THE INFLUENCE OF MEASUREMENT ERRORS IN TUMOR MARKERS

by

Adeya Shontelle Powell

(Under the Direction of Kevin Dobbin and William McCormick)

ABSTRACT

Measurement error is inherent in the collection of Tumor Markers. In general, when measurement error is present we know that the regression parameters are bias, but for a logistic model there are other concerns. This research sought to answer what happens to Specificity, Area under the curve (AUC), Sensitivity, and the classification accuracy when measurement error was present. We found that there was better discrimination for tumor markers highly correlated with the dichotomous outcome variable; and Specificity, or true negatives, decreased as measurement error increased indicating an increase in the number of false negatives in the presence of measurement error.

INDEX WORDS: Measurement error, Tumor Markers, AUC, Specificity, Sensitivity, Logistic Regression, Classification, Accuracy, Bias

THE INFLUENCE OF MEASUREMENT ERRORS IN TUMOR MARKERS

by

Adeya Shontelle Powell

B.S, Georgia State University, 2005

M.Ed., The University of Georgia, 2006

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2012

THE INFLUENCE OF MEASUREMENT ERRORS IN TUMOR MARKERS

by

ADEYA SHONTELLE POWELL

Major Professors:	Kevin Dobbin
	William McCormick
Committee:	Cheolwoo Park
	T. N. Sriram

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

December 2012

ACKNOWLEDGEMENTS

Thanks to Professor Dobbin, Dr. Tiffany Aholou, Stephanie Cooke, Dr. Sherre Bush, the Statistics department at UGA, and all my friends and family who helped finish his degree. Without your love and support success would be an empty shell-- void without marrow.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The use of medical biopsies and tumor marker tests can provide doctors with combined scientific evidence as to the presence or absence of cancer and tumors. Biopsies are invasive, time consuming, and expensive procedures, but they are also definitive Gherardi (2009). Tumor marker tests are less invasive and cost less than biopsies, but the outcomes are not always precise or conclusive. There are many elements that can affect the viability of a marker, namely consistency of laboratory procedures and processes, and the internal characteristics of the marker itself. Statistically, the presence of measurement error internal and external to tumor markers can cause them to be less precise predictors of cancer.

Measurement error is a common issue for all statisticians. Carroll et al. (2006) defined two types of errors: Berkson and classical measurement errors, the two errors that statisticians most commonly face. According to Carroll et al. (2006), when we have a linear regression model there is attenuation (lessening or reduction) of the regression coefficients under the classical measurement error model. My study seeks to explore the impact on the logistic model when classical measurement error is introduced. Additionally, this research aims to examine the effects of that introduction on all aspects of the model such as regression weights, area under the curve (AUC), specificity, and sensitivity. We hypothesize that the regression coefficients will decrease

as measurement error increases and that accuracy will decrease as the measurement error increases, regardless of the sample size.

CHAPTER 2

LITERATURE REVIEW

Tumor Markers. Tumor markers are chemicals produced by cancer or by other cells in response to cancer and can be found in bodily fluids, such as blood, urine and body tissues (National Cancer Institute, 2011). Tumor marker levels can be measured and used to screen for early stages of various forms of tumors and cancers (National Cancer Institute, 2011). Doctors consider markers to be surrogate indicators for disease, so they test patients for the presence or absence of certain markers to predict cancer or likelihood of disease. Testing for tumor markers is less expensive and less invasive alternative to having a biopsy (the physical removal of cells to observe microscopically and make definitive cancer diagnoses). Although tumor marker tests cannot be used as sources of definitive cancer diagnoses, they can help doctors detect the presence of some cancers so that biopsies, if necessary, can be more precise.

An example of a common tumor marker test is the prostate-specific antigen (PSA) blood test for men, which tests for prostate cancer. Men with prostate cancer usually have high PSA levels in their blood; however, for men without prostate cancer, the PSA levels can also be found to be high. Also, normal PSA levels do not indicate the absence of prostate cancer. Other factors such as age, body mass index (BMI) and race can affect PSA test results. PSA levels naturally increase with age, so the National Comprehensive Cancer Network (NCCN)

recommends all men begin receiving yearly PSA tests at age 40 to establish a baseline for which future tests are compared (Loeb & Catalona, 2008). Despite the fact that elevated PSA levels aren't definitive tumor markers, since PSA screening has become more routine, prostate mortality has declined by 35%, and there has been a 75% reduction in the incidence of late stage disease by 1992 (Etzioni et al., 2008). Although age, BMI and race are known to affect PSA levels, other factors such as prior ejaculation within 48 hours of the PSA test are also known to affect the outcomes (Loeb & Catalona, 2008). Therefore, various internal factors can affect PSA test results.

In general, high levels of PSA do not equate to a prostate cancer diagnosis, but they should be a cause for concern. Other marker tests such as the Alpha fetoprotein (AFP) test for liver cancer are considered to be reliable because they are less affected by internal factors Diamandis, E. P. (2002). High levels observed for this marker can be used as a standalone diagnosis of cancer. This standalone ability to diagnose is also true for the carbohydrate antigen 125 (CA 125) test, which is used for diagnosing ovarian cancer. In essence, markers are used to detect the presence of a disease (and in some cases to diagnose disease), such as various forms of cancers and tumors; and a high level of such markers does not always indicate cancer or a tumor depending on the particular marker. In some cases no one marker can be used as a conclusive means for detecting the presence of cancer; multiple markers might serve as indicators. Therefore, combining several tumor marker tests into a single test adds more evidence that a particular cancer is present than does a standalone test (Yarbro, Foodman, & Frogge (2005).

