CHARACTERIZATION OF TECHNICAL UNCERTAINTY IN THE CLASSIFICATION OF CENTROID-BASED MULTIVARIATE ASSAYS

by

Mark Tyler Wilkinson Ebbert

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Biomedical Informatics

The University of Utah

August 2012

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of **Mark Tyler Wilkinson Ebbert**

has been approved by the following supervisory committee members:

Julio Facelli	, Chair	**04/25/2012** Date Approved
Philip Bernard	, Member	**04/25/2012** Date Approved
Kenneth Boucher	, Member	**04/25/2012** Date Approved
Karen Eilbeck	, Member	**04/25/2012** Date Approved
Lewis Frey	, Member	**04/25/2012** Date Approved

and by **Joyce A. Mitchell** , Chair of

the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

Multivariate assays using gene expression as their contributing factors, such as the centroid-based PAM50 Breast Cancer Intrinsic Classifier, are becoming commonly used in assisting treatment decisions in medicine, especially in oncology. Although physicians may rely on these multivariate assays for planning treatment, little is known about the effects on the results of an assay due to the intrinsic error in the laboratory process and measuring its contributing factors. While we expect that classification of samples in proximity to one of the centroids defining the tumor classes will be stable with respect to experimental errors in the gene expression measurements, what happens to the samples not in proximity to a single centroid is unknown. Results reported to the attending physician may be misleading because he or she is receiving no information about the probability for sample misclassification. Given the serious consequences due to ambiguous results in clinical classifications, methods to measure the effects of a multivariate assay's intrinsic errors need to be established and communicated to attending physicians. In this study, a method to characterize the technical uncertainty in the classification of centroid-based multivariate assays, is developed and described, using the PAM50 Breast Cancer Intrinsic Classifier as the model multivariate assay. Furthermore, the described method provides a general and individual classification confidence measurement that advances multivariate assays towards personalized healthcare by providing personalized confidence measurements on the assay's result. Finally, this study explores whether using parametric versus nonparametric distance measurements is most effective when using a single gene expression platform, such as microarray or Real-time, quantitative Polymerase Chain Reaction.

To my wife, Cheri Michelle, and my children and parents for their incredible support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

During the course of my graduate work in the Department of Biomedical Informatics I have received support, encouragement, and guidance from numerous individuals, without whom I would not have been able to complete my degree. My experience has been enlightening and educational, and I would like to specifically acknowledge those to whom I am indebted.

I would first like to acknowledge my committee comprised of Dr. Julio Facelli, Dr. Philip Bernard, Dr. Kenneth Boucher, Dr. Karen Eilbeck, and Dr. Lewis Frey. While seeking committee members I quickly identified each member as one who could augment my education with their wisdom and contribute to the research I hoped to accomplish. Each member has made valuable contributions and it has been an honor to work under their tutelage.

I would also like to acknowledge the other faculty and staff of the Biomedical Informatics Department who have been willing to give their time to help me, at times when it was not convenient for them. Specifically I would like to acknowledge Dr. John Hurdle whose contagious excitement for research has inspired me. Kate Handziuk and Jo Ann Thompson, staff within the department, have also been especially helpful while navigating through graduate school requirements. Of course, there have certainly been other faculty and staff whom have contributed to my education in ways I am not even aware.

One of the most important sources of guidance, inspiration, and desire for further education has come from my parents. Years ago while I was in grade school, my mother wondered whether she would ever get me through high school successfully, since my educational interests were somewhat lacking. Throughout those unsettling years, my parents showed extraordinary patience by continually encouraