Snail expression status on outcomes (70) and another found no effect (74).

We found no evidence that Snail expression status was associated with survival or was an independent prognostic marker. Indeed, in our study Snail failed to meet minimal expectations based on the EMT mechanism when normal tissue exhibited greater average expression than tumor tissue. Unlike Integrin beta-6, for which there is only one prior study, our finding of lack of association between Snail and outcomes can be compared to inconsistent findings across several previous studies. Although the lack of association in our study might be due to nonspecific staining as could be the case with Integrin beta-6, taken together, prior and present results suggest that Snail would not be a useful clinical marker for CRC.

Our study had a number of important strengths. First, it used tumor specimens from a population-based prospective observational study. This gives our results greater external validity than the hospital-based samples typically used in studies of EMT markers and patient outcomes (75). The only previous population-based study of this topic in CRC used tumor specimens from a single hospital that served virtually every CRC case in a defined geographic region (72). As far as we know, ours is the first study of EMT markers and CRC outcomes that collected tumor specimens from multiple clinics, with all of the variation in specimen handling and transport that

implies. This makes our finding that E-cadherin expression was associated with CRC outcomes particularly striking.

Another strength is our approach to cut point determination. After collecting continuous marker expression data, for each expression variable we performed an exhaustive automated search along the continuum for the cut point that yielded the largest difference by model fit between the hazard functions (i.e. survival experiences) of marker-positive and marker-negative subjects. Most studies have only examined a small number of possible cut points, often because they collected ordinal data but even when they collected continuous data. In addition, cut point decisions have typically been arbitrary, often using convenient percentiles or subjective judgments about high versus low staining. In contrast, our approach fully exploits the richness of continuous data, and identifies a cut point based on an objective criterion (best model fit) applied to an examination of how marker expression relates to the outcome of interest in the observed data.

Notably, of our six continuous marker expression variables, the optimal cut point of only one (E-cadherin weighted average) yielded an association with time-to-death. For a given continuum, the optimization macro identifies the observed cut point with the best model fit compared to all of the other observed cut points, but this does not guarantee that the statistically-optimal cut point will be associated with outcomes. This suggests that the optimization technique is ideal for evaluating associations between cancer biomarkers and patient outcomes. If the marker is associated with outcomes, it will find the strongest statistical association that exists in the data. If the marker is not associated with outcomes, the exhaustiveness of the procedure provides especially strong evidence that no association could be detected.

Finally, to promote standardization across studies, we selected our modeling adjustment variables based on the results of past studies and an appropriate conceptual framework of directed acyclic graph theory. Estimates of the effect of marker expression on time-to-death are only comparable across studies if the studies adjusted for the same set of covariates, preferably with the same variable coding. Thus, it is important that researchers working on studies of EMT markers and patient outcomes standardize their covariate adjustment sets across investigative teams.

Several limitations must be noted. Our study did not sample tumors in a consistent way. Ideally, each tumor would have been sampled at the invasive front, tumor center, and an edge of the tumor away from the invasive front. However, for any given tumor in our dataset, we do not know from which part of the tumor our tissue specimens came. EMT marker expression could vary throughout a tumor and it may be that, for clinical purposes, physicians should always sample a particular portion (e.g. the invasive front). Not knowing from which part of the tumor each core came, we could not calculate portion-specific estimates of, say, the association between invasive front E-cadherin and time-to-death, and separately, the association between tumor center E-cadherin and time-to-death.

A second limitation is that our outcome in statistical models was time to all-cause mortality. This may have led to weaker associations than would have been observed with an outcome of time to cancer-specific mortality or time to recurrence.

Finally, our cut point optimization technique is an informative method for determining whether an association exists between marker expression and patient outcomes. However, finding an association using the statistically-optimal cut point based on best model fit does not automatically mean that one has found the best cut point for clinical purposes. The patients most

likely to benefit from introducing EMT markers into clinical practice are those diagnosed with local disease according to lymph-node evaluation and radiologic imaging, but whose EMT marker measurements suggest poor prognosis. These patients generally would not receive chemotherapy based on conventional staging (15), but their stage and treatments might be reconsidered in light of their EMT marker status.

For example, consider how the cross-tabulation of E-cadherin status and stage changes as the cut point varies (Table 7). Of 99 subjects with E-cadherin measurements and diagnosed with local disease, 16 were E-cadherin-negative at a cut point of 0.60 versus 6 such subjects at the statistically-optimal cut point of about 0.52. The hazard ratio point estimate for a cut point of 0.60 is nearly as strong as the point estimate for the best-fitting cut point. Primary tumors with E-cadherin values between 0.52 and 0.60 probably are not biologically much different with respect to cancer cell detachment from tumors with values below 0.52. Nevertheless, a notable number of subjects diagnosed with local disease had E-cadherin values just above the statistically-optimal cut point and would not have their treatments changed if the clinical cut point were set at the statistically-optimal value. In addition, the cross-tabulation of stage and E-cadherin status fits what is known about disease recurrence better at a cut point of 0.60 than 0.52, as about 25% of diagnosed Stage I/II CRC patients experience recurrence after surgery (4). Altogether, these considerations suggest that cut point model fit should be one of several criteria used to determine a clinical cut point, but not the sole criterion.

In sum, we believe that our methods for marker expression measurement, cut point identification, and covariate selection provide informative results and hope that future studies of EMT markers and patient outcomes will adopt them. By doing so, investigators would make

possible meaningful comparisons of results from different research groups and valid meta-analyses.

Our results suggest that E-cadherin could be a useful marker to identify colorectal cancer patients at risk for metastatic disease, even among those who appear to have local disease according to lymph-node evaluation and radiologic imaging. Clinical measurement of E-cadherin expression in primary tumor cancer cells might increase the accuracy of stage diagnosis, thereby altering the treatments that some patients receive and improving their outcomes.

Table 3. Subject characteristics for overall North Carolina CanCORS and subset in whose primary tumors EMT markers were measured

Characteristic Mean (SD) or N (%) ^a	Overall NC-CanCORS (N=990)		EMT Study Sample (N=190)		P-value
Age at Baseline (years)	66.3	13.2	66.6	13.1	0.8
Sex					0.6
Male	488	49.3%	90	47.4%	
Female	502	50.7%	100	52.6%	
Race					0.6
Non-Hispanic White	762	77.0%	150	78.9%	
Other	228	23.0%	40	21.1%	
Tumor Stage					0.9
Local	437	51.1%	100	52.6%	
Regional	311	36.4%	66	34.7%	
Distant	107	12.5%	24	12.6%	
Missing/Noninformative	135		0		
Tumor Location					0.07
Proximal	279	42.9%	82	50.9%	
Distal	371	57.1%	79	49.1%	
Missing/Noninformative	340		29		
Tumor Grade					0.4
Well/Moderately Differentiated	539	83.7%	129	81.1%	
Poorly/Not Differentiated	105	16.3%	30	18.9%	
Missing/Noninformative	346		31		
Neoadjuvant Chemotherapy					0.3
Yes	46	8.5%	8	5.7%	
No	496	91.5%	132	94.3%	
Missing/Noninformative	448		50		
Neoadjuvant Radiation					0.09
Yes	41	7.6%	5	3.6%	
No	501	92.4%	135	96.4%	
Missing/Noninformative	448		50		
Died Within 5 Years of Baseline	366	37.0%	62	32.6%	0.3

^aPercentages are for informative non-missing data.

Table 4. Average continuous EMT marker expression in tumor tissue compared to normal adjacent tissue

Marker	Algorithm	Scale	Tissue Type	N	Marker Expression					
					<u>Weighted Average^a</u>			<u>Worst Core^b</u>		
					Mean	SD	P-value	Mean	SD	P-value
E-cadherin	Membrane+ Cytoplasm	Average Intensity (0-3)	Tumor	188	0.84	0.22	<0.0001	0.74	0.24	<0.0001
			Normal	181	0.98	0.22		0.87	0.28	
Integrin Beta-6	Membrane	Percent Positive Cells (0-100)	Tumor	181	58.1	26.8	<0.0001	69.9	25.4	<0.0001
			Normal	177	38.5	25.6		52.1	29.2	
Snail	First Nuclear ^c	Percent Positive Nuclei (0-100)	Tumor	185	41.7	22.1	0.02	52.3	24.9	0.0004
			Normal	173	47.9	27.9		62.1	27.2	
Snail	Second Nuclear ^c	Percent Positive Nuclei (0-100)	Tumor	185	44.4	26.5	0.08	57.0	29.5	0.02
			Normal	173	49.5	27.7		64.1	28.0	
Snail	Whole Cell ^c	Percent Positive Cells (0-100)	Tumor	185	34.3	24.0	0.005	46.6	28.7	0.004
			Normal	173	41.9	26.4		55.3	28.3	

^aMarker expression values assigned as weighted average of cores by tissue type (weighted by area analyzed for E-cadherin, number of cells for Integrin, and number of nuclei/cells for Snail).

^bMarker expression values assigned as expression by tissue type of the core with lowest expression for E-cadherin and highest expression for Integrin and Snail.

^cSeveral different Snail scoring algorithms were developed and applied to all tissue microarrays. See Section 4.2.3 (pp. 62-64) for details.

Table 5. Optimally-dichotomized marker expression status cross-tabulated with tumor stage and with risk of dying within 5 years of surgery^{a,b}

Stage	<u>E-cadherin</u>				<u>Integrin beta-6</u>				<u>Snail</u>			
	<u>Weighted Average</u>		<u>Worst Core</u>		<u>Weighted Average</u>		<u>Worst Core</u>		<u>Weighted Average</u>		<u>Worst Core</u>	
	<u>Neg.</u>	<u>Pos.</u>	<u>Neg.</u>	<u>Pos.</u>	<u>Neg.</u>	<u>Pos.</u>	<u>Neg.</u>	<u>Pos.</u>	<u>Neg.</u>	<u>Pos.</u>	<u>Neg.</u>	<u>Pos.</u>
Local	6	93	3	96	10	82	5	87	29	69	54	44
Regional	5	61	3	63	3	63	1	65	14	51	45	20
Distant	0	23	0	23	1	22	0	23	1	21	13	9
<u>Mortality Risk</u>												
Died within 5 Years	7	55	3	59	6	55	0	61	17	44	43	18
Total	11	177	6	182	14	167	6	175	44	141	112	73
5-Year Risk of Death Post-Surgery (%)	64	31	50	32	43	33	0	35	39	31	38	25

^aOf 190 overall subjects in the EMT study, the number with tumor tissue marker data available was 188 for E-cadherin, 181 for Integrin, and 185 for Snail.

^bMarker-negative status is expression in tumor tissue below the statistically-optimal cut point (low expression) and marker-positive status is expression in tumor tissue at or above the statistically-optimal cut point (high expression). The *a priori* hypotheses were that E-cadherin-negative subjects would have worse outcomes while Integrin-positive and Snail-positive subjects would have worse outcomes.

Neg.=marker-negative status, Pos.=marker-positive status

Table 6. Unadjusted and adjusted Cox proportional hazards models of the effect of optimally-dichotomized marker expression status on time-to-death censored at 5 years after surgery (n=190)^a

Variable	Weighted Average ^b				Worst Core ^c			
	Unadjusted		Adjusted ^d		Unadjusted		Adjusted ^d	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
E-cadherin^e	2.84	1.29, 6.28	2.57	1.10, 6.03	1.53	0.47, 4.95	1.56	0.32, 7.63
Integrin beta-6^f	0.68	0.29, 1.59	0.45	0.19, 1.08	-- ^g	--	-- ^g	--
Snail^f	0.83	0.48, 1.45	0.61	0.33, 1.13	0.62	0.36, 1.08	0.57	0.32, 1.02

^aPrior to dichotomization, continuous E-cadherin measured as core average intensity (0-3) and continuous Integrin beta-6 and Snail measured as core percent positive cells (Integrin) or nuclei (Snail) (0-100). Subjects with missing data for a given marker expression variable had values imputed to retain full sample size. For the value of the statistically-optimal cut point for each marker expression variable, see Materials and Methods section.

^bMarker expression values assigned as weighted average of tumor cores (weighted by area analyzed for E-cadherin, number of cells for Integrin, and number of nuclei for Snail).

^cMarker expression values assigned as average intensity of the subject's tumor core with the lowest average intensity for E-cadherin, or as percent positive cells/nuclei of the subject's tumor core with the highest percent positive cells/nuclei for Integrin or Snail.

^dAdjusted for age (continuous), TNM stage (local/regional/distant), neoadjuvant chemotherapy (yes/no), and neoadjuvant radiation therapy (yes/no). Expression status for a given marker was not adjusted for the other markers.

^eComparison is E-cadherin-negative (low expression) to E-cadherin-positive (high expression).

^fComparison is marker-positive (high expression) to marker-negative (low expression).

^gNot estimable due to no deaths within 5 years of surgery among those classified as Integrin-negative.

Table 7. Trade-offs between strength of cut-point/time-to-death association and number of patients whose treatments would be reassigned, by stage distribution and 5-year risk of death for E-cadherin weighted average status^a

Cox Model Estimate^b	Cut Point 0.52^c		Cut Point 0.60		Cut Point 0.85	
	<u>HR</u>	<u>95% CI</u>	<u>HR</u>	<u>95% CI</u>	<u>HR</u>	<u>95% CI</u>
	2.57	1.10, 6.03	2.40	1.29, 4.49	1.75	0.99, 3.08
Stage	<u>E-cad Negative</u>	<u>E-cad Positive</u>	<u>E-cad Negative</u>	<u>E-cad Positive</u>	<u>E-cad Negative</u>	<u>E-cad Positive</u>
Local (n=99)	6	93	16	83	56	43
Regional (n=66)	5	61	11	55	41	25
Distant (n=23)	0	23	1	22	11	12
<u>Mortality Risk</u>						
Died Within 5 Years (n=62)	7	55	14	48	42	20
Total (n=188)	11	177	28	160	108	80
5-Year Risk of Death Post-Surgery (%)	63.6	31.1	50.0	30.0	38.9	25.0
<u>Mortality Risk Effect Estimates^d</u>	<u>Estimate</u>	<u>95% CI</u>	<u>Estimate</u>	<u>95% CI</u>	<u>Estimate</u>	<u>95% CI</u>
Risk Difference (%)	32.6	3.3, 61.8	20.0	0.2, 39.8	13.9	0.7, 27.1
Risk Ratio	2.05	1.25, 3.37	1.67	1.07, 2.59	1.56	0.99, 2.43

^aContinuous E-cadherin measured as weighted average (weighted by area analyzed) of tumor core average intensities (0-3) prior to dichotomization. Of 190 subjects, 2 had missing data for E-cadherin, but multiple imputation enabled retention of these 2 in Cox models.

^bFor effect of dichotomous E-cadherin expression (negative versus positive) on time-to-death, adjusted for age (continuous), TNM stage (local/regional/distant), neoadjuvant chemotherapy (yes/no), and neoadjuvant radiation treatments (yes/no).

^cStatistically-optimal cut point by best model fit.

^dUnadjusted comparisons of marker-negative versus marker-positive expression status.

Figure 11. Kaplan-Meier survival stratified by levels of dichotomous E-cadherin weighted average expression status

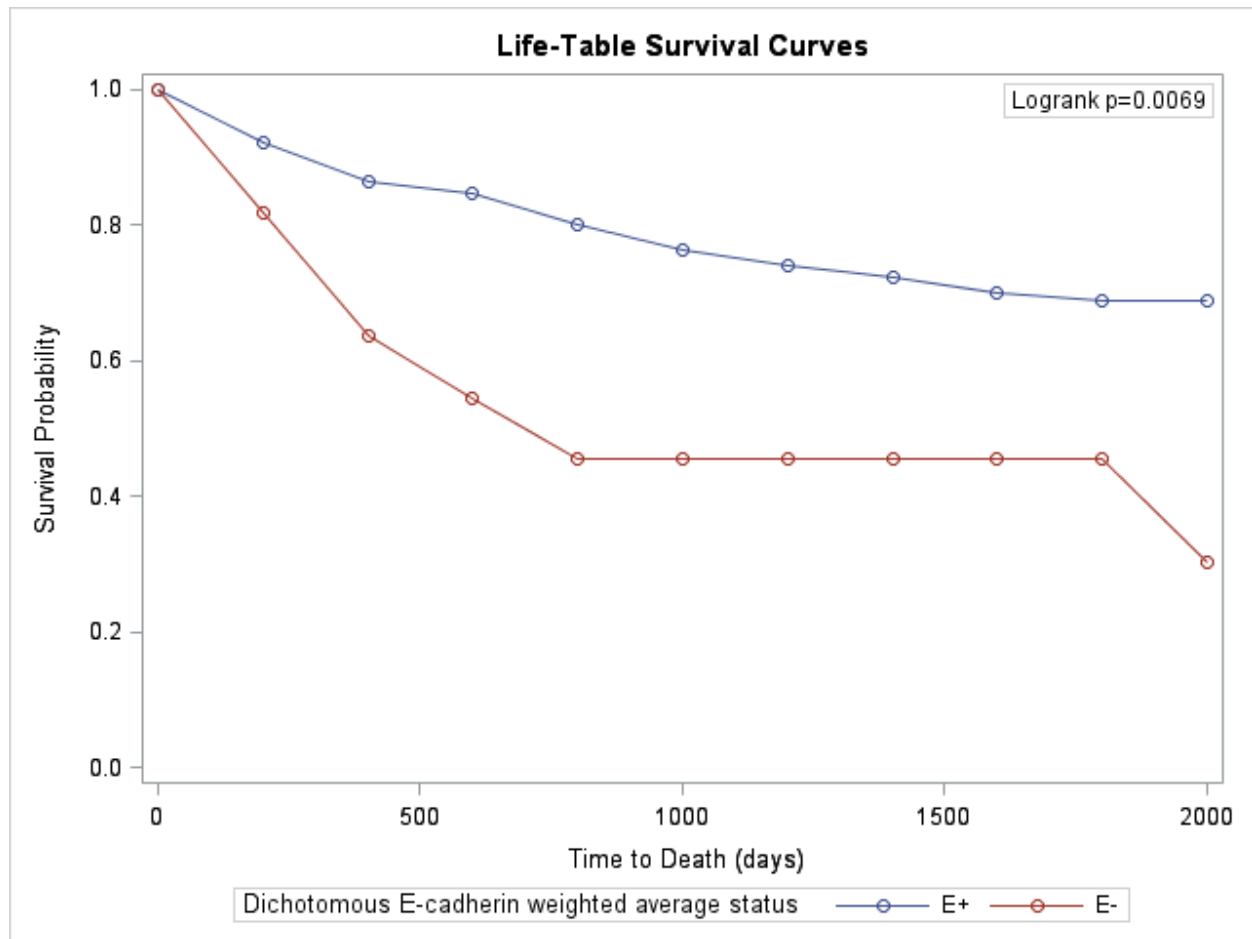


Figure 12. Kaplan-Meier survival stratified by levels of dichotomous E-cadherin worst core expression status

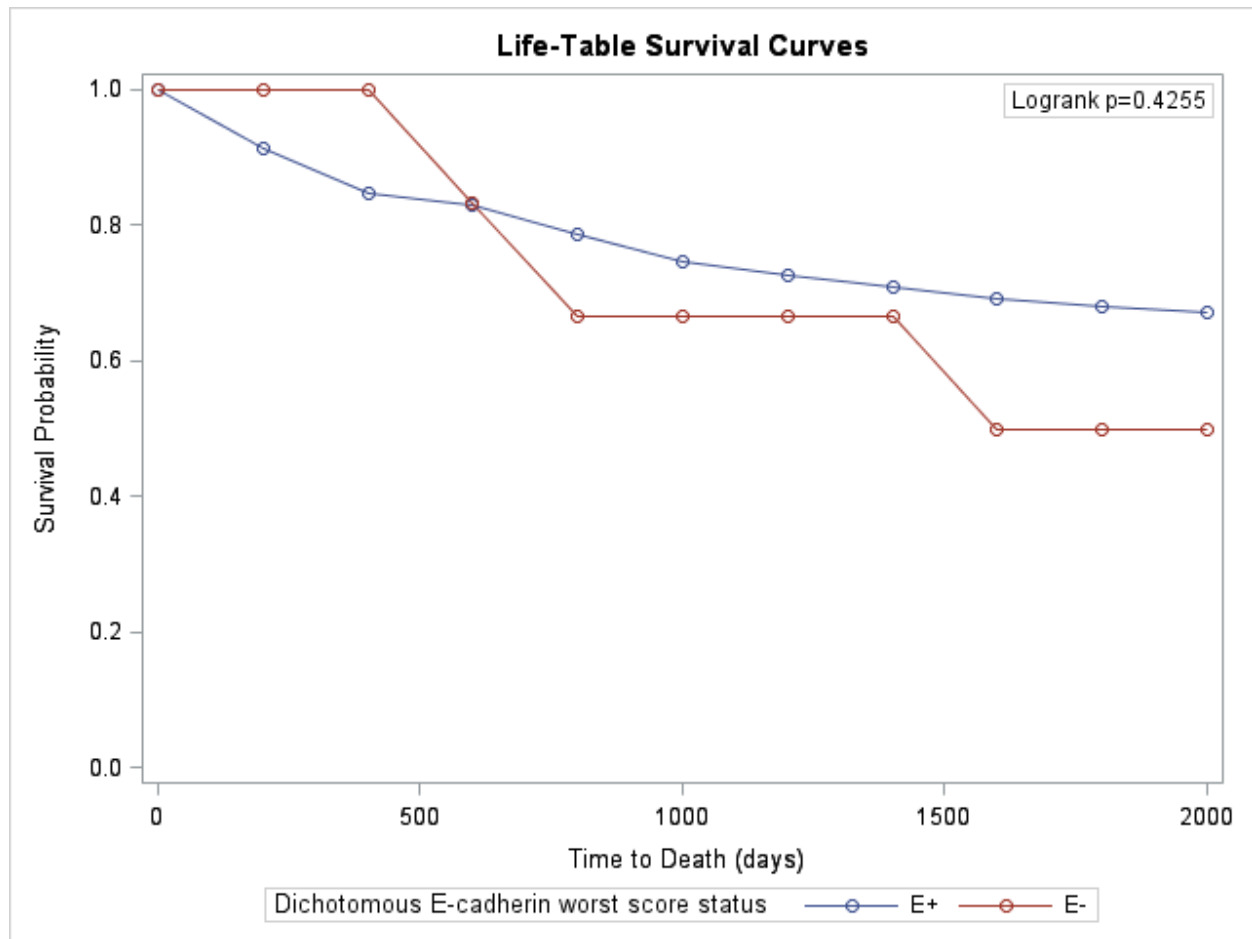


Figure 13. Kaplan-Meier survival stratified by levels of dichotomous Integrin beta-6 weighted average expression status

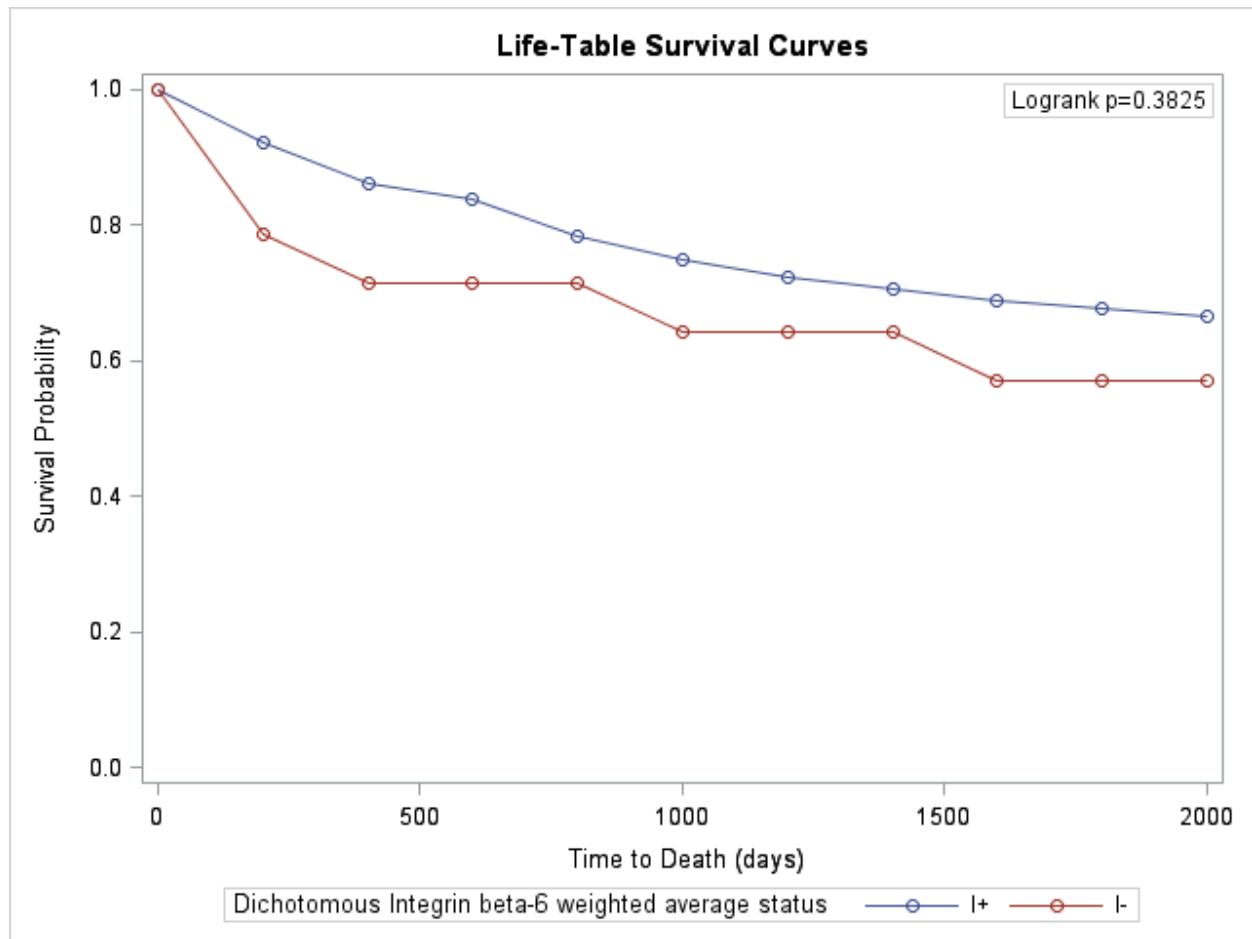


Figure 14. Kaplan-Meier survival stratified by levels of dichotomous Integrin beta-6 worst core expression status

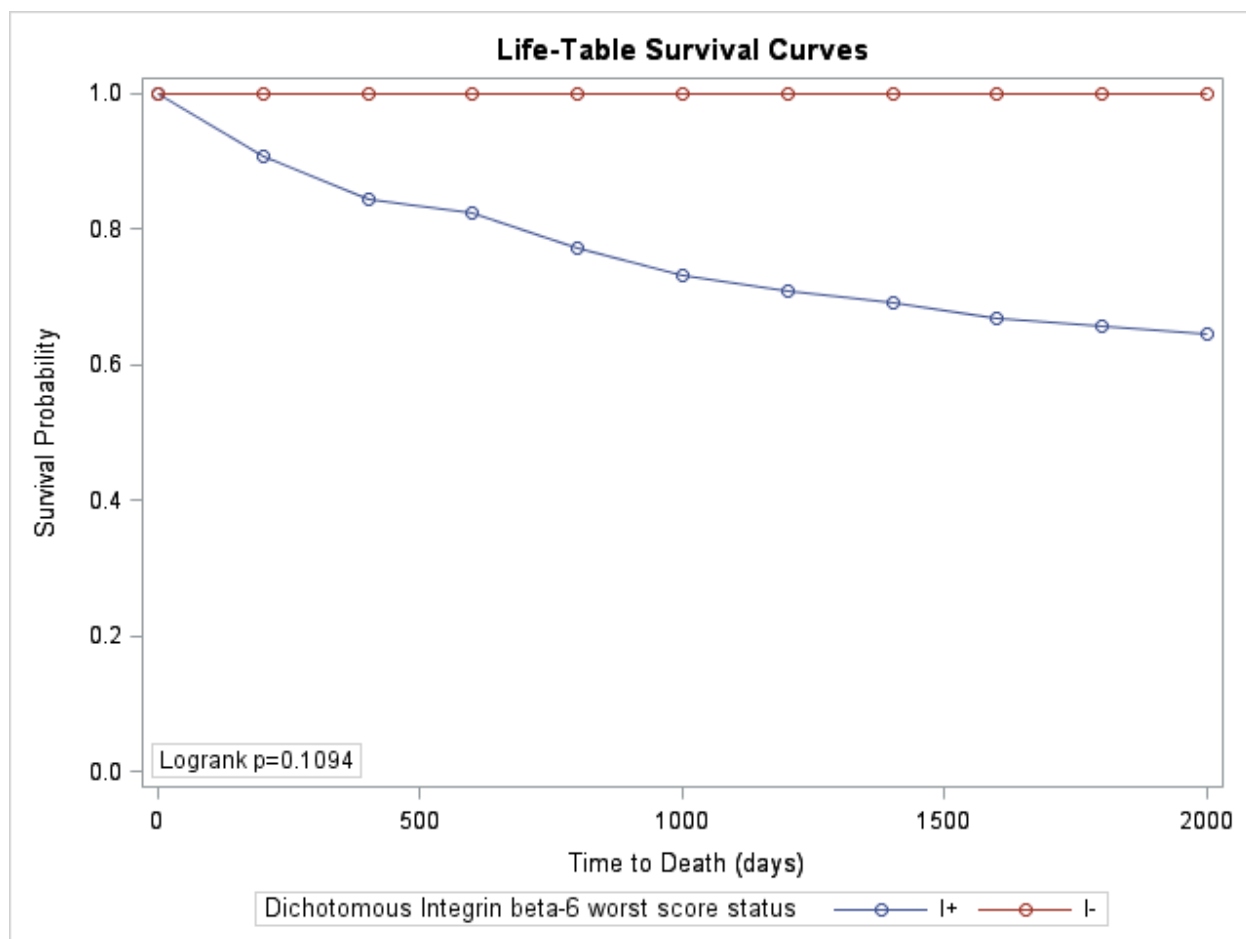


Figure 15. Kaplan-Meier survival stratified by levels of dichotomous Snail weighted average expression status

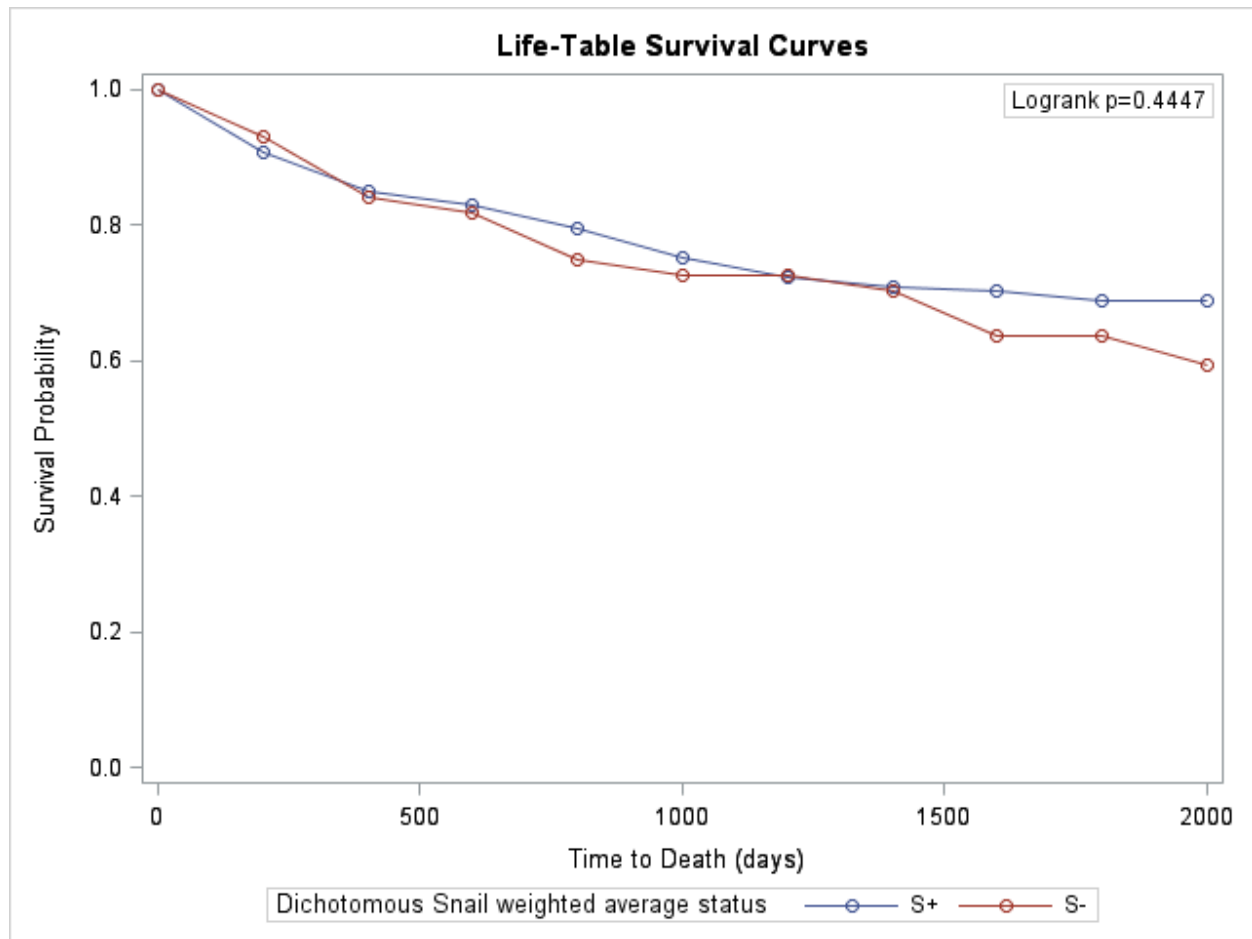


Figure 16. Kaplan-Meier survival stratified by levels of dichotomous Snail worst core expression status

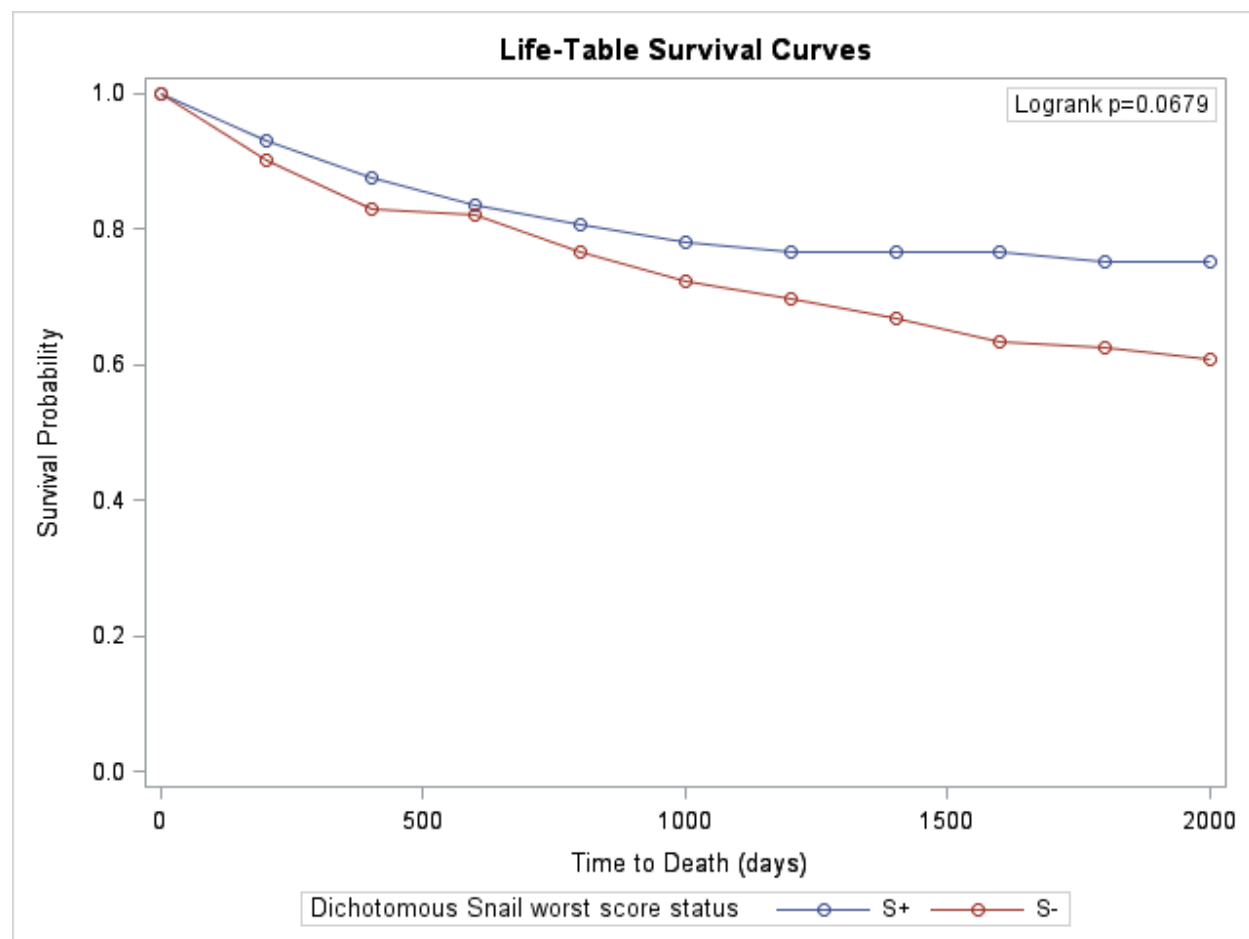


Figure 17. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin weighted average expression status and dichotomous Integrin beta-6 weighted average expression status

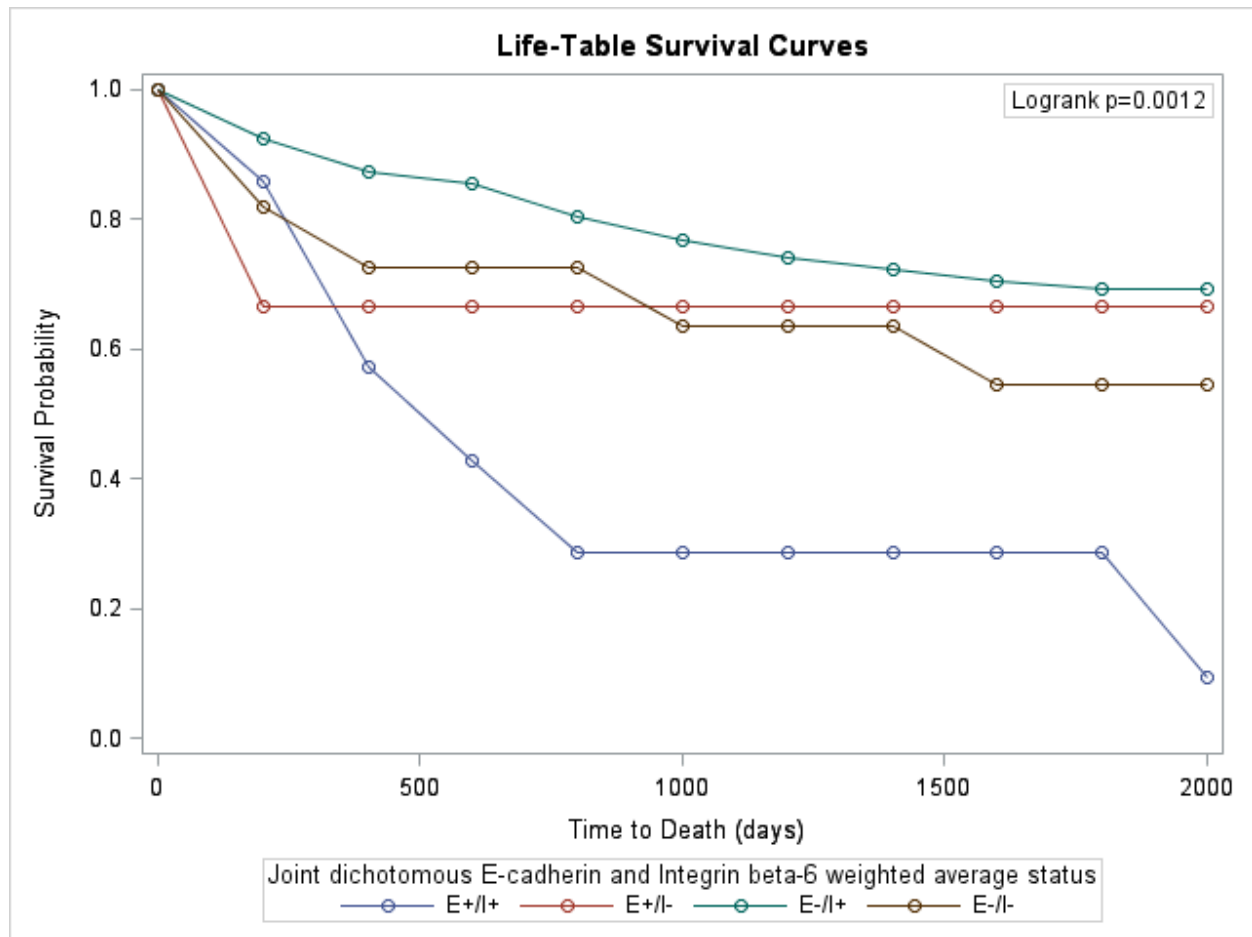


Figure 18. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin worst core expression status and dichotomous Integrin beta-6 worst core expression status

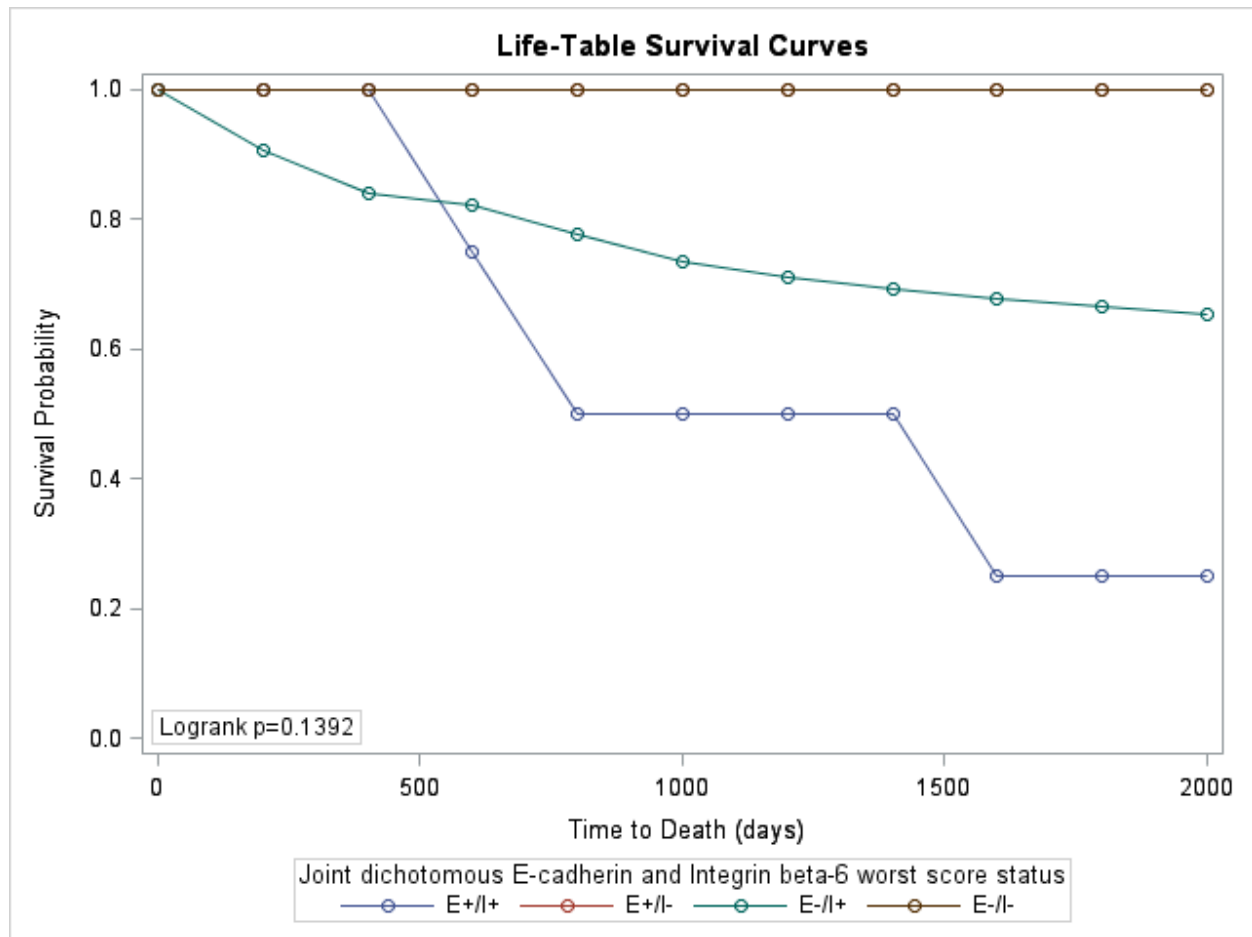


Figure 19. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin weighted average expression status and dichotomous Snail weighted average expression status

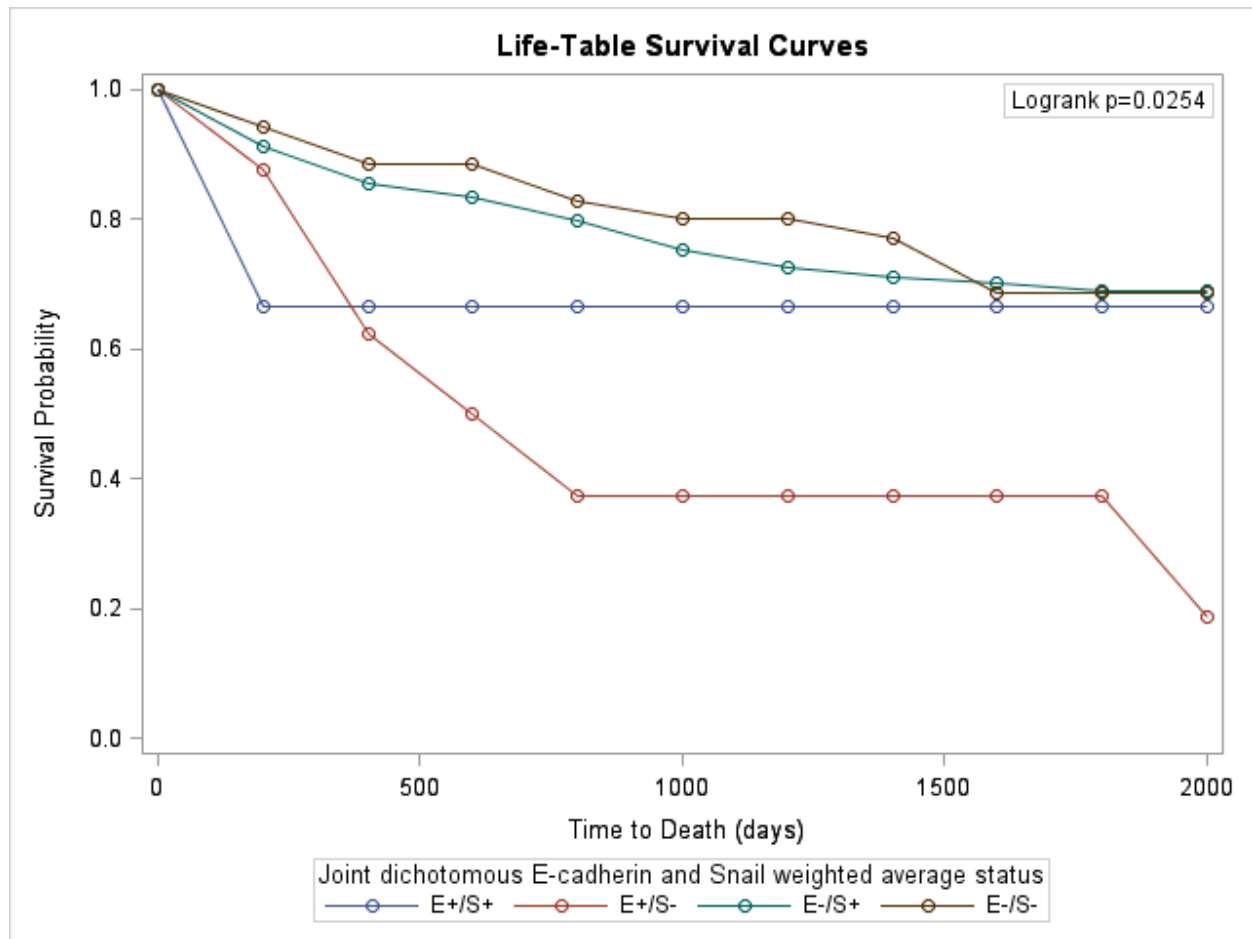


Figure 20. Kaplan-Meier survival jointly stratified by levels of dichotomous E-cadherin worst core expression status and dichotomous Snail worst core expression status