The last section detailed how tumor markers can be used to aid diagnosis, but tumor markers can also be used for prognosis. With some markers, depending on the tumor maker

levels, doctors can predict the behavior or outlook of certain cancers and determine whether certain treatments or drugs are likely to work. Examining the fluctuation of the tumor marker serves as an indicator of responsiveness to treatment for cancer patients. For the patients in remission, some tumor markers are solely used to test for the return of certain cancers. Tumor marker levels can change over time due to various factors; therefore, it is best to have multiple test results. Multiple test results allow for the tracking of changes in the markers to follow the course of the disease.

Because of unnecessary cost and stress caused by mistakes, reliability of tumor marker tests is very important. The reliability of any tumor marker test is determined by the test's *sensitivity* and *specificity*. The sensitivity of a tumor marker test is the ability for the presence of a marker to identify people who have the disease. If the tumor marker test is not sensitive enough then many patients who have disease will not be diagnosed, leading to false negative test results. A false negative occurs when someone who has the disease is diagnosed as not having it. Tumor marker test specificity measures the opposite by identifying individuals without the disease when the disease is not present. A false positive test result can occur if the tumor marker test lacks specificity, and a patient who does not have the disease is diagnosed with it. Similar to Type I and Type II statistical errors, the best tumor marker tests maintain a healthy balance between sensitivity and specificity. The goal is to prevent such errors from occurring by utilizing and standardizing guidelines on how obtain, when to obtain and how to handle/care for the proposed tumor markers.

According to Sturgeon et al. (2008), the pre-analytical stage is the most important stage because errors occur up to 10 times as often during this stage. During this stage, doctors have to

identify the most appropriate tumor marker test to screen for disease and if it will result in a viable result. For example, if the prostate has been manipulated prior to taking blood for a PSA test or if a patient has menstruated prior to a CA125 test, then these tests might indicate elevated marker levels. Medications and cannabis are also known to alter test results of some markers. Since internal and external factors can alter test results, pre-analysis of the patient is necessary. Pre-analysis involves knowing the appropriate test to use, the appropriate procedure you can and cannot do before conducting a test, and screening the patient for known confounding variables. Included in this stage is good laboratory handling technique, which involves: clear instructions for the appropriate tube necessary for each test, and correct handling of the material such as maintaining consistent temperatures since some tumor markers are sensitive to high ambient temperature (Sturgeon et al, 2008). The pre-analytical stage is an additional factor that could cause variability in test results. Besides the pre-analytical stage factor, there is intra-individual variation (within person variation), assay imprecision, and also, cut-point decision (Badrick, Hawkins, Wilson, and Hickman, 2005).

There are many reasons why tumor marker test results vary. This variation can lead to inconsistent scientific conclusions about the viability of tumor marker to be indicators of disease. By examining the within-subject variation, one can consider an individual's dispersion and compare it with the group such as the Cochran C test – a one-sided upper limit variance outlier test named after William Cochran – which tests for outliers and can be used to check for causes in variation. Another reason for tumor marker variations is the use of assays that are not standardized or lack reproducibility (McShane et al., 2005). Finney (1978) defined a biological assay (bioassay) as an experiment for estimating the nature, constitution or potency of a material

by means of the reaction that follows its application to a living matter (as cited in Rand, 1995, p.36). Even if the assay technique is standardized by quality controls, random variation (measurement error) in assay results may persist due to assay imprecision, and variation between people. For example, an immunohistochemical (IHC) assay is a process for detecting proteins in tissue cells. Many of these assays require selection of 'best' regions to score, and subjective assessments of staining intensity and percentage of stained cells (Altman et al., 2012). Each step of the test allows for subjective judgment to introduce measurement error in the test. According to Altman et al.(2012), "the impact of measurement error is attenuation of the estimated prognostic effect of the marker. Good prognostic performance of a marker cannot be achieved in the presence of a large amount of imprecision." Therefore they recommend reporting all strategies used to reduce the measurement error, "such as taking the average of two or three readings to produce a measurement with less error, potentially increasing the power of the study and hence the reliability of the findings." McShane (2005) noted that tumor marker studies have not been reported in a rigorous fashion and generally lack generalizability. Tovey (2011) explains that the choice of cutpoint used helps define the results of the test. The type of cut point should be based on the assay variability. Even more important is the appropriate use of a sharply defined cutpoint. Because there is uncertainty associated with laboratory measurement, sharply interpretive cutpoint can be inappropriate. For some diseases such as diabetes there cannot be such a boundary value since there is a gray zone between being normal and having full-blown diabetes (Badrick et al., 2005).

Others have defined the need to have reproducibility, repeatability and reliability for good test measures, which means that the results of the test are identical or closely related.

Kanchanaraksa (2008) defines variation as a measure of intra-subject variation, intra-observer and inter-observer variation. Intra-subject variation is the variation in tests conducted over time in the same individual. The causes of these types of variations depend on a number of factors. Variations can be caused by physiological or environmental reasons. Inter-observer variation is the variation that occurs when multiple observers or researchers examine the result/material. Intra-observer variation occurs when the same person examines the same result at different times. Differences are due to the extent to which the observers agree or disagree. A test for rater or observer agreement (or percentage of agreement for the different times) can be used as test for reliability. One of the few studies to look at the explanation of the lack of reproducibility of lab results was conducted by Dobbin et al. (2008), when they compared biological variation between patients, between-lab measurement error, and within-lab measurement error to find which resulted in the most variation. They were able to show that if the biological variation between patients were large relative to the between lab and within measurement errors then the quantitative assay was reproducible. They determined this by looking at the intra-class correlation (ICC). Intra-class correlation describes the tendency for those in the same group to have the same behavior.