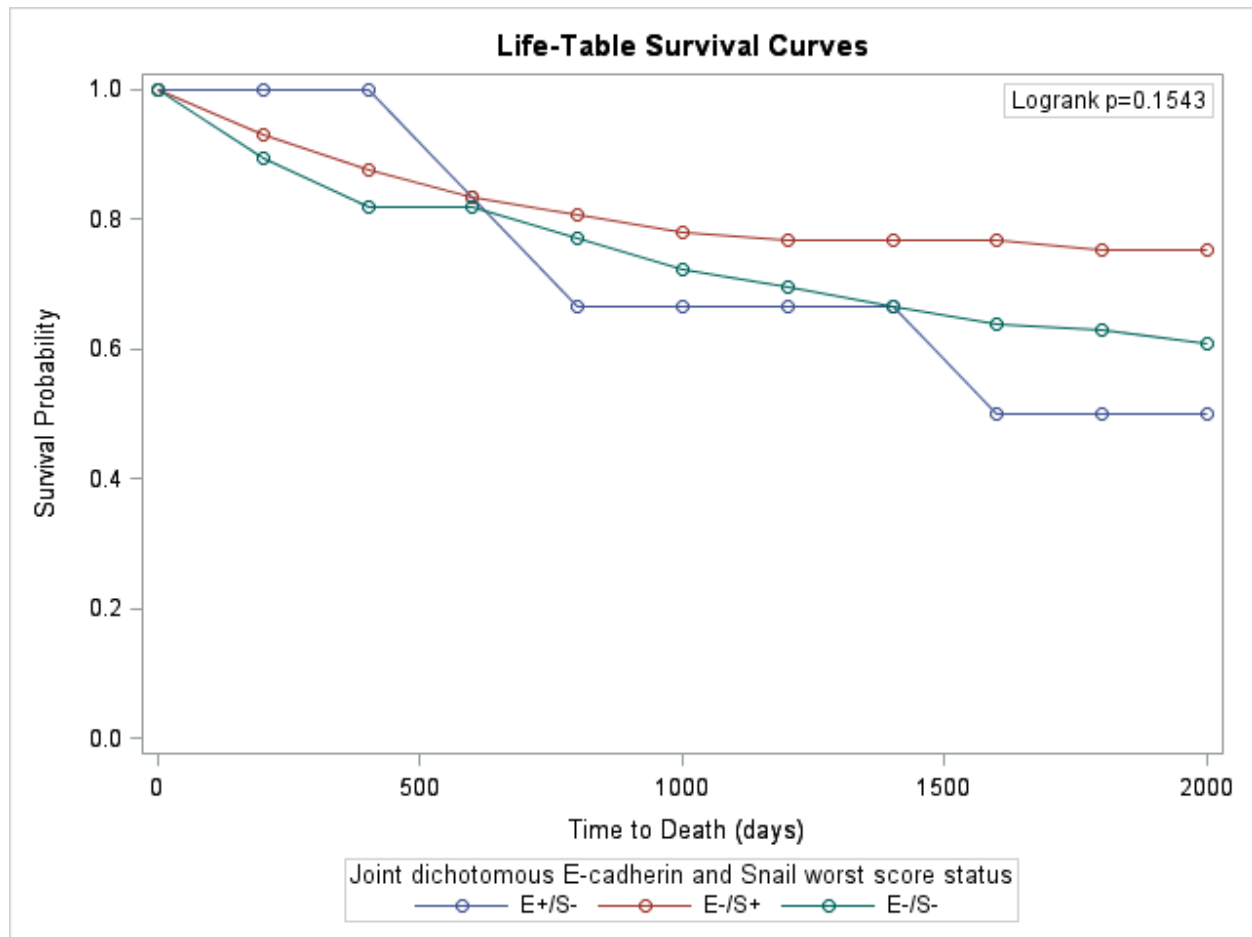


Figure 21. Kaplan-Meier survival jointly stratified by levels of dichotomous Integrin beta-6 weighted average expression status and dichotomous Snail weighted average expression status

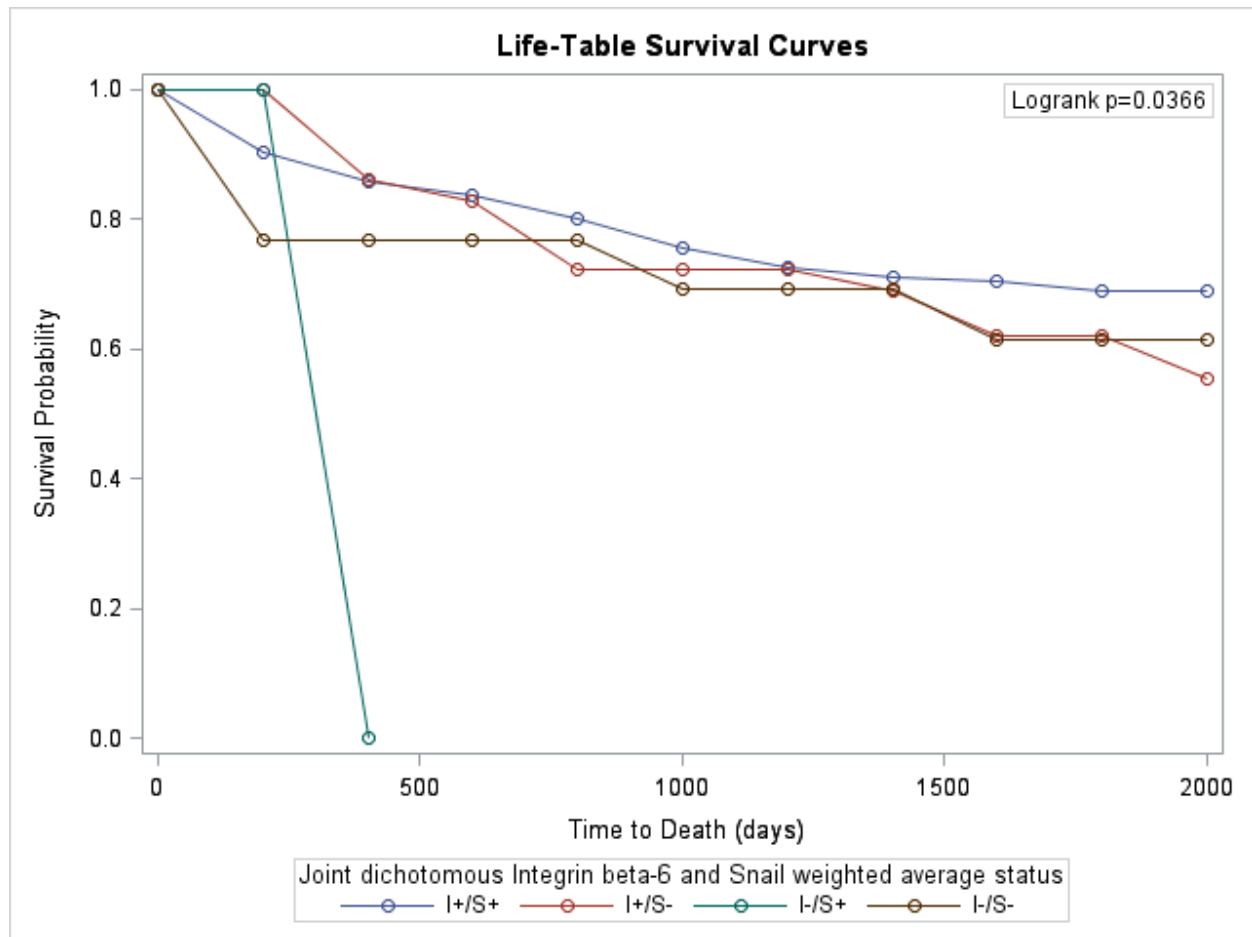
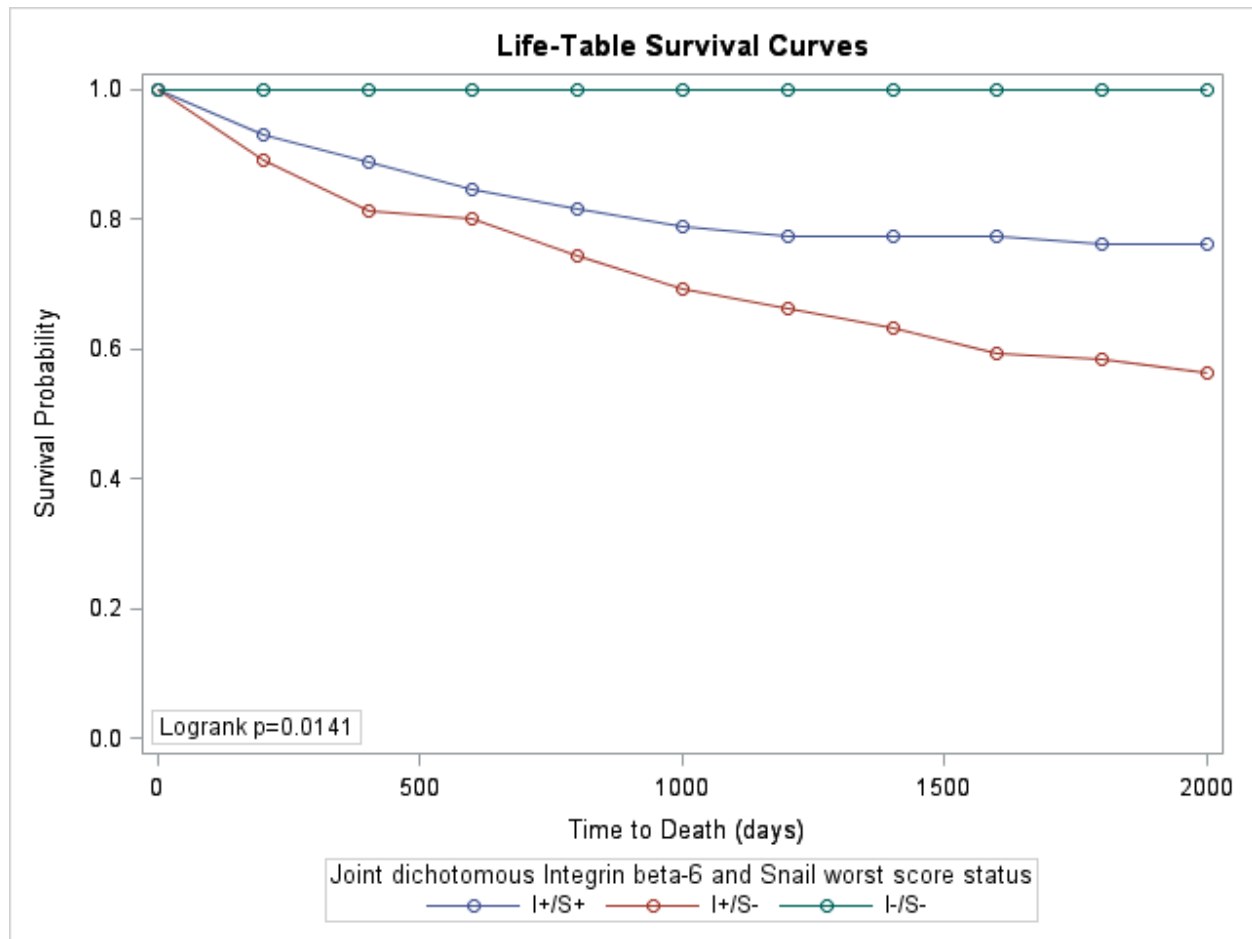


Figure 22. Kaplan-Meier survival jointly stratified by levels of dichotomous Integrin beta-6 worst core expression status and dichotomous Snail worst core expression status



CHAPTER 5. ESTIMATING SENSITIVITY AND SPECIFICITY OF MARKERS OF EPITHELIAL-MESENCHYMAL TRANSITION TO EVALUATE CANCER CELL DETACHMENT FROM PRIMARY TUMORS: A LATENT CLASS ANALYSIS

5.1. Introduction

Metastases are responsible for about 90% of cancer deaths (1). An essential component of cancer diagnosis is accurately assessing whether cancer cells have detached from the primary tumor because this impacts adjuvant therapy decisions.

Currently, physicians use two diagnostic tests to assess cancer cell detachment: examination of lymph nodes near the primary tumor and radiologic imaging. While highly useful, these methods do not always successfully detect metastases. For example, roughly 25% of colorectal cancer patients diagnosed with local disease later experience recurrence (4). Most of these recurrences are likely due to metastases that were present at the time of diagnosis but were too small to be detected by imaging or lymph node evaluation. Adding a third test could substantially reduce the number of joint false negative results across the entire panel of tests.

Markers of epithelial-mesenchymal transition (EMT), a mechanism of metastasis, might be able to serve this role (5). Roughly 80% of cancer is epithelial (1) and EMT involves epithelial cells temporarily transforming into mesenchymal cells by decreasing expression of epithelial markers and increasing expression of mesenchymal markers (2, 14). EMT marker expression levels measured in primary tumor cancer cells could suggest whether the tumor contains a substantial number of cells capable of breaking off from the tumor, thereby putting the patient at risk for metastatic disease (5).

Previous studies have shown that EMT marker levels in primary tumor cancer cells are associated with patient outcomes (5, 24, 29, 43, 56, 68). These studies did not, however, evaluate the diagnostic accuracy of EMT markers in terms of sensitivity and specificity, which is important to decide whether to add them to the panel of tests of cancer cell detachment.

Standard calculations of sensitivity and specificity require comparing the new test to a gold standard (58). However, this condition cannot be reasonably met for EMT markers because imaging and lymph node evaluation are unsuccessful too often to qualify as gold standard measures of metastatic disease. Latent class models provide a way to estimate sensitivity and specificity of each item in a panel of diagnostic tests without assuming that any of the tests is a gold standard (61). In this analysis, we used latent class models to estimate the sensitivity and specificity of EMT markers to assess cancer cell detachment from primary tumors in a cohort of colorectal cancer patients.

5.2. Methods

5.2.1. Study population

Subjects were 188 cancer patients enrolled in the Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) for whom the EMT marker E-cadherin was measured in primary tumor specimens (76). CanCORS was a population-based, case-only, multi-site prospective cohort study of lung and colorectal cancer that enrolled subjects during 2003-06 (48). All subjects in this analysis were colorectal cancer patients from North Carolina. Survey and medical records data were collected.

5.2.2. Latent class models

Latent variable models attempt to identify subgroups within a population by postulating that the subgroups represent different levels of an unobserved, error-free latent variable (62).

The latent variable is measured indirectly by multiple observed indicator variables (also called manifest variables), each of which is conceived as being generated jointly by its own random error term and the latent variable. Individuals' response patterns to the indicators can be used to estimate the prevalence of each latent subgroup (gamma parameter or class-membership probability) and, within each latent subgroup, to estimate the probability of a given response to each indicator (rho parameter or item-response probability). Latent class models are latent variable models that use categorical indicators and postulate a categorical latent variable. An important practical issue in implementing any given latent class model is deciding which parameters to freely estimate and which, if any, to assign a fixed, a priori value.

Figure 10 shows our conceptual model of the relationship between cancer cell detachment from the primary tumor and the tests used to assess it. Detachment was the latent variable, that is, the phenomenon of interest that was not observed directly. EMT markers, lymph node evaluation, and imaging were the indicators.

5.2.3. Diagnostic tests of cancer cell detachment from primary tumors

Each test was treated in models as binary test-positive versus test-negative. Test-positive meant evidence supporting cancer cell detachment from the primary tumor while test-negative meant no evidence supporting cancer cell detachment.

EMT

EMT in primary tumor cancer cells was measured using E-cadherin (76), an epithelial membrane protein that plays a crucial role in adhesion between adjacent epithelial cells (1, 2). An epithelial cancer cell undergoing EMT will downregulate E-cadherin expression, making low expression suggestive of increased risk for cell detachment from the tumor (2). E-cadherin was measured using immunohistochemistry in tissue microarrays. We reported previously that, when

E-cadherin expression was measured as a weighted average of tumor cores among these subjects, low expression was associated with greater risk of death within 5 years of surgery than was high expression (76).

E-cadherin was measured on a continuous average intensity scale of 0-3. This allowed us to explore the impact on EMT sensitivity and specificity of setting different cut points to define high expression versus low expression. For any given cut point, EMT-positive status meant low E-cadherin expression (below the cut point) and EMT-negative status meant high E-cadherin expression (at or above the cut point). Both EMT sensitivity and specificity were freely estimated in all models.

Lymph Node Evaluation and Radiologic Imaging

Each latent class model had seven parameters: one class-membership and six item-response parameters (sensitivity and specificity for each of EMT, lymph node evaluation, and imaging). Freely estimating all seven parameters would have led to poor model specification (see Analysis section below). In most models, we fixed (i.e. set or restricted) two of the six item-response parameters—either the sensitivities of lymph node evaluation and imaging or the specificities of lymph node evaluation and imaging—while freely estimating the other four as well as the class-membership parameter. This approach required observed test results for lymph node evaluation and imaging for freely estimated parameters as well as determination of fixed values for restricted parameters.

Observed lymph node evaluation and imaging results were not available in CanCORS. As an approximation, we inferred them from subject tumor stage using the rules presented in Table 2. These inferred test results were used for freely estimated lymph node evaluation and imaging sensitivity and specificity parameters.

In models where the sensitivities of lymph node evaluation and imaging were fixed, we assigned a value of 60% or 80% for each fixed parameter. Nationwide clinical data suggested that, between lymph node evaluation and imaging, the average sensitivity of each of these two tests by itself is about 65% (see Discussion for details) (4, 63). The values of 60% and 80% were chosen to represent bounds to assess the influence on model results of different combinations of assumptions about high or low sensitivity for lymph node evaluation and imaging. For models in which the specificities of lymph node evaluation and imaging were fixed, we assigned bounds of 90% or 100% for each fixed parameter. These specificity bounds reflected our assumption that false positive results for each of these tests are rare.

Across models, we varied the combinations of fixed values and of which lymph node evaluation and imaging parameters were restricted to observe how doing so impacted the EMT estimates. For each EMT cut point, we also ran a model in which all seven parameters were freely estimated.

Whether freely estimated or restricted, every lymph node evaluation or imaging parameter represented a binary test. For interpretation, lymph node-positive meant that cancer cells were found in lymph nodes near the primary tumor and lymph node-negative that no cancer cells were found there. Imaging-positive meant detection of a metastasis via imaging and imaging-negative indicated no such detection of metastasis via imaging.

5.2.4. Analysis

Every model included all three tests and estimated a 2-class solution distinguishing high risk from low risk for metastatic disease. Making parameter estimates interpretable as sensitivity or specificity required binary tests and grouping the data into two latent classes.

A properly specified latent class model has a positive number of degrees of freedom (df), given by $df = W - P - 1$, where W is the number of possible response patterns and P is the number of parameters estimated (62). Three binary indicators yielded $2^3 = 8$ possible response patterns. When all class-membership and item-response parameters were freely estimated, a 2-class model with 3 binary indicators estimated 7 parameters (1 class-membership and 6 item-response), leaving 0 df.

Using fixed sensitivity or specificity values for lymph node evaluation and imaging reduced the number of parameters estimated in a given model, thereby giving positive degrees of freedom. Fixing the values of two item-response parameters in a model yielded 2 df. Across models, different combinations of fixed lymph node evaluation and imaging values were used to assess the impact of different assumptions about low or high accuracy of those tests on EMT estimates. We also evaluated models in which all parameters were freely estimated, i.e. 0 df.

We have shown previously that, when E-cadherin expression is measured as a weighted average of tumor cores on a continuous average intensity scale of 0-3, values of 0.52, 0.60, and 0.85 are three possibilities for selection of a clinical cut point to distinguish low-risk from high-risk patients (76). We ran separate models using each of these values to create the EMT indicator.

All analyses were performed using SAS 9.3 (SAS Institute, Cary, NC), with PROC LCA used for the latent class analyses (64). The SAS procedure used an iterative expectation-maximization algorithm to obtain maximum likelihood estimates for freely-estimated parameters.

5.3. Results

The study sample was mainly non-Hispanic whites, about evenly divided by sex, and had a mean age close to the national average for colorectal cancer (63) (Table 8). Of 188 subjects, 90 (48%) tested positive for at least one of lymph node evaluation or imaging. This is comparable to the approximately 56% of colorectal cancer cases in the United States diagnosed with regional or distant disease (63).

At an EMT cut point of 0.52, the specificity of the EMT diagnostic test was over 90% across all variations in lymph node evaluation and imaging parameter restrictions (Table 9). However, the sensitivity under these conditions was never greater than 6%.

For an EMT cut point of 0.60, specificity was somewhat lower than at a cut point of 0.52, but remained greater than 80% in all variations tested (Table 10). Sensitivity was consistently higher in these models than those at a cut point of 0.52, but peaked at 14%.

At an EMT cut point of 0.85, specificity dropped substantially from what it was at a cut point of 0.60, falling to about 40% across variations in parameter restrictions (Table 11). Sensitivity was greater at this cut point than at the others, generally around 50-60%.

The model at which the EMT marker could reduce the number of false local diagnoses to the greatest extent used a cut point of 0.85. When lymph node evaluation and imaging specificity were both fixed at 90%, EMT sensitivity and specificity were estimated to be 59% and 44%, respectively. Notably, though 59% is low for optimal sensitivity, the freely-estimated sensitivities of lymph node evaluation and imaging in this model were 97% and 26%, respectively.

For a given set of lymph node evaluation and imaging parameter restrictions, the estimated prevalence of the latent class at high risk for metastatic disease was very consistent

across EMT cut points (Tables 9-11). The prevalence of the high risk group was low when lymph node evaluation and imaging sensitivities were fixed (about 16-21%) but increased substantially when lymph node evaluation and imaging specificities were fixed (41-53%). For each EMT cut point, freely estimating all parameters (i.e. no lymph node evaluation and imaging parameter restrictions) produced results roughly equivalent to the worst-performing model with restrictions at that cut point.

5.4. Discussion

We used latent class analysis to estimate the sensitivity and specificity of EMT markers to evaluate cancer cell detachment from primary tumors under varying assumptions about the accuracy of lymph node evaluation and imaging to assess the same. EMT outperformed the freely-estimated parameters of the other tests in some scenarios and EMT specificity was over 90% in several models, while the peak sensitivity was 59%.

In the United States there are around 140,000 new cases of colorectal cancer annually (77). Of these, about 40% (~56,000 cases) are diagnosed as having local disease and 60% (~84,000 cases) as having some form of metastatic disease, whether regional or distant (63). Among those diagnosed with local disease, roughly 25% (~14,000 cases) later experience recurrence and therefore likely had undetected metastases at the time of diagnosis (4). This implies that about 10% of all new colorectal cancer cases each year consist of patients who have metastatic disease at diagnosis but who test falsely negative for both lymph node evaluation and imaging. A further implication is that, of 140,000 annual incident cases, about 98,000 truly have metastatic disease at the time of diagnosis rather than the 84,000 currently diagnosed as such. Thus, among those who truly have metastatic disease, false diagnoses would account for

14,000/98,000=14%. In other words, the joint false negative proportion for lymph node evaluation and imaging would be 14% and therefore their joint sensitivity would be 86%.

Since a diagnosis of local disease requires testing negative on every measure of cancer cell detachment, adding a third test to the panel could substantially reduce the number of these diagnostic failures. If the third test has a sensitivity of 70%, and therefore a false negative proportion of 30%, it would cut the joint false negative proportion for the entire panel from 14% to 4%. In other words, the third test would reduce the annual number of false diagnoses of local disease from 14,000 to a little over 4,000.

A necessary trade-off for this benefit is that the number of false positives would increase. Some patients who truly do not have metastatic disease and who test negative for both lymph node evaluation and imaging would test positive for EMT. However, the clinical consequences of false negative and false positive diagnoses of metastatic disease must be weighed against each other. False positives lead to administration of chemotherapy when it is unlikely to benefit the patient. False negatives generally lead to withholding chemotherapy when the patient truly has metastatic disease, which could be fatal.

Our analysis had several important strengths. First, we estimated EMT sensitivity and specificity under a variety of reasonable assumptions about the diagnostic accuracy of lymph node evaluation and imaging. The empirical estimate that about 10% of all colorectal cancer patients jointly test falsely negative for lymph node evaluation and imaging suggests that each of these two tests individually has an average sensitivity of about 65%. This is because the probability of testing falsely negative for lymph node evaluation and also testing falsely negative for imaging is, as independent events, the product of the false negative proportions of each test.

To show this, assume that lymph node evaluation and imaging have the same false negative proportion X . It follows that $(X)(X) = 0.14$ (i.e. 14%, the proportion of colorectal cancer patients we estimate to truly have metastatic disease at diagnosis that is not detected by either test), and therefore $X = 0.37$. Since the sensitivity of a single test is equal to $1 - (\text{false negative proportion})$, for each test $\text{sensitivity} = 1 - 0.37 = 0.63 = 63\%$ (58). We used this conclusion to choose our values of 60% and 80% for fixed lymph node evaluation and imaging sensitivity parameters. These choices set up reasonable bounds to assess the impact on estimates of EMT sensitivity and specificity of assuming high or low sensitivity of lymph node evaluation and imaging.

A second strength is that latent class analysis provided valid estimates of EMT sensitivity and specificity because it did not require the assumption that either lymph node evaluation or imaging constitutes a gold standard measure of cancer cell detachment from primary tumors. Standard calculations of sensitivity and specificity are based on a cross-tabulation between a new test and a gold standard. The only tests that EMT can be compared to in this fashion are lymph node evaluation and imaging. The calculation in the preceding paragraph supports the conclusion that neither of the conventional tests of cancer cell detachment is perfectly accurate or nearly so. Therefore, standard calculations of sensitivity and specificity would not be valid. These quantities must be estimated using a method that does not assume that any of the tests is a gold standard. Latent class analysis is precisely suited to this task (61).

Third, testing the impact of several different marker expression cut points to define EMT status allowed us to assess whether the estimates from our latent class models reflected how one would expect sensitivity and specificity to change as marker cut point changes. At a cut point of 0.52, few subjects were classified as EMT-positive, and one would expect the test to have high

specificity and low sensitivity. As the cut point was raised to 0.60 and then 0.85, the proportion of subjects classified as EMT-positive increased, and one would expect test specificity to decrease and sensitivity to increase accordingly. Our models across different cut points matched this pattern exactly (Tables 9-11), supporting the validity of latent class analysis as an estimation technique for EMT sensitivity and specificity.

Our analysis had several limitations. First, our sample size was small and there was a slight overrepresentation of subjects diagnosed with local disease compared to the national average. Second, our dataset lacked information on lymph node evaluation and imaging test results, leading us to infer them based on tumor stage diagnosis. Although this was probably a good approximation, we cannot verify this and it is likely that our inferred test results do not exactly match what was observed clinically. A third limitation was the inherently constrained setting of latent class estimation with three binary indicators. Free estimation of all model parameters led to model saturation (i.e. 0 df), and our saturated models performed about as well as our worst-performing models with lymph node evaluation and imaging parameter restrictions. Conversely, models restricting all 4 lymph node evaluation and imaging item-response parameters essentially make it impossible to estimate latent classes because only EMT can vary. Therefore, the best compromise was to fix some, but not all, of the lymph node evaluation and imaging parameters in each model.

Future research on this topic is warranted. Latent class analysis of EMT marker sensitivity and specificity should be carried out in larger datasets with information available on directly observed lymph node evaluation and imaging results. In addition, Bayesian approaches to latent class estimation may provide more model flexibility and allow for deeper exploration of the latent class structure than was possible with our relatively simple models (78-80). For

example, an important assumption in latent class analysis is the local independence assumption: within a given latent class, that the indicators are statistically independent of each other (62). This is likely reasonable for tests of cancer cell detachment from primary tumors because lymph node evaluation, imaging, and EMT would be measured at different sites (lymph nodes, distant sites, and primary tumor, respectively) using different technologies and evaluated by different personnel (radiologist and pathologist). Bayesian analysis could allow for evaluation of this assumption. In addition, Bayesian approaches could provide more informative estimates of the “value added” by adding a third measure to the panel of tests of cancer cell detachment. However, an effective Bayesian analysis would probably require a much larger sample size than was available in our dataset.

Latent class analysis provides an important strategy to obtain valid estimates of EMT marker sensitivity and specificity in a setting without a true gold standard. Our results suggest that EMT markers could help to substantially reduce the number of cancer patients incorrectly diagnosed as having local disease.

Table 8. Subject characteristics (n=188)

<u>Characteristic</u>		
Age (years) (Mean, SD)	67	13
Sex	<u>N</u>	<u>%</u>
Male	89	47%
Female	99	53%
Race		
Non-Hispanic White	149	79%
Other	39	21%
Tumor Stage		
Local	99	53%
Regional	66	35%
Distant	23	12%
Lymph Node Diagnostic Status^a		
Positive	86	46%
Negative	102	54%
Radiologic Imaging Diagnostic Status^a		
Positive	23	12%
Negative	165	88%
EMT Diagnostic Status, Cut Point=0.52^b		
Positive	11	6%
Negative	177	94%
EMT Diagnostic Status, Cut Point=0.60^b		
Positive	28	15%
Negative	160	85%
EMT Diagnostic Status, Cut Point=0.85^b		
Positive	108	57%
Negative	80	43%

^aInferred from tumor stage using rules given in Table 2.

^bBased on E-cadherin expression measured as a weighted average of tumor cores on a continuous average intensity scale of 0-3. Low E-cadherin expression (below the cut point) is evidence of EMT (EMT-positive).

Table 9. Estimated sensitivity and specificity of EMT markers, and prevalence of high risk for metastatic disease, at EMT cut point of 0.52 and varying fixed values of sensitivity or specificity for other tests (n=188)^a

Fixed Lymph-Node Parameter	Fixed Imaging Parameter	Model DF	Estimates at EMT cut point 0.52		
			Sensitivity	Specificity	Prevalence of High Risk for Metastatic Disease ^b
Sensitivity 60%	Sensitivity 60%	2	0%	93%	21%
Sensitivity 80%	Sensitivity 60%	2	0%	93%	21%
Sensitivity 60%	Sensitivity 80%	2	0%	93%	16%
Sensitivity 80%	Sensitivity 80%	2	0%	93%	16%
Specificity 90%	Specificity 90%	2	6%	94%	41%
Specificity 100%	Specificity 90%	2	6%	94%	47%
Specificity 90%	Specificity 100%	2	5%	94%	47%
Specificity 100%	Specificity 100%	2	6%	94%	53%
None	None	0	0%	93%	20%

^aEMT measured by immunohistochemistry as a weighted average of cancer cell E-cadherin expression in primary tumor cores on a continuous average intensity scale of 0-3, then dichotomized into values below 0.52 (EMT-positive) and at or above 0.52 (EMT-negative). Low E-cadherin expression suggests occurrence of EMT in cancer cells.

^bPrevalence of low-risk group for metastatic disease is (100 – prevalence of high-risk group).

DF=degrees of freedom

Table 10. Estimated sensitivity and specificity of EMT markers, and prevalence of high risk for metastatic disease, at EMT cut point of 0.60 and varying fixed values of sensitivity or specificity for other tests (n=188)^a

Fixed Lymph-Node Parameter	Fixed Imaging Parameter	Model DF	Estimates at EMT cut point 0.60		
			Sensitivity	Specificity	Prevalence of High Risk for Metastatic Disease ^b
Sensitivity 60%	Sensitivity 60%	2	5%	82%	21%
Sensitivity 80%	Sensitivity 60%	2	5%	82%	21%
Sensitivity 60%	Sensitivity 80%	2	5%	83%	16%
Sensitivity 80%	Sensitivity 80%	2	5%	83%	16%
Specificity 90%	Specificity 90%	2	13%	84%	41%
Specificity 100%	Specificity 90%	2	14%	84%	47%
Specificity 90%	Specificity 100%	2	12%	83%	46%
Specificity 100%	Specificity 100%	2	13%	84%	53%
None	None	0	5%	83%	20%

^aEMT measured by immunohistochemistry as a weighted average of cancer cell E-cadherin expression in primary tumor cores on a continuous average intensity scale of 0-3, then dichotomized into values below 0.60 (EMT-positive) and at or above 0.60 (EMT-negative). Low E-cadherin expression suggests occurrence of EMT in cancer cells.