Intra-class correlation (ICC)

$$ICC_{BetweenLab} = \frac{\sigma^2_{Patient}}{\sigma^2_{Patient} + \sigma^2_{Lab} + \sigma^2_{Error}}$$

$$ICC_{WithinLab} = \frac{\sigma^2_{Patient}}{\sigma^2_{Patient} + \sigma^2_{Error}} \quad (1)$$

The assay was reproducible when the ICC was close to one. When the ICC was close to zero, the assay was not reproducible. The model (or score) on a particular assay was a combination of person effect, lab effect and measurement error such as biological (DNA).

Measurement Error. The presence of measurement error tends to increase the variability of independent variables thereby adding discriminatory difficulties. There are two types of measurement errors: classical and Berkson measurement errors.

In classical measurement error theory, measurements are assumed to be random and independently distributed about the true value. So the classical measurement error model represents measurement error as an additive model where the observed value W is a function of the true score plus error. Observed W= X + E (unobserved true score plus error). W is considered an unbiased estimate of the true score. An example of classical measurement error occurs when we measure depression using a questionnaire. The observed score on depression, W, is not fixed and has a true score on depression and measurement error. When classical measurement errors are present in the regression models, they are known to bias the linear regression coefficients.

In Berkson's measurement error theory, the true score is a function of observed score plus error, X= W+ E. With the Berkson model, X and W are linearly related with a non-constant variance. A classic example of a Berkson measurement error, when trying to measure the amount of herbicide absorption by a plant we can't look at the amount absorbed, X, directly, but we can observe W, the concentration of the herbicide in the plant. Plants can receive the same amount of herbicide but absorb differently due to the makeup of the plant (type of plant) (Carroll, Ruppert,

Stefanski, and Crainiceanu, 2006). Therefore, the amount of concentration of herbicide is fixed because of makeup of the plant. Certain plants will absorb a certain fixed amount. The true concentration of X varies and is a function of makeup of the plant and the amount of concentration. In the presence of Berkson errors, there is no bias in the linear regression coefficients and little to no bias in logistic regression coefficients. Therefore, Carroll et al. (2006) advise researchers to know what types of measurement may be present when starting a study.

In a simple regression model, measurement error bias leads to bias slope estimates and loss of power for the classical measurement error model. Suppose we have a simple regression equation where Y is the dependent variable, X is the independent variable, β_X is the slope regression coefficient, β_0 is the intercept, and ε is the error.

$$(2)\ Y= \beta_0 + \beta_X (X) + \varepsilon$$

If the data is contaminated and we are using a substitute for X, such as W, where W= X+ E (as in the classical additive measurement error model), then additional variance is added into the model. Suppose the Var (X), Var (ε), and Var (E) are all standard normal and independent then the variance W, instead of being one, is actually two thereby causing increased variability in the data modeling (Y, W), which causes loss of power when trying to estimate the Y hat line (Carroll et al., 2006, p.42). For the classical measurement error model, the slope regression coefficient, β_X, is attenuated. When you model Y on W, you are not actually getting, β_X, as the regression coefficient but the reliability ratio, λ

$$(3)\ \lambda= \sigma^2_X / (\sigma^2_X + \sigma^2_E).$$

Lamba is a function of the new beta that lessens the original beta, $\beta^* = \lambda\beta_X$, and is less than one. Therefore the new equation becomes:

$$(4)\ Y = \beta_0 + \beta^* (X) + \varepsilon$$

Rewriting the equation in terms of the observed variable W adds an additional error component to the model:

$$(5)\ X = W - E, \text{ therefore } Y = \beta_0 + \beta^* (W) + (\varepsilon - \beta^* (E))$$

$$(6)\ \sigma^2_{\varepsilon} + (\beta^*)^2\sigma^2_E > \sigma^2_{\varepsilon}$$

The new error is $(\varepsilon - \beta^* (E))$ and the residual variance of the regression Y on W becomes

$$(7)\ \text{Var}(Y|W) = \sigma^2_{\varepsilon} + (\beta^*)^2\sigma^2_E = \sigma^2_{\varepsilon} + (\lambda\beta_X)^2\sigma^2_E$$

Not only is there loss of power, but there is also more error about the line with the additional error variance and the slope is attenuated. Attenuation is not present in the Berkson model; however, tumor makers are an example of classical measurement error model. The effect of measurement error is gravely important because as shown above, naive estimate of the slope (Y, W) is less variable than the true data estimator (Y, X) (Buzas, Stefanski and Tosteson, 2004, as cited in Carroll et al., 2006). Therefore, a particular marker that might be as good as an estimator for that particular disease might have measurement issues due to the marker itself or the assay process.

Logistic Regression. When scientists want to assess whether marker tests can determine the presence of disease, they collect data on people who have disease and those who do not have the disease and correlate the various tumor markers scores with disease status. One or a combination of markers may be found to be the best predictors of disease based on which markers produce the best sensitivity. The disease status is a categorical dependent variable, so it is best for logistic regression to be used. The independent variables are categorical and/or continuous. For example, Hwa et al. (2008) sampled 55 female patients with breast cancer and 39 without breast cancer and collected serum levels of "carcinoembryonic antigen, breast cancer-specific cancer antigen 15.3 (CA15-3), tissue polypeptide-specific antigen (TPS), soluble interleukin-2 receptor (sIL-2R) and insulin-like growth factor binding protein-3 (IGFBP-3)." After performing univariate and multivariate logistic regressions to evaluate the association between biomarkers and breast cancer, they found that the serum level of TPS had the best predictive value, with a sensitivity of 80% at an optimal cut-off value of 69.1 UL.

Logistic regression is a type of generalized linear model, which includes a broad class of models that allows for non-normal response variables and predicts group membership based on a set of normal and non-normal independent variables. Linear regression is also a type of generalized linear model, which assumes four things: linearity, independence, homoscedasticity, and normality. The independent and dependent variables are assumed to be linear. The errors are assumed to be independent. The variance of the errors are assumed to be constant no patterns (homoscedasticity) and the errors (and dependent variable is assumed to be normal). Logistic regression violates a number of these assumptions, namely the linearity assumption, the assumption of constant error variance, and normally distributed error variance. The assumptions

of independence in the errors are still held. What is not so obvious is that for both types of models we assume we have all the explanatory variables needed and that these variables are measured without error.

The relationship between the predictors and the response variables can be explained by a probability risk function, $\hat{P}(X)$, also known as the logistic model (Kleinbaum and Klein, 2010).

$$(8)\quad P(X) = \frac{1}{1+e^{-(\alpha+\sum\beta x)}}$$

$$(9)\quad \frac{P(X)}{1-P(X)} = e^{-(a+\sum b_i x_i)}$$

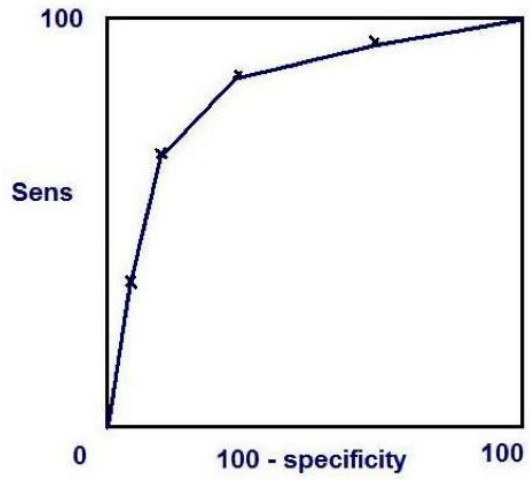
In (log for individual X)= $a+\sum b_i x_i$

The logistic model predicts the probability of getting the disease for a given set of independent variables. The log odds is a linear function of the beta weights. In fact, the intercept, a, is the log odds for a person who has zero on all the explanatory variables (no presence of any tumor markers). Beta, for a categorical response variable, represents one unit change in the log odds.

In order to assess the performance of a logistic model, it is helpful to use the receiver operating curve (ROC). The ROC is used for fitting models and computing the AUC to measure discriminatory performance (Kleinbaum and Klein, 2010). The discriminatory performance can be determined by looking at the proportion of observed and non-observed cases that were

correctly predicted. Cut-points allow scientists to classify people who were not in the study as having the disease or not having the disease. For example, if the cut-point was established as .05, then if the probability risk function is greater than .05, the person is said to have the disease. If a person's probability risk function is less than .05, then he would not have the disease. Usually the cut-points are not chosen randomly, but instead they are based on literature or some mathematical formulation. After the cut-point is carefully chosen, a classification table can be obtained. The classification is a cross table of the predicted outcomes based on the cut-points (number who had the disease and number who did not have the disease) versus the actual observed outcomes (who had the disease and who did not have the disease). The sensitivity represents the proportion of those who had the disease who our cut point predicted would have the disease. Specificity is the amount of true non-cases, or the proportion of people without the disease that we correctly predicted would not have the disease. One minus the specificity is the false positive rate and one minus sensitivity is the proportion of false negatives. The goal is to have the proportion of false positives to be close to zero and also have high sensitivity (when the proportion of true positive is greater than the proportion of false positives). Thus, the ROC plots the sensitivity vs one minus specificity on the graph. The larger the area under a ROC curve, the better the discrimination. The ROC displays all possible cut-points and allows for identification of the optimal cut-point.

Figure I: ROC Graph



When the area under the ROC curve (AUC) is large, a randomly chosen true case has higher probability of being predicted than a randomly chosen non-case (Kleinbaum and Klein, 2010). The values for the AUC can range from 0 to 1. If the area is between 0 and .5, there is negative discrimination. If the AUC is exactly .50 then there is no discrimination at all; if the AUC is between .50 and 1 then there is positive discrimination. Finally, if the area is one then there is perfect discrimination. We roughly want to have discrimination above .60. According to Kleinbaum and Klein (2010), an excellent discriminator has an AUC score of .90 and above, while a good discriminator has an AUC score ranging from .80 to .90. A fair discriminator has a score between .70 and .80. A poor discriminator has an AUC score between .60 and .70, and a failed discriminator has an AUC score between .50 and .60.

CHAPTER 3

METHODS

Choosing the correct cut point was vital to the success of this research study. There were many options, but the best options were given in Efron (1975), which recommended a logistic cut point of .50. Hosmer and Lemeshow (2000) describe this as a commonly used cut point in logistic regression. Using .50 as the static cutpoint, I aimed to find out what occurs to the coefficients under different sample sizes, when error is present in the data, under various alpha and beta values.

Using R Software version 2.13, a set of normal values were generated for the linear part of the logistic equation using the rnorm function. To obtain the probabilities, plogis R function – the cumulative distribution for a logistic random variable that returns the probabilities, exp(x)/(1+exp(x)) – was used (Chihara and Hesterberg, 2011). Error was generated by adding a random normal variable to the generated normal variable. The new random normal variable, called xerr, with standard error (tau) varied throughout my model as a way of controlling how much error was added. A new logistic regression model was fitted using the original set of zeros and ones (outcomes) with the new xerr variable. This produced a new alpha and beta (intercept and slope) when error was present. Using the new model with the error, a new set of probabilities based on the error were generated.

Classification of the new probabilities was based on the use of an if-else logical statement and the predicted probabilities were compared to the static cutpoint. The area under the curve was obtained by using the performance function in R. Accuracy was calculated as the percentage of time the original classification matched my new classification. The number of true positives, false positives, true negatives, and false negatives were calculated manually. The number of true negatives was defined as the proportion of times the person had the disease (original classification was zero) and my new classification was actually zero.