^bPrevalence of low-risk group for metastatic disease is (100 – prevalence of high-risk group).

DF=degrees of freedom

Table 11. Estimated sensitivity and specificity of EMT markers, and prevalence of high risk for metastatic disease, at EMT cut point of 0.85 and varying fixed values of sensitivity or specificity for other tests (n=188)^a

Fixed Lymph-Node Parameter	Fixed Imaging Parameter	Model DF	<u>Estimates at EMT cut point 0.85</u>		
			Sensitivity	Specificity	Prevalence of High Risk for Metastatic Disease ^b
Sensitivity 60%	Sensitivity 60%	2	49%	40%	21%
Sensitivity 80%	Sensitivity 60%	2	50%	41%	21%
Sensitivity 60%	Sensitivity 80%	2	48%	41%	16%
Sensitivity 80%	Sensitivity 80%	2	49%	41%	16%
Specificity 90%	Specificity 90%	2	59%	44%	41%
Specificity 100%	Specificity 90%	2	59%	44%	47%
Specificity 90%	Specificity 100%	2	58%	43%	47%
Specificity 100%	Specificity 100%	2	58%	44%	53%
None	None	0	51%	41%	21%

^aEMT measured by immunohistochemistry as a weighted average of cancer cell E-cadherin expression in primary tumor cores on a continuous average intensity scale of 0-3, then dichotomized into values below 0.85 (EMT-positive) and at or above 0.85 (EMT-negative). Low E-cadherin expression suggests occurrence of EMT in cancer cells.

^bPrevalence of low-risk group for metastatic disease is (100 – prevalence of high-risk group).

DF=degrees of freedom

CHAPTER 6. DISCUSSION

6.1. Summary of findings

In this study I explored the potential for EMT markers to improve CRC outcomes by serving as a test of cancer cell detachment from primary tumors in addition to lymph node evaluation and radiologic imaging. This exploration involved three steps. First, I reviewed the literature to identify the most promising EMT markers. Second, I measured the selected markers in a set of CRC primary tumors from a population-based prospective cohort study and used time-to-event analysis to estimate associations between marker expression and time from surgery to death. Third, using the one marker that was associated with mortality in the second step, I used LCA to estimate the diagnostic accuracy of EMT while accounting for lymph node evaluation and imaging, but without assuming that any of the tests was a gold standard.

6.1.1. Associations between EMT marker expression and time-to-death

I decided to study the most promising marker from each of the three categories of EMT markers: epithelial markers, mesenchymal markers, and inducers. Based on the literature, I selected E-cadherin, Integrin beta-6, and Snail, respectively. The expression of each marker was measured on a continuous scale in CRC primary tumors from a population-based prospective cohort study. For E-cadherin, this was average intensity ranging from 0-3. For Integrin and Snail, I used the scale of percent positive cells or nuclei, respectively (range: 0-100). In addition, for a given marker the expression score for each subject was assigned in two different ways: as a weighted average of tumor cores and as the worst tumor core, where “worst” meant lowest expression for E-cadherin and highest expression for Snail and Integrin. For each of the

resulting six continuous marker expression variables, I examined every possible dichotomization of the data to identify the cut point with the best model fit in a bivariate proportional hazards model between dichotomous marker expression and time-to-death.

When marker expression was dichotomized at the continuum-specific statistically-optimal cut point, I found for E-cadherin weighted average that, compared to high expression, low expression was associated with a greater hazard of dying within 5 years of surgery. Optimally-dichotomous expression was not associated with outcomes for any of the other marker expression variables, including E-cadherin worst core. Adjusting for age, tumor stage, and neoadjuvant treatments did not change any of these findings.

To further explore how E-cadherin weighted average data might be implemented clinically, I compared several different cut points along this continuum that could each predict outcomes: the statistically-optimal cut point of about 0.52 as well as 0.60 and 0.85. I found that, in moving from 0.52 to 0.85, the strength of the proportional hazards association between dichotomous marker expression and time-to-death attenuated but precision increased. The same pattern appeared on the risk-difference and risk-ratio scales.

Cross-tabulating dichotomous E-cadherin weighted average expression status and tumor stage further complicated the picture. The statistically-optimal cut point would lead to changes in treatment for only 6 of 99 subjects diagnosed with local disease and having E-cadherin measurements available. The corresponding numbers for 0.60 and 0.85 were 16 and 56, respectively. These results should be compared to the clinical observation that about 25% of CRC patients diagnosed with local disease later experience recurrence (4). This suggests that about 25 of the 99 subjects diagnosed with local disease and having E-cadherin measurements would ideally be classified as E-cadherin-negative and therefore as EMT-positive.

6.1.2. Latent class estimates of EMT sensitivity and specificity to evaluate cancer cell detachment from primary tumors

I used LCA to estimate the diagnostic accuracy of EMT markers to evaluate cancer cell detachment from primary tumors. Based on the finding from time-to-event analysis that only E-cadherin weighted average was associated with outcomes, in the LCA I classified subjects as EMT-positive or EMT-negative based on their E-cadherin weighted average measurements. I examined results when varying each of several different conditions: assumptions about the sensitivity or specificity of lymph node evaluation and radiologic imaging, as well as the E-cadherin weighted average cut point used to classify subjects as EMT-positive or EMT-negative.

The LCA results reproduced the general patterns one would expect to see when varying the cut point of a continuous variable. When the cut point was at a low value of 0.52, such that few subjects were classified as EMT-positive, the specificity of the marker was high (over 90%) but the sensitivity was low (under 10%). As the cut point was progressively raised and the proportion of subjects classified as EMT-positive increased accordingly, the specificity decreased and sensitivity increased. The fact that the LCA models produced results that fit the expected pattern of changes in sensitivity and specificity as cut point changed supported the validity of LCA as an estimation procedure in this setting.

For metastatic disease diagnosis, false negatives have more serious clinical consequences than false positives. A false negative in this setting is a case that truly has metastatic disease but receives a negative result for every diagnostic test. In other words, the patient would need chemotherapy but likely would not receive it. A false positive would occur when the patient truly does not have metastatic disease but receives a positive result for at least one test. This patient would not benefit much from chemotherapy but would probably receive it. The consequences of a false positive would generally be that the patient would unnecessarily suffer

the serious side effects of chemotherapy. In contrast, the consequences of a false negative could easily be fatal.

I estimated that there are about 14,000 cases each year in the United States of colorectal cancer patients with metastatic disease at the time of diagnosis who nonetheless are diagnosed with local disease (4, 77). In addition, I estimated that there are about 42,000 colorectal cancer patients in the United States each year who are diagnosed with local disease and truly do not have metastatic disease (4, 63, 77).

The most plausible set of assumptions that I explored using LCA assumed was that, between lymph node evaluation and imaging, one of them has a sensitivity of 80% and the other a sensitivity of 60%. Under these conditions, the best-performing EMT cut point that I examined—E-cadherin weighted average cut point of 0.85 on a continuous average intensity scale of 0-3—had an EMT sensitivity of about 50% and specificity of about 40%. This implies an EMT false negative proportion of 50% and false positive proportion of 60%.

Combining these EMT diagnostic accuracy estimates with the population estimates for true and false diagnoses of local disease allows us to see the overall clinical trade-off that EMT markers offer per the results of my analyses. In short, implementing EMT as a test of metastatic disease at an E-cadherin weighted average cut point of 0.85 could eliminate 50% of the false diagnoses of local disease each year (about 7,000 cases). This gain would come at the expense of over-treating 60% of those who are currently accurately diagnosed with local disease (about 25,200 cases).

The best-performing EMT estimates that I found involved a cut point of 0.85 and assumed a lymph node specificity of 90% and imaging specificity of 90%. Under these conditions, EMT had a sensitivity of 59% and specificity of 44%, implying a false negative

proportion of 41% and false positive proportion of 56%. These estimates translated to elimination of about 8,300 false diagnoses of local disease per year, but at the cost of over-treating about 23,500 cases that are currently correctly diagnosed with local disease.

In sum, under the best estimates of sensitivity and specificity that I obtained, EMT markers presented a clinical trade-off between eliminating false negatives and creating new false positives that would probably not be considered a net improvement. Under the most plausible assumptions, even the best-performing EMT cut point produced sensitivity and specificity estimates that would not represent a net improvement.

6.2. Strengths and limitations

The project had a number of important strengths. First, it used a population-based sample, meaning it was based on a well-defined source population and therefore should be regarded as having greater external validity than would a hospital-based sample (75). Although we were only able to measure EMT markers in primary tumor specimens for roughly 20% of NC CanCORS subjects, the EMT study sample seemed to closely resemble overall NC CanCORS, preserving the interpretation of the EMT study sample as population-based.

A second strength is that we measured continuous marker expression data using automated procedures. Continuous data enabled a more detailed examination of clinically-informative cut points and exploration of how marker expression varied with time-to-death than would have been possible with ordinal data, such as that typically produced by manual scoring. In addition, we assigned continuous expression scores in two ways: as a weighted average of cores by tissue type (normal or tumor) and as the “worst” core by tissue type. This allowed us to assess the impact of marker expression heterogeneity throughout a tumor. As will be discussed

later, this assessment of heterogeneity was severely limited by the lack of consistent, informative sampling of primary tumors.

Third, we assessed the reliability of our automated annotations of tissue specimens by comparing them to gold-standard manual annotations. We found correlations of about 90% between the automated scores produced by automated annotation and the automated scores produced by manual annotation. This quality-control step supported the accuracy of our marker expression measurements.

Fourth, we developed an automated procedure to examine, for each marker expression continuum, every possible cut point and its relationship with time to patient death. In addition to exhaustiveness, we also identified an objective criterion that could be used to find a clinically-informative cut point based on best model fit. This approach allowed us to fully exploit the richness of the continuous data that we measured. It also made it possible to consider the impact of choosing clinical cut points that differed from the statistically-optimal cut point. For example, we found for E-cadherin weighted average that the statistically-optimal cut point would lead to a number of patients whose treatments would be reassigned that was probably too low, given the proportion of subjects diagnosed with local disease who later experience recurrence. However, by exploring multiple cut points, we identified a cut point with a point estimate only a little weaker than the one at the statistically-optimal cut point, but with better precision and a more appropriate number of patients whose treatments would be reassigned. In short, while the statistically-optimal cut point might not be the clinically-optimal cut point, our procedures towards data collection and analysis provided maximum flexibility to evaluate such trade-offs.

Finally, I estimated EMT marker sensitivity and specificity using LCA to avoid the assumption that any diagnostic test of metastatic disease constituted a gold standard, which was

likely not a valid assumption. Diagnostic accuracy as expressed by sensitivity and specificity is an important measure of the clinical benefit of any biomarker. In the setting of metastatic disease diagnosis, no gold standard exists to which EMT markers can be compared using standard calculations. A true gold standard would have both a sensitivity and specificity of 100%. While lymph node evaluation and imaging could well have specificities close to 100%, I estimated that the average sensitivity for each of them is only around 65%. Using LCA allowed me to obtain valid accuracy estimates for EMT while still taking into account the accuracy of lymph node evaluation and imaging.

The study had several limitations. First was the small sample size. This led to poor precision even among the statistically-significant estimates. The low power of the full sample size also prevented me from dividing the dataset into training and validation sets. It also impeded a richer consideration of different E-cadherin weighted average cut points to better explore which one might be best for clinical purposes. Aside from examining a cut point of about 0.52 because it had the best model fit, I examined the alternative cut points of 0.60 and 0.85 because they either remained statistically significant (0.60) or were very nearly so (0.85). Most cut points in the range of 0.50 to 1.05 had strong hazard ratio point estimates between 1.60 and 2.60. However, the lower 95% confidence limit would often dip to around 0.85 or 0.90. Though this was almost certainly due to the small sample size rather than a lack of effect, such imprecision made it difficult to consider cut points such as 0.70 or 0.90 more deeply.

Second, the goat antibodies that we used for Snail and Integrin beta-6 could have been rendered ineffective by nonspecific staining. My analyses suggested that Snail and Integrin expression levels simply had no effect on patient time-to-death. However, this could have been due to poor performance of the particular antibodies that were used rather than a reflection of a

genuine lack of association between these markers and patient outcomes. We were not able to examine this possibility but it is a potentially serious flaw in these marker measurements.

Third, some amount of selection bias was potentially present in the sample because not all eligible CRC cases enrolled in NC CanCORS. During the enrollment period, there were 1,899 nominally-eligible and 1,506 fully-eligible incident CRC cases in the study catchment area (Figure 3). Of these, 990 patients enrolled in the study. Thus, NC CanCORS included 66% of fully-eligible cases and 52% of nominally-eligible cases. Any systematic differences between those who did not enroll and those who did would bias our results away from the true estimates for the source population. For example, 85 eligible cases could not give consent to enroll and had no proxy to consent for them. These cases were probably sicker at the time of diagnosis than those who enrolled. Their inability to consent means we have no study data for them, including tumor specimens. Exclusion of these cases meant that NC CanCORS, and therefore the EMT study sample, differ at least somewhat from the targeted source population.

Fourth, the dataset had a large amount of missing data. This problem arose in two different ways: as substantial proportions of missing data on variables that were collected, and also lack of variables that were relevant to the analysis. For the first issue, variables for neoadjuvant treatments—which were part of our multivariate Cox modeling adjustment set—had missing or non-informative values for 50 of 190 subjects (26%) in the EMT study sample. The time-to-event outcome variable of time from surgery until all-cause mortality was missing for 27 of 190 subjects (14%). Missing values for these variables had to be imputed using multiple imputation before running proportional hazards models. Other variables relevant to characterizing cancer patients also had substantial proportions of missing or non-informative data, such as tumor grade (16%) and tumor location (15%). Although multiple imputation is as

robust a method for handling missing data as is available, and though our effect estimates were stable in sensitivity analyses comparing complete cases to all cases with imputed values for missing data, imputation is still inferior to complete information on all subjects.

The problem of crucial variables not being available at all was manifest in the latent class analysis. Observed test results were needed for all three tests: lymph node evaluation, radiologic imaging, and EMT. Results for the first two were not part of the CanCORS medical records data collection. Therefore, results for these two variables were inferred from tumor stage, which likely did not exactly match what was observed clinically.

Fifth, in the latent class analysis, tests of metastatic disease presented a highly constrained setting that complicated estimation of meaningful results. The reason is that the flexibility of latent class models varies directly with the number of possible response patterns. In a setting with three binary tests, there are only 8 possible response patterns. Since free estimation of all parameter yields a poorly-performing model with 0 df, better model specification and performance were achieved when fixing the values of two parameters (of 7 possible) per model. In some models with restrictions these parameters were the sensitivities of lymph node evaluation and imaging; in others they were the specificities of these two tests. While reasonable fixed values were assigned, clearly our results are limited to the sets of assumptions implemented in the models that were run. Had we used different combinations of fixed values, we would have observed different results.

The sixth and final limitation of the study was probably the most important. EMT marker expression likely is not homogeneous throughout the cancer cells in a primary tumor. Rather, certain portions of the tumor are probably much more likely to have cancer cells likely to undergo EMT compared to other parts of the tumor, with the invasive front thought to be the

chief source of transitioning cells (81). Ideally, every primary tumor in the study would have been sampled in an informative, consistent manner, such as systematically sampling the invasive front, tumor center, and an edge of the tumor away from the invasive front (5). This would have enabled us to estimate, say, not just the association between E-cadherin weighted average expression and outcomes, but between invasive front E-cadherin weighted average expression and outcomes. Separately, we could also have estimated the association between tumor center E-cadherin weighted average and outcomes. This would have allowed us to assess marker expression heterogeneity most directly.

However, informative tumor sampling was not carried out in CanCORS. This meant that we were looking at some unknown mixture of invasive front samples, tumor center samples, and samples from other parts of the tumors. The most probable impact of this problem would have been to attenuate our effect estimates compared to what they would be if we only had invasive front expression. I would expect marker expression from parts of the tumor other than the invasive front to have a weaker association with outcomes than what would be seen with invasive front expression. A further implication is that we could not evaluate tumor expression heterogeneity. This could explain why we saw no effect of E-cadherin worst core expression on outcomes but did see a relationship for E-cadherin weighted average expression. “Worst core” may not mean much if the tumors are not sampled in an informative way. The weighted average at least had the virtue of being more representative of marker expression for the overall tumor because it generally accounted for more cells than did the worst core.