After defining all four possibilities, those values were used to calculate sensitivity and specificity. Sensitivity was defined as the number of true positives to the number of true positives and false negatives. Specificity was defined as the number of true negatives to the number of true negatives and false positives. The program was built with a loop so that the program could run with numerous iterations. The number of iterations needed was determined to be 500 and the average and the variance for the returned AUC, intercept*, slope*, accuracy, sensitivity, and specificity were calculated.

The standard error, tau, ranged from .25, .50, 1, 1.5, 2, 3, and 5. The Beta (slope) varied from 0 to 6, in increments of 1. For a given tau and beta, I varied alpha (intercept) from 1 to 3 (in increments of 1). The intercept and slope were the original values used to create the logistic regression function. The values returned were the alpha star and beta star, which was the new estimated alpha and beta (or intercept * and slope*) under the condition of error. The sample size varied as well with ranges from 50, 100, 200, to 1,000. Table I (below) shows the number of variables that varied and their respective quantities.

Table I. Replications

Replications			
Sample Sizes	Tau	Beta	Alpha
1000	.25,.50,1,1.5,2,3,5	0,1,2,3,4,5,6	1,2,3
200	.25,.50,1,1.5,2,3,5	0,1,2,3,4,5,6	1,2,3
100	.25,.50,1,1.5,2,3,5	0,1,2,3,4,5,6	1,2,3
50	.25,.50,1,1.5,2,3,5	0,1,2,3,4,5,6	1,2,3

For each sample size, a total of 147 conditions were created — with a total of 588 conditions.

CHAPTER 4

RESULTS

The study results comprised expected and unexpected findings, such as linear coefficients being attenuated in all cases no matter the sample size.

Table II. Sample Size (SS) 1000 Attenuation

		Tau =.25; Beta=0			Tau =.25; Beta =1			Tau =.25; Beta =2		
		alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Mean	Intercept	1.0040	2.0014	3.0259	0.9876	1.9889	2.9946	0.9619	1.9286	2.9121
	Slope	0.0015	0.0090	-0.0083	0.9320	0.9437	0.9422	1.8085	1.8192	1.8289
	CutPoint	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
	Accuracy	0.7314	0.8802	0.9529	0.7315	0.8483	0.9312	0.7852	0.8345	0.8916
	AUC	0.5000	0.5000	0.5000	0.6117	0.5353	0.5055	0.7487	0.7042	0.6452
	Sensitivity	1.0000	1.0000	1.0000	0.9156	0.9896	0.9993	0.8715	0.9403	0.9768
	Specificity	0.0000	0.0000	0.0000	0.3078	0.0810	0.0116	0.6258	0.4680	0.3137
Var	Intercept	0.0057	0.0095	0.0217	0.0067	0.0135	0.0284	0.0083	0.0154	0.0337
	Slope	0.0044	0.0091	0.0203	0.0071	0.0116	0.0188	0.0137	0.0155	0.0240
	CutPoint	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Accuracy	0.0002	0.0001	0.0000	0.0002	0.0001	0.0001	0.0002	0.0001	0.0001
	AUC	0.0000	0.0000	0.0000	0.0005	0.0004	0.0001	0.0002	0.0004	0.0007
	Sensitivity	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0002	0.0001	0.0000
	Specificity	0.0000	0.0000	0.0000	0.0028	0.0017	0.0003	0.0010	0.0019	0.0032

As seen in Table II, when error was introduced, the intercept and slope, were less than the original alpha and beta. As alpha or beta increased, more attenuation occurred. Sensitivity increased as alpha increased, and decreased as beta increased. Specificity performed in the opposite way decreasing as alpha increased, and increasing as beta increased. The AUC

increased as beta increased, but decreased mostly as alpha increased. Accuracy increased as alpha increased, but there was no clear pattern with beta (sometimes there was an increase and sometimes not and the differences were usually very slight).

A major concern was what would happen to the accuracy, parameters, AUC, and other variables once more error was introduced. Table III, below, depicts what happened in this case.

Table III. SS1000. Attenuation with Error Increase

		Tau =.25; Beta =2			Tau =.50; Beta =2			Tau =1.5; Beta =2		
		alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Mean	Intercept	0.9619	1.9286	2.9121	0.8819	1.7741	2.6916	0.6840	1.4067	2.1454
	Slope	1.8085	1.8192	1.8289	1.4078	1.4295	1.4569	0.4236	0.4410	0.4592
	CutPoint	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
	Accuracy	0.7852	0.8345	0.8916	0.7629	0.8200	0.8842	0.6833	0.7816	0.8709
	AUC	0.7487	0.7042	0.6452	0.7185	0.6689	0.6105	0.5985	0.5388	0.5098
	Sensitivity	0.8715	0.9403	0.9768	0.8665	0.9428	0.9799	0.8846	0.9777	0.9969
	Specificity	0.6258	0.4680	0.3137	0.5705	0.3950	0.2410	0.3124	0.1000	0.0227
Var	Intercept	0.0083	0.0154	0.0337	0.0070	0.0124	0.0262	0.0049	0.0075	0.0134
	Slope	0.0137	0.0155	0.0240	0.0096	0.0137	0.0175	0.0018	0.0022	0.0033
	CutPoint	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Accuracy	0.0002	0.0001	0.0001	0.0002	0.0002	0.0001	0.0002	0.0002	0.0001
	AUC	0.0002	0.0004	0.0007	0.0003	0.0005	0.0007	0.0003	0.0003	0.0001
	Sensitivity	0.0002	0.0001	0.0000	0.0002	0.0001	0.0000	0.0004	0.0001	0.0000
	Specificity	0.0010	0.0019	0.0032	0.0013	0.0023	0.0032	0.0026	0.0016	0.0004

No matter the sample size, the following pattern emerged: As the standard error increased, there was more attenuation in the intercept and the slope. In the above table, the actual slope was 2, but when the standard error was 1.5 the predicted slope was .4592. Although not presented here, when tau was 5 and beta was 2, the predicted slope was .05 – a very far cry from the original value. The intercept was attenuated as well, but not as fast and sharply as the slope. As one might suspect, accuracy decreased as error increased. The AUC decreased as error

increased. Specificity decreased as error increased. Sensitivity increased and sometimes decreased as the error increased with no clear pattern. The findings indicate that as the error increased, more people were classified in one particular class (given all ones).