6.3. Public health implications and future directions

The potential clinical and public health benefit of EMT markers is as an additional diagnostic test of metastatic disease to complement conventional tests. Specifically, the markers could identify as high risk a substantial number of patients who truly have metastatic disease at diagnosis but that test negative according to lymph node evaluation and imaging. This would enable these patients to receive more appropriate treatments than they would otherwise.

I found that E-cadherin measured as a weighted average of tumor cores was associated with time from surgery to patient death. However, the picture was complicated when I estimated the diagnostic accuracy of E-cadherin weighted average as marker sensitivity and specificity. I examined a variety of combinations of E-cadherin cut point and assumptions about the diagnostic accuracy of lymph node evaluation and imaging. I did not find E-cadherin accuracy estimates that simultaneously resulted from plausible assumptions and provided an acceptable trade-off between the number of false negatives that would be eliminated and the number of new false positives that would result.

These mixed results and the strengths and limitations of the study suggest important considerations for future work on EMT markers and patient outcomes. First, much larger sample sizes are needed. Specifically, sample sizes must be big enough that the sample can be divided into training and validation sets, each of which having much better precision than the sample used in the present study.

Second, informative tumor sampling is essential. I recommend that every primary tumor be sampled for at least two cores from each of the invasive front, tumor center, and an edge of the tumor away from the invasive front. Cores from different parts of the tumor should not be

averaged together. Different analyses should be run using marker expression data from each of the various tumor portions separately.

Third, future studies should collect data on all-cause mortality, cancer-specific mortality, and time to recurrence. This would enable comparison between different time-to-event analyses using each of the three outcomes to assess whether this makes a difference. In our study, we only had all-cause mortality available and therefore could not compare the results to what would have been obtained with cancer-specific mortality or time to recurrence as the outcome.

Fourth, future latent class estimates of EMT marker sensitivity and specificity should be based on direct observations of lymph node evaluation and imaging test results, not inferred from tumor stage or other information.

Finally, latent class analyses to estimate EMT marker diagnostic accuracy should be carried out within a Bayesian framework. This would allow for more informative estimates and greater flexibility in testing assumptions than was possible with our simple frequentist models. An effective Bayesian analysis would almost certainly require a substantially greater sample size than was available in the present work.

In this study we have developed numerous ideas about how best to measure EMT markers in primary tumors and analyze the data to provide clinically-informative estimates. In turn, these estimates can be used to decide whether and how to use the markers in clinic. Though we lacked the resources to implement all of our ideas, we hope that future studies will be able to do so and thereby realize our intention in undertaking this work to improve cancer outcomes.

APPENDIX A. SEARCH TERMS FOR SYSTEMATIC LITERATURE REVIEW OF EMT MARKERS AND OUTCOMES

Below is the search we ran in PubMed. Search terms were reformatted and run in EMBASE via Elsevier and BIOSIS via Thomson Reuters Web of Science.

(epithelial mesenchymal[tw] OR epithelial to mesenchymal[tw] OR emt[tw] OR epithelium mesenchyme[tw]) AND (tumor markers, biological[mesh] OR tumor marker*[tw] OR twist*[tw] OR e-cadherin[tw] OR ecadherin[tw] OR n-cadherin[tw] OR ncadherin[tw] OR vimentin*[tw] OR fibronectin*[tw] OR zeb1[tw] OR zeb 1[tw] OR zeb2[tw] OR zeb 2[tw] OR Cytokeratin*[tw] OR snail[tw] OR slug[tw] OR beta catenin[tw] OR integrin*[tw] OR microRNA*[tw] OR micro RNA*[tw] OR miRNA*[tw] OR TGF-beta*[tw] OR TGFbeta*[tw] OR TGFB*[tw] OR transforming growth factor beta*[tw]) AND (mortality[mesh] OR mortality[subheading] OR mortalit*[tw] OR prognosis[mesh] OR prognosis[tw] OR prognost*[tw] OR surviv*[tw] OR hazard[tw] OR rate[tw] OR risk[tw] OR ratio[tw] OR analys*[tw] OR analyz*[tw]) AND ("Colorectal Neoplasms"[Mesh] OR (tumor[tw] OR tumors[tw] OR tumour*[tw] OR cancer[tw] OR cancers[tw] OR carcinoma*[tw] OR adenoma*[tw] OR adenocarcinoma*[tw] OR polyp[tw] OR polyps[tw])) AND (colorectal[tw] OR colonic[tw] OR rectal[tw] OR colon[tw] OR rectum[tw] OR anal[tw] OR anus[tw] OR large bowel[tw]))

The search consisted of four groups of terms: EMT, tumor markers, outcomes, and colorectal cancer. To be included, an item had to contain at least one term from each of the groups. The tumor-marker group was designed to capture any marker using the generalized “tumor markers, biological” MeSH term and “tumor marker*” text-word term. Particular markers were specified as a “safety net” for markers prominent in the EMT literature but did not exclude other markers. The search can be modified from colorectal cancer to other kinds of cancer by replacing the colon- and rectum-specific terms in the last group with analogous terms for other tumor sites.

APPENDIX B. MARKER CUT POINT OPTIMIZATION MACRO: SAS CODE AND DETAILS

B.1. Introduction

At the end of this supplement we present the SAS code for the macro (“opt”) used to find the cut point for a continuous marker expression variable that yields the best model fit for a Cox regression of the association between the dichotomized marker expression variable and time from surgery to patient death.

For anyone who wishes to use the macro, we describe the features of the code that can be adapted to the investigator’s data and desire to control the output. For ease of reference, every fifth line of code has been numbered. Skipped lines and annotations do not count in the line numberings. The code was written using SAS version 9.3 (SAS Institute, Cary, NC).

B.2. Features of the code

The macro input is a dataset with one observation per subject and containing the following information: continuous marker expression variables, length of time from surgery to patient death, and any censoring variable to be used in Cox proportional hazards modeling. In our example, the input dataset is called “markercore7” and is read-in in line 3.

For any continuous marker expression variable, the macro orders the expression values of all subjects from least to greatest, identifying the observed range of values. For a given marker expression value, the program dichotomizes the variable at that value, thereby establishing distinct marker-positive and marker-negative groups as defined by the particular cut point.

The dichotomized marker expression variable is fit as the only independent variable in a Cox proportional hazards model of time from surgery to death (PROC PHREG in lines 41-45). In our example, TTE5 is the time-to-event variable, with administrative censoring at 5 years.

CENSOR5 is an indicator of whether the subject was censored at 5 years (0=not censored, 1=censored). The Cox model produces a fit statistic for the current iteration of the program.

The macro repeats this process for every continuous expression value in the observed data. The dichotomous marker expression variable is named “cut.” The macro output is a list of expression values and corresponding model fit statistics when marker expression is dichotomized at that particular expression value. The list is ordered from lowest fit statistic (best fit) to highest fit statistic (worst fit). See Section 3 below for an example.

Investigators using the macro may wish to tailor the output at several points. First, the investigator can control what kind of model fit statistics are produced by setting the value of “_n_” in line 49. Setting the value to 1 requests -2 log likelihood statistics, a value of 2 requests AIC statistics, and 3 requests BIC statistics. We chose to work with BIC statistics and so used a value of 3, as well as naming the fit statistic output variable “bic.”

Second, the macro incorporates a “switch” variable that can be set to 0 or 1. Setting switch=0 produces a complete list of all expression values and corresponding model fit statistics, ordered from lowest to highest fit statistics. Switch=1 restricts the output to the lowest model fit statistic and its corresponding expression value.

The final line of the code calls the macro for a particular continuous marker expression variable: “%opt(EAIWAV_T,0);” (line 85). The items in parentheses are particular values of the general form (continuous marker expression variable, switch variable). Our example expression variable is EAIWAV_T, which stands for “continuous E-cadherin expression measured on the average intensity scale, values assigned as weighted averages of cores, for tumor cores.” We request that the program show the full macro output for this variable and so have set switch=0.

B.3. Sample macro output

Below is the first 10 rows of output for marker expression variable EAIWAV_T with BIC model fit statistics. Since we had 190 subjects in our dataset, the actual list of output results has as many rows as the number of subjects, minus ties for continuous expression values. These first 10 results are the 10 lowest (best-fitting) BIC values for this expression continuum in our study. E-cadherin was measured on a continuous average intensity scale ranging from 0 to 3.

Result for EAIWAV_T	
cut	bic
0.51537	624.441
0.56128	624.976
0.50443	625.234
0.47189	625.410
0.52104	625.965
0.56590	626.022
0.47018	626.048
0.58048	626.208
0.82207	626.332
1.55779	626.346

The top result sets the cut point at an expression value of 0.51537 (what we refer to in the main text as “about 0.52”) and has the best fit of all bivariate associations between dichotomous marker expression status and time-to-death in our observed data. We thus refer to 0.51537 as the statistically-optimal cut point for E-cadherin weighted average. A dichotomous positive/negative status variable was created for E-cadherin weighted average using this cut point. We used this dichotomous variable in the E-cadherin weighted average Cox models presented in Tables 5 and 6.

B.4. Comparison of optimization macro to receiver operating characteristic (ROC) curves

Our approach is an alternative to ROC curves. Both methods can be used to select a cut point along a continuum of marker expression values based on a criterion that relates marker expression to patient outcomes. The methods differ in two important respects.

First, the form of subject outcomes is different. ROC curves use a binary outcome of whether a subject died (yes/no) while our approach uses continuous time-to-death. It matters whether a patient died 5 months or 50 months after surgery. ROC curves do not account for such distinctions whereas our approach does.

Second, the criterion used to identify a cut point differs between the methods. In ROC curves, the cut point selected is typically the one corresponding to the most upper-left-hand point on a plot of sensitivity versus $(1 - \text{specificity})$ (i.e. true positive rate versus false positive rate). In our approach, the statistically-optimal cut point is the one yielding the best model fit in a bivariate proportional hazards model of dichotomous marker expression and continuous time-to-death.

This difference between cut point selection criteria implies a difference in interpretation between the two methods. Being based on measures of sensitivity and specificity, the cut point selected by an ROC curve is usually interpreted as the one that should be implemented clinically. However, selecting the upper-left-hand corner of the ROC curve implies that false positives and false negatives have clinical consequences of roughly equal importance, which is rarely true. In contrast, the optimal cut point in our approach is a statistical measure of the largest difference in the observed data between hazard functions for marker-positive and marker-negative subjects. Thus, the direct application of our method is to determine whether an association exists between marker expression and time-to-death. The statistically-optimal cut point might or might not be

judged to be best for clinical use, but that determination requires consideration of additional information besides model fit alone.

B.5. SAS code for marker cut point optimization macro

```
%macro opt(var,switch);
/* no missing */
  data _internal_;
    set markercore7;
    if &var^=.;
    run;
/* values of &VAR ordered */
  proc freq data=_internal_ noprint;
    tables &var/out=_table_(keep=&var);
  run;
/* number unique values of &VAR to &NN */
  data _null_;
    set _table_ end=end;
    if end then do;
      nn=put(_n_,8.);
      call symput("nn",nn);
    end;
  run;
/* initialize the output data set to _NULL_ */
  data _rsq_;
    set _null_;
  run;
  %do i=1 %to &nn;
/* get the present cut point */
    data _cut_;
      set _table_;
      rename &var=cut;
      if _n_=&i then do;
        xx=put(&var,25.10);
        call symput("cut",xx);
        output;
      end;
    run;
/* merge and assign to groups */
    data _use_;
      if _n_=1 then set _cut_;
      set _internal_;
      group=(&var<=cut);
      x1=group*&var;
      x0=(1-group)*&var;
    run;
    proc datasets;
      delete _cut_;
    run;
    quit;
/* model */
    ods listing close;
    proc phreg data=_use_;
      model TTE5*CENSOR5(1)=x1;
```



```

        ods output fitstatistics=_fit_(keep=withcovariates);
run;
quit;                                /* line 45 */
ods listing;
data _fit_;
set _fit_;
if _n_=3;
rename withcovariates=rsquare;        /* line 50 */
run;
data _rsq_;
set _rsq_ _fit_(in=in);
keep cut bic;
if in then do;                        /* line 55 */
    cut=&cut;
    bic=rsquare;
end;
run;
proc datasets;                        /* line 60 */
    delete _use_ _fit_;
run;
quit;
%end;
proc datasets;                        /* line 65 */
    delete _table_ _internal_;
run;
quit;
proc sort data=_rsq_ out=_rsq_;
    by bic;                            /* line 70 */
run;
title "Result for %upcase(&var)";
proc print noobs data=_rsq_
    %if &switch=1 %then %do;
    (obs=1)                            /* line 75 */
%end;
;
run;
title;
proc datasets;                        /* line 80 */
    delete _rsq_;
run;
quit;
%mend;

/* call macro for each continuous expression variable */

%opt(EAIWAV_T,0);                      /* line 85 */

```

APPENDIX C. LATENT CLASS ANALYSIS: SAS CODE AND DETAILS

Below is sample code for one of the latent class models that I ran. Running a model with parameter restrictions using PROC LCA in SAS 9.3 required three steps.

First, I created a dataset with starting values for each parameter (`lca_start_85_2`). Since I modeled 2-class solutions for three binary indicators—EMT, lymph node evaluation, and radiologic imaging—each model had seven parameters: 1 gamma (class-membership probability) and 6 rho (item-response probabilities, specifically sensitivity and specificity for each of the three indicators). In the example, I used EMT status defined as E-cadherin weighted average expression dichotomized at a cut point of 0.85 on a continuous average intensity scale of 0-3 (EMT85). The gamma parameter was set to “random” start values of 0.5 for each of the two classes. Class 1 was the “healthy” class, that is, those who tested negative for a given indicator and corresponded to those at low risk for metastatic disease. Therefore, it was the group in whom we obtained estimates for specificity. Class 2 was the “sick” class, that is, those who tested positive for a given indicator and corresponded to those at high risk for metastatic disease. Sensitivity was estimated among Class 2.

The second step was to create a dataset that designated which parameters would be restricted (i.e. given fixed values and not estimated) and which would be freely estimated (`lca_restr_85_2`). In the example, the fixed parameters were lymph node (LN3) sensitivity of 0.8 and radiologic imaging (RI) sensitivity of 0.6. The rest were freely estimated. Fixed parameters were designated with a 0 and freely estimated parameters with a 1. However, PROC LCA required that restricted parameters be designated with a 0 in only one of the two classes. Therefore, each fixed parameter is given a value of 0 for one of the two classes and a value of 1 for the other class.

The third step is to run the analysis using PROC LCA and incorporating the start-value and restriction datasets set up in the first two steps. The RHO PRIOR statement is used to ensure smooth model estimation in the event that any of the estimates would have values of 0 or 1, making it difficult for the algorithm to compute an overall solution and appropriate standard errors.

```
/* set up parameter start values */

DATA lca_start_85_2;
INPUT param $ group variable $ respcat estlc1 estlc2;
DATALINES;
  GAMMA      1      .      .      0.5      0.5
  BETA       1      .      .      0.0      0.0
  RHO        1      LN3    1      0.6      0.2
  RHO        1      RI     1      0.6      0.4
  RHO        1      EMT85  1      0.7      0.3
  RHO        2      LN3    2      0.4      0.8
  RHO        2      RI     2      0.4      0.6
  RHO        2      EMT85  2      0.3      0.7
;
RUN;

/* assign parameters as restricted or freely estimated */

DATA lca_restr_85_2;
INPUT param $ group variable $ respcat estlc1 estlc2;
DATALINES;
  GAMMA      1      .      .      1      1
  BETA       1      .      .      1      1
  RHO        1      LN3    1      1      0
  RHO        1      RI     1      1      0
  RHO        1      EMT85  1      1      1
  RHO        2      LN3    2      1      1
  RHO        2      RI     2      1      1
  RHO        2      EMT85  2      1      1
;
RUN;

/* run latent class model using designated start values and restrictions */

PROC LCA data=lca OUTEST=lca_out_85_2 START=lca_start_85_2
  RESTRICT=lca_restr_85_2;
  TITLE1 "EMT85 Model 2: LN fixed Se 0.80, RI fixed Se 0.60";
  NCLASS 2;
  ITEMS LN3 RI EMT85;
  CATEGORIES 2 2 2;
  MAXITER 5000;
  CRITERION 0.000001;
  RHO PRIOR=1;
RUN;
```

REFERENCES

1. Weinberg RA. The Biology of Cancer. New York: Garland Science; 2007.
2. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. The Journal of clinical investigation. 2009;119:1420-8.
3. Kang H, O'Connell JB, Leonardi MJ, Maggard MA, McGory ML, Ko CY. Rare tumors of the colon and rectum: a national review. International journal of colorectal disease. 2007;22:183-9.
4. Young PE, Womeldorph CM, Johnson EK, Maykel JA, Brucher B, Stojadinovic A, et al. Early detection of colorectal cancer recurrence in patients undergoing surgery with curative intent: current status and challenges. Journal of Cancer. 2014;5:262-71.
5. Busch EL, McGraw KA, Sandler RS. The potential for markers of epithelial-mesenchymal transition to improve colorectal cancer outcomes: a systematic review. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2014;23:1164-75.
6. NCI. Common Cancer Types. 29 January 2014]; Available from: <http://www.cancer.gov/cancertopics/types/commoncancers>
7. Siegel R, DeSantis C, Virgo K, Stein K, Mariotto A, Smith T, et al. Cancer treatment and survivorship statistics, 2012. CA: a cancer journal for clinicians. 2012;62:220-41.
8. SEER. Fast Stats. 29 January 2014]; Available from: <http://seer.cancer.gov/faststats/>
9. Edwards BK, Ward E, Kohler BA, Ehemann C, Zauberg AG, Anderson RN, et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer. 2010;116:544-73.
10. NCI. What You Need To Know About Cancer of the Colon and Rectum. 29 January 2014]; Available from: <http://www.cancer.gov/cancertopics/wyntk/colon-and-rectal>
11. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57-70.
12. Patz EF, Jr., Goodman PC, Bepler G. Screening for lung cancer. The New England journal of medicine. 2000;343:1627-33.
13. Ueno H, Kajiwara Y, Shimazaki H, Shinto E, Hashiguchi Y, Nakanishi K, et al. New criteria for histologic grading of colorectal cancer. The American journal of surgical pathology. 2012;36:193-201.