Two unexpected errors emerged when sample size varied: error in the prediction or glm.fit error for the fitted probabilities/algorithm did not converge. This occurs we have a groups that is not completely separate such that we can draw a line with a probability of .50. This means the results obtained yield unstable or unreliable results (Faraway, 2006).

Table IV. Convergence Error within Sample Size 50

		Sample Size =50 Convergence Errors								
		Tau =.25; Beta=0			**Tau =.25; Beta =1**			**Tau =.25; Beta =2**		
		alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Mean	Intercept	1.0386	1.0386	1.0386	0.8720	1.7992	1.7992	0.6540	1.3167	2.0982
	Slope	0.0061	0.0061	0.0061	0.0342	0.0479	0.0479	0.0488	0.0610	0.0542
	CutPoint	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
	Accuracy	0.7300	0.7300	0.7300	0.6992	0.8404	0.8404	0.6622	0.7753	0.8702
	AUC	0.5072	0.5072	0.5072	0.5145	0.5051	0.5051	0.5270	0.5138	0.5057
	Sensitivit	0.9897	0.9897	0.9897	0.9773	0.9979	0.9979	0.9378	0.9902	0.9984
	Specificity	0.0247	0.0247	0.0247	0.0517	0.0123	0.0123	0.1162	0.0373	0.0131
Var	Intercept	0.1278	0.1278	0.1278	0.1162	0.2360	0.2360	0.1178	0.1499	0.3559
	Slope	0.0051	0.0051	0.0051	0.0052	0.0085	0.0085	0.0048	0.0063	0.0102
	CutPoint	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Accuracy	0.0041	0.0041	0.0041	0.0042	0.0030	0.0030	0.0044	0.0034	0.0025
	AUC	0.0006	0.0006	0.0006	0.0013	0.0005	0.0005	0.0022	0.0013	0.0008
	Sensitivit	0.0014	0.0014	0.0014	0.0032	0.0001	0.0001	0.0156	0.0007	0.0001
	Specificity	0.0054	0.0054	0.0054	0.0112	0.0024	0.0024	0.0324	0.0081	0.0033

When the sample size was 50, as beta increased, convergence depended on the value of alpha. Notice in Table IV (above) that for beta=0, when alpha>2, there is a replication of the

values (the same thing occurs when beta=1 and alpha=3). The R program generated an error when these replicated values emerged. In Table IV, when beta=0, convergence problems occurred when alpha≥2. However, when beta=1, convergence problems emerged when alpha≥3. These problems were consistent despite the number of time the program ran and no matter the value of tau (see Table V). The reason for this error is too few classifications in one group, which occurs when the probability of being in one group is greater than the other. The beta weight is a function of the probabilities, so as the beta weight increases, the probability does as well.

Table V. Convergence Errors across tau for sample size 100

		Sample Size= 100 Convergence Errors across tau								
		Tau =.25; Beta=0			**Tau =.50; Beta=0**			**Tau =1.5; Beta=0**		
		alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Mean	Intercept	1.0219	2.1000	2.1000	1.0312	2.0636	2.0636	1.0229	2.0906	2.0636
	Slope	-0.0347	-0.0154	-0.0154	0.0071	-0.0027	-0.0027	-0.0038	0.0026	-0.0027
	CutPoint	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
	Accuracy	0.7311	0.8823	0.8823	0.7322	0.8789	0.8789	0.7314	0.8819	0.8789
	AUC	0.5012	0.5002	0.5002	0.5013	0.5007	0.5007	0.5015	0.5002	0.5007
	Sensitivity	0.9981	0.9999	0.9999	0.9983	0.9999	0.9999	0.9984	1.0000	0.9999
	Specificity	0.0043	0.0006	0.0006	0.0042	0.0014	0.0014	0.0047	0.0003	0.0014
Var	Intercept	0.0549	0.1237	0.1237	0.0560	0.1171	0.1171	0.0527	0.1192	0.1171
	Slope	0.0505	0.1201	0.1201	0.0468	0.1006	0.1006	0.0169	0.0356	0.1006
	CutPoint	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Accuracy	0.0019	0.0010	0.0010	0.0019	0.0010	0.0010	0.0018	0.0010	0.0010
	AUC	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
	Sensitivity	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
	Specificity	0.0005	0.0001	0.0001	0.0003	0.0002	0.0002	0.0004	0.0000	0.0002

A finding outside the scope of this study can be found when comparing Tables IV and V. The point where the error occurs depends on the sample size. For example, when beta=0 and the sample size is 50, then for alpha≥2, more convergence problems surfaced. When the sample size was increased to 100 and when alpha≥3, there were more convergence problems. These convergence problems are due to the beta weight affecting the probability. Also, in a small sample size, the probability determines the amount of people in each class. When the sample size was 1,000, there were no convergence issues for the alphas chosen, but with an alpha≥5 and beta=0, the algorithm would not converge.