14. Zeisberg M, Neilson EG. Biomarkers for epithelial-mesenchymal transitions. *The Journal of clinical investigation*. 2009;119:1429-37.
15. Benson AB, 3rd, Schrag D, Somerfield MR, Cohen AM, Figueredo AT, Flynn PJ, et al. American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2004;22:3408-19.
16. Bhangu A, Wood G, Mirnezami A, Darzi A, Tekkis P, Goldin R. Epithelial mesenchymal transition in colorectal cancer: Seminal role in promoting disease progression and resistance to neoadjuvant therapy. *Surgical oncology*. 2012;21:316-23.
17. Hashimoto Y, Skacel M, Adams JC. Association of loss of epithelial syndecan-1 with stage and local metastasis of colorectal adenocarcinomas: An immunohistochemical study of clinically annotated tumors. *BMC Cancer*. 2008;8.
18. Lv Q, Wang W, Xue J, Hua F, Mu R, Lin H, et al. DEDD interacts with PI3KC3 to activate autophagy and attenuate epithelial-mesenchymal transition in human breast cancer. *Cancer Research*. 2012;72:3238-50.
19. Matsushita K, Toiyama Y, Tanaka K, Saigusa S, Hiro J, Uchida K, et al. Soluble CXCL16 in preoperative serum is a novel prognostic marker and predicts recurrence of liver metastases in colorectal cancer patients. *Annals of Surgical Oncology*. 2012;19:S518-S27.
20. Meng HM, Zheng P, Wang XY, Liu C, Sui HM, Wu SJ, et al. Overexpression of nanog predicts tumor progression and poor prognosis in colorectal cancer. *Cancer Biology and Therapy*. 2010;9:295-302.
21. Zhou WJ, Geng ZH, Chi S, Zhang W, Niu XF, Lan SJ, et al. Slit-Robo signaling induces malignant transformation through Hakai-mediated E-cadherin degradation during colorectal epithelial cell carcinogenesis. *Cell research*. 2011;21:609-26.
22. Zhu W, Cai MY, Tong ZT, Dong SS, Mai SJ, Liao YJ, et al. Overexpression of EIF5A2 promotes colorectal carcinoma cell aggressiveness by upregulating MTA1 through C-myc to induce epithelial-mesenchymal transition. *Gut*. 2012;61:562-75.
23. Bates RC, Bellovin DI, Brown C, Maynard E, Wu B, Kawakatsu H, et al. Transcriptional activation of integrin (beta)6 during the epithelial-mesenchymal transition defines a novel prognostic indicator of aggressive colon carcinoma. *Journal of Clinical Investigation*. 2005;115:339-47.
24. Bellovin DI, Bates RC, Muzikansky A, Rimm DL, Mercurio AM. Altered localization of p120 catenin during epithelial to mesenchymal transition of colon carcinoma is prognostic for aggressive disease. *Cancer Research*. 2005;65:10938-45.

25. Bellovin DI, Simpson KJ, Danilov T, Maynard E, Rimm DL, Oettgen P, et al. Reciprocal regulation of RhoA and RhoC characterizes the EMT and identifies RhoC as a prognostic marker of colon carcinoma. *Oncogene*. 2006;25:6959-67.
26. Fan XJ, Wan XB, Huang Y, Cai HM, Fu XH, Yang ZL, et al. Epithelial-mesenchymal transition biomarkers and support vector machine guided model in preoperatively predicting regional lymph node metastasis for rectal cancer. *British Journal of Cancer*. 2012;106:1735-41.
27. Fan XJ, Wan XB, Yang ZL, Fu XH, Huang Y, Chen DK, et al. Snail promotes lymph node metastasis and Twist enhances tumor deposit formation through epithelial-mesenchymal transition in colorectal cancer. *Human Pathology*. 2012.
28. Franci C, Gallen M, Alameda F, Baro T, Iglesias M, Virtanen I, et al. Snail1 protein in the stroma as a new putative prognosis marker for colon tumours. *PLoS ONE*. 2009;4.
29. Fujikawa H, Tanaka K, Toiyama Y, Saigusa S, Inoue Y, Uchida K, et al. High TrkB expression levels are associated with poor prognosis and EMT induction in colorectal cancer cells. *Journal of Gastroenterology*. 2012;47:775-84.
30. Gomez I, Pena C, Herrera M, Munoz C, Larriba MJ, Garcia V, et al. TWIST1 is expressed in colorectal carcinomas and predicts patient survival. *PLoS ONE*. 2011;6.
31. Harbaum L, Pollheimer MJ, Kornprat P, Lindtner RA, Rehak P, Hoefler G, et al. Keratin 7 expression in colorectal cancer - Freak of nature or significant finding? *Annals of Oncology*. 2010;21:i38.
32. Kahlert C, Klupp F, Brand K, Lasitschka F, Diederichs S, Kirchberg J, et al. Invasion front-specific expression and prognostic significance of microrna in colorectal liver metastases. *Cancer Science*. 2011;102:1799-807.
33. Kahlert C, Lahes S, Radhakrishnan P, Dutta S, Mogler C, Herpel E, et al. Overexpression of ZEB2 at the invasion front of colorectal cancer is an independent prognostic marker and regulates tumor invasion in Vitro. *Clinical Cancer Research*. 2011;17:7654-63.
34. Kevans D, Wang LM, Sheahan K, Hyland J, O'Donoghue D, Mulcahy H, et al. Epithelial-mesenchymal transition (EMT) protein expression in a cohort of stage II colorectal cancer patients with characterized tumor budding and mismatch repair protein status. *International Journal of Surgical Pathology*. 2011;19:751-60.
35. Knosel T, Chen Y, Hotovy S, Settmacher U, Altendorf-Hofmann A, Petersen I. Loss of desmocollin 1-3 and homeobox genes PITX1 and CDX2 are associated with tumor progression and survival in colorectal carcinoma. *International journal of colorectal disease*. 2012;1-9.
36. Knosel T, Emde V, Schluns K, Schlag PM, Dietel M, Petersen I. Cytokeratin profiles identify diagnostic signatures in colorectal cancer using multiplex analysis of tissue microarrays. *Cellular Oncology*. 2006;28:167-75.

37. Kouso H, Yano T, Maruyama R, Shikada Y, Okamoto T, Haro A, et al. Differences in the expression of epithelial-mesenchymal transition related molecules between primary tumors and pulmonary metastatic tumors in colorectal cancer. *Surgery Today*. 2012;1-8.
38. Liu Y, Zhang N, Wang Y, Xu M, Liu N, Pang X, et al. Zinc finger E-box binding homeobox 1 promotes invasion and bone metastasis of small cell lung cancer in vitro and in vivo. *Cancer Science*. 2012;103:1420-8.
39. Loboda A, Nebozhyn MV, Watters JW, Buser CA, Shaw PM, Huang PS, et al. EMT is the dominant program in human colon cancer. *BMC Medical Genomics*. 2011;4.
40. Mesker WE, Liefers GJ, Junggeburst JMC, Van Pelt GW, Alberici P, Kuppen PJK, et al. Presence of a high amount of stroma and downregulation of SMAD4 predict for worse survival for stage I-II colon cancer patients. *Cellular Oncology*. 2009;31:169-78.
41. Ono H, Imoto I, Kozaki K, Tsuda H, Matsui T, Kurasawa Y, et al. SIX1 promotes epithelial-mesenchymal transition in colorectal cancer through ZEB1 activation. *Oncogene*. 2012.
42. Oshiro R, Yamamoto H, Takahashi H, Ohtsuka M, Wu X, Nishimura J, et al. C4.4A is associated with tumor budding and epithelial-mesenchymal transition of colorectal cancer. *Cancer Science*. 2012;103:1155-64.
43. Shioiri M, Shida T, Koda K, Oda K, Seike K, Nishimura M, et al. Slug expression is an independent prognostic parameter for poor survival in colorectal carcinoma patients. *British Journal of Cancer*. 2006;94:1816-22.
44. Spaderna S, Schmalhofer O, Hlubek F, Berx G, Eger A, Merkel S, et al. A Transient, EMT-Linked Loss of Basement Membranes Indicates Metastasis and Poor Survival in Colorectal Cancer. *Gastroenterology*. 2006;131:830-40.
45. Toquet C, Colson A, Jarry A, Bezieau S, Volteau C, Boisseau P, et al. ADAM15 to (alpha)5(beta)1 integrin switch in colon carcinoma cells: A late event in cancer progression associated with tumor dedifferentiation and poor prognosis. *International Journal of Cancer*. 2012;130:278-87.
46. Zlobec I, Bihl MP, Foerster A, Ruffle A, Lugli A. The impact of CpG island methylator phenotype and microsatellite instability on tumour budding in colorectal cancer. *Histopathology*. 2012.
47. Qu P, Chu H, Ibrahim JG, Peacock J, Shen XJ, Tepper J, et al. Statistical strategies to improve the efficiency of molecular studies of colorectal cancer prognosis. *Br J Cancer*. 2008;99:2001-5.

48. Ayanian JZ, Chrischilles EA, Fletcher RH, Fouad MN, Harrington DP, Kahn KL, et al. Understanding cancer treatment and outcomes: the Cancer Care Outcomes Research and Surveillance Consortium. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2004;22:2992-6.
49. Ayanian J FM, Kahn K, Sandler R, Wallace R, Weeks J, Provenzale D, van Ryn M. Share thoughts on care: a population-based study of care and outcomes in lung and colorectal cancer patients (CanCORS).
50. Kang M, Shen XJ, Kim S, Araujo-Perez F, Galanko JA, Martin CF, et al. Somatic gene mutations in African Americans may predict worse outcomes in colorectal cancer. *Cancer biomarkers : section A of Disease markers*. 2013;13:359-66.
51. Kim S SR, Galanko J, Martin C, Sandler DP, Keku TO. Microsatellite instability and colorectal cancer survival in African Americans and whites.
52. CanCORS University of North Carolina at Chapel Hill Procedure Manual. UNC Center for Gastrointestinal Biology & Disease; 2002.
53. North Carolina Colorectal Cancer Study II (NCCCS II) TMA Protocol. Keku Laboratory.
54. Protocol for Blocks [1]. Keku Laboratory.
55. How to Request Blocks. Keku Laboratory.
56. He X, Chen Z, Jia M, Zhao X. Downregulated E-cadherin expression indicates worse prognosis in Asian patients with colorectal cancer: evidence from meta-analysis. *PLoS One*. 2013;8:e70858.
57. Allison PD. *Survival Analysis Using SAS: A Practical Guide*. Second ed. Cary, NC: SAS Press; 2010.
58. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Third ed. Philadelphia: Wolters Kluwer, Lippincott Williams & Wilkins; 2008.
59. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology (Cambridge, Mass)*. 1999;10:37-48.
60. Allison PD. *Missing Data*. Thousand Oaks, CA: Sage Publications; 2002.
61. Gaffikin L, McGrath JA, Arbyn M, Blumenthal PD. Visual inspection with acetic acid as a cervical cancer test: accuracy validated using latent class analysis. *BMC medical research methodology*. 2007;7:36.
62. Collins LM, Lanza ST. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, New Jersey: Wiley; 2010.

63. SEER. SEER Stat Fact Sheets: Colon and Rectum Cancer. 30 January 2015]; Available from: <http://seer.cancer.gov/statfacts/html/colorect.html>
64. Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS Procedure for Latent Class Analysis. *Structural equation modeling : a multidisciplinary journal*. 2007;14:671-94.
65. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646-74.
66. Thiery JP, Acloque H, Huang RY, Nieto MA. Epithelial-mesenchymal transitions in development and disease. *Cell*. 2009;139:871-90.
67. Malin JL, Ko C, Ayanian JZ, Harrington D, Nerenz DR, Kahn KL, et al. Understanding cancer patients' experience and outcomes: development and pilot study of the Cancer Care Outcomes Research and Surveillance patient survey. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*. 2006;14:837-48.
68. Yun JA, Kim SH, Hong HK, Yun SH, Kim HC, Chun HK, et al. Loss of E-Cadherin expression is associated with a poor prognosis in stage III colorectal cancer. *Oncology*. 2014;86:318-28.
69. Gao ZH, Lu C, Wang ZN, Song YX, Zhu JL, Gao P, et al. ILEI: a novel marker for epithelial-mesenchymal transition and poor prognosis in colorectal cancer. *Histopathology*. 2014.
70. Kim YH, Kim G, Kwon CI, Kim JW, Park PW, Hahm KB. TWIST1 and SNAIL as markers of poor prognosis in human colorectal cancer are associated with the expression of ALDH1 and TGF-beta1. *Oncology reports*. 2014;31:1380-8.
71. Kroepil F, Fluegen G, Vallbohmer D, Baldus SE, Dizdar L, Raffel AM, et al. Snail1 expression in colorectal cancer and its correlation with clinical and pathological parameters. *BMC Cancer*. 2013;13.
72. Bruun J, Kolberg M, Nesland JM, Svindland A, Nesbakken A, Lothe RA. Prognostic Significance of beta-Catenin, E-Cadherin, and SOX9 in Colorectal Cancer: Results from a Large Population-Representative Series. *Frontiers in oncology*. 2014;4:118.
73. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Archives of pathology & laboratory medicine*. 2013;137:1710-22.
74. Hwang WL, Jiang JK, Yang SH, Huang TS, Lan HY, Teng HW, et al. MicroRNA-146a directs the symmetric division of Snail-dominant colorectal cancer stem cells. *Nature cell biology*. 2014;16:268-80.

75. Szklo M. Population-based cohort studies. *Epidemiologic reviews*. 1998;20:81-90.
76. Busch EL, Keku TO, Richardson DB, Cohen SM, Eberhard DA, Avery CL, et al. Markers of epithelial-mesenchymal transition and mortality in a population-based prospective cohort of colorectal cancer patients. Unpublished manuscript ed2015.
77. NCI. Colon and Rectal Cancer. 30 January 2015]; Available from: <http://www.cancer.gov/cancertopics/types/colon-and-rectal>
78. Chu H, Chen S, Louis TA. Random Effects Models in a Meta-Analysis of the Accuracy of Two Diagnostic Tests Without a Gold Standard. *Journal of the American Statistical Association*. 2009;104:512-23.
79. Chu H, Zhou Y, Cole SR, Ibrahim JG. On the estimation of disease prevalence by latent class models for screening studies using two screening tests with categorical disease status verified in test positives only. *Statistics in medicine*. 2010;29:1206-18.
80. Zhang J, Cole SR, Richardson DB, Chu H. A Bayesian approach to strengthen inference for case-control studies with multiple error-prone exposure assessments. *Statistics in medicine*. 2013;32:4426-37.
81. Zlobec I, Lugli A. Invasive front of colorectal cancer: dynamic interface of pro-/anti-tumor factors. *World journal of gastroenterology : WJG*. 2009;15:5898-906.