Table VI. Accuracy within and across Tau

Accuracy/SS=1000									
	Beta=1			Beta=2			Beta=6		
	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Tau=.25	0.7315	0.8483	0.9312	0.7852	0.8345	0.8916	0.8828	0.8869	0.8941
Tau=.50	0.7248	0.8463	0.9309	0.7629	0.8200	0.8842	0.8302	0.8387	0.8487
Tau= 1	0.7086	0.8451	0.9309	0.7152	0.7925	0.8738	0.7422	0.7546	0.7741
Tau=1.5	0.7015	0.8443	0.9301	0.6833	0.7816	0.8709	0.6870	0.7028	0.7309
Tau=2	0.6974	0.8450	0.9307	0.6679	0.7767	0.8701	0.6505	0.6737	0.7085
Tau=3	0.6964	0.8444	0.9305	0.6530	0.7751	0.8702	0.6109	0.6453	0.6920
Tau=5	0.6970	0.8443	0.9310	0.6486	0.7751	0.8699	0.5812	0.6292	0.6859

Based on previous known research, one of the hypothesis stated that that the accuracy would decrease as the error increased, which is what happened. I defined accuracy as the percentage of times the original classification matches the new classification when error is introduced into the model using the common cut point (.50). However, the decrease in accuracy was very slight and it did not consistently decrease when beta was small. In fact, when beta was large the decrease in

accuracy was more pronounced and consistent. This pattern was observed regardless of the sample size. The accuracy did decrease slightly as sample size increased (see Table VII below).

Table VII. Accuracy across sample sizes

Accuracy across sample sizes									
	Beta=1			Beta=2			Beta=6		
	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Tau=.25									
N=100	0.7388	0.8514	0.9324	0.7872	0.8379	0.8950	0.8841	0.8918	0.8967
N=1000	0.7315	0.8483	0.9312	0.7852	0.8345	0.8916	0.8828	0.8869	0.8941
Tau=1									
N=100	0.7155	0.8450	0.8450	0.7188	0.8019	0.8788	0.7475	0.7596	0.7745
N=1000	0.7086	0.8451	0.9309	0.7152	0.7925	0.8738	0.7422	0.7546	0.7741

As seen in Table VII, as sample size increases, the accuracy decreases. The only times when there was discrepancy in the decrease (such as when beta=1 and alpha=2) was when the R software reported convergence problems due to sample size and beta. For the most part, a clear pattern was observed.

Table VIII. AUC

AUC									
	Beta=1			Beta=2			Beta=6		
	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Tau=.25									
N=100	0.6142	0.5493	0.5207	0.7452	0.7033	0.6537	0.8794	0.8802	0.8732
N=1000	0.6117	0.5353	0.5055	0.7487	0.7042	0.6452	0.8801	0.8768	0.8716
Tau=1									
N=100	0.5561	0.5190	0.5190	0.6502	0.5977	0.5508	0.7332	0.7228	0.6994
N=1000	0.5510	0.5052	0.5006	0.6506	0.5894	0.5391	0.7341	0.7224	0.7052

Table VIII summarizes what happens to the AUC as the sample size increases, beta and alpha increase, and as error increases. Note, as previously stated the AUC decreases with alpha and increases with the beta value. When error was introduced, the AUC was less. Also, the AUC had somewhat inconsistent patterns, decreasing and sometimes increasing with sample size. What appears to impact the AUC values most are the beta weight and the amount of error.

Table IX. Sensitivity

Sensitivity									
	Beta=1			Beta=2			Beta=6		
	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Tau=.25									
N=100	0.9160	0.9856	0.9978	0.8730	0.9402	0.9766	0.9007	0.9189	0.9331
N=1000	0.9156	0.9896	0.9993	0.8715	0.9403	0.9768	0.9000	0.9167	0.9320
Tau=1									
N=100	0.9481	0.9934	0.9934	0.8635	0.9589	0.9882	0.8004	0.8486	0.8893
N=1000	0.9503	0.9980	0.9999	0.8665	0.9583	0.9903	0.7954	0.8484	0.8904

Table X. Specificity

Specificity									
	Beta=1			Beta=2			Beta=6		
	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3	alpha=1	alpha=2	alpha=3
Tau=.25									
N=100	0.3125	0.1130	0.0435	0.6175	0.4664	0.3308	0.8581	0.8416	0.8132
N=1000	0.3078	0.0810	0.0116	0.6258	0.4680	0.3137	0.8602	0.8368	0.8112
Tau=1									
N=100	0.1641	0.0447	0.0447	0.4368	0.2364	0.1135	0.6660	0.5969	0.5094
N=1000	0.1516	0.0125	0.0013	0.4348	0.2205	0.0880	0.6728	0.5965	0.5200

In Table IX, sensitivity increased as alpha increased, and decreased as beta decreased. The sensitivity results were inconsistent across sample sizes and with increase in error. Interestingly,

the specificity (Table X) decreased as alpha increased and increased with beta. It shows consistent patterns as error increased, but there were no patterns when sample size increased.

CHAPTER 5

CONCLUSION

When tumor markers are highly correlated with disease and error is introduced through the assay process, the accuracy of the test decreases. Sample size does not impact this trend. There is better discrimination the more the tumor marker is correlated with the outcome (dichotomous indicator of presence of a tumor), but the discrimination ability is hampered by the presence of measurement error. As Carroll (2006) predicted, the behavior of the linear coefficients were attenuated in the presence of error. As more error was introduced, more attenuation surfaced. As Sample size increase accuracy decreased. For Specificity (which indicates the number of true negatives) and Sensitivity (which indicates the number of true positives), as the sample size increased neither quantity showed a clear pattern. However, Specificity decreased as error increased.

In summary, as more error is introduced, there is less accuracy, which makes it more likely that doctors will falsely identify a person who does not have the disease. This study was limited by the fact that only one predictor was included in the model. The presence of multiple predictors (possibly with differing errors) would be very useful in future studies since tumor markers are supposed to combat the presence of error.

CHAPTER 6

REFERENCES

Altman, D., McShane, L., Sauerbrei, W., & Taube, S. (2012). Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): Explanation and Elaboration. *BMC medicine*, *10*(1), 51. Retrieved from http://www.biomedcentral.com/1741-7015/10/51/.

Badrick, T., Hawkins, R., Wilson, S., & Hickman, P. (2005). Uncertainty of Measurement: What it is and What it Should Be. *Clinical Biochemist Reviews*, *26*(4), 155-158.

Carroll, R. J., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective*. Boca Raton [etc.: Chapman & Hall, Taylor & Francis.

Chihara, L., & Hesterberg, T. (2011). *Mathematical statistics with resampling and R*. Hoboken, N.J: Wiley.

Diamandis, E. P. (2002). Tumor markers: Physiology, pathobiology, technology, and clinical applications. Washington, DC: AACC Press.

Dobbin K., Hamilton S., Thibodeau S., Redston M., Taube S., Jessup J., Wang Z., and the Program for the Assessment of Clinical Cancer Tests (PACCT) (2008). Inter-laboratory assay reproducibility study for loss of heterozygosity on chromosome 18 (18q LOH) in colon cancer. Paper presented at National Cancer Institute: Rockville, MD

Efron, B. (1975). The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. *Journal of the American Statistical Association*, *70*(352), 892-898.

Etzioni , R., Tsodikov, A., Mariotto , A., Szabo, A., Falcon , S., Wegelin, J., . . . Feuer, E. (2008). Quantifying the role of PSA screening in the US prostate cancer mortality decline. *Cancer Causes Control*, *19*(2), 175-181.

Gherardi, G. (2009). Fine-needle biopsy of superficial and deep masses: Interventional approach and interpretation methodology by pattern recognition. Dordrecht: Springer.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Hwa, H., Kuo, W., Chang, L., Wang, M., Tung, T., Chang, K., & Hsieh, F. (2008). Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models. *Journal of evaluation in clinical practice*, *14*(2), 275-280.

Kanchanaraksa, S., & Johns Hopkins University (2008). *JHSPH OCW*. Retrieved from http://ocw.jhsph.edu/courses/fundepiii/PDFs/Lecture16.pdf

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text*. New York: Springer.

Loeb, S., & Catalona, W. J. (2008). What to Do with an Abnormal PSA Test. The Oncologist, 13(3).

McShane, L., Altman, D., Sauerbrei, W., Taube, S., Gion, M., Clark, G. (2005). Reporting Recommendations for Tumor Marker Prognostic Studies. *American Society of Clinical Oncology*, *23*(36), 9067-9072.

National Cancer Institute. (2011). Comprehensive Cancer Information - National Cancer Institute. Retrieved from http://www.cancer.gov/cancertopics/factsheet/detection/tumor-markers

Rand, G. M. (1995). *Fundamentals of aquatic toxicology: Effects, environmental fate, and risk assessment*. Washington, D.C: Taylor & Francis.

Sturgeon, C., Hoffman, B., Chan, D., Ch'ng, S., Hammond, E., Hayes, D., . . . Diamandis, E. (2008). National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines for Use of Tumor Markers in Clinical Practice: Quality Requirements. *Clinical Chemistry, 54*(8), 1-10.

Tovey, M. G. (2011). *Detection and quantification of antibodies to biopharmaceuticals: Practical and applied considerations*. Hoboken, N.J: Wiley.

Yarbro, C. H., Goodman, M., & Frogge, M. H. (2005). Cancer nursing: Principles and practice. Sudbury, Mass: Jones and Bartlett Publishers.

CHAPTER 7

APPENDIX

RCode

```
library(ROCR)

n <- 1000

fitglm <- function(iteration,sigma,tau,beta){
x <- rnorm(n,0,sigma)
intercept <- 4
ystar <- intercept+beta*x
z <- rbinom(n,1,plogis(ystar))
xerr <- x + rnorm(n,0,tau)
model<-glm(z ~ xerr, family=binomial(logit))
int<-coef(model)[1]
slope<-coef(model)[2]
pred<-predict(model,type="response")
cutp<-.5
result<-ifelse(pred>cutp,1,0)
rocpreds<-prediction(result,z)
auc<-performance(rocpreds,"auc")@y.values
accuracy<-length(which(result==z))/length(z)

tn<- sum(z==0 & result==0)  # True Negative
fp<- sum(z==0 & result==1)  # False Positive
tp<- sum(z==1 & result==1)  # True Positive
fn<- sum(z==1 & result==0)  # False Negative

sensitivity<- tp/(tp+fn)
specificity<-tn/(tn+fp)

output<-
c(int,slope,cutp,accuracy,auc,sensitivity,specificity,iteration)
names(output)<-
c("Intercept","Slope","CutPoint","Accuracy","AUC","Sensitivity","Speci
ficity","iteration")
return(output)
}

y<-fitglm(1,1,.25,0)
y
```

```
# Function designed to work with apply function
output<-t(sapply(1:500, function(x) fitglm(x,
sigma=1,tau=.25,beta=0)))

# sapply returns columns as lists.  Unlist before taking mean.
data1<-apply(output,2, function(x) mean(unlist(x)))
data2<-apply(output,2, function(x) var(unlist(x)))
data3<-cbind(data1,data2)

data4<-melt(data3,id='')

data5<-data4[c(-8,-16),]

write.table(data5, 'test1.xls')
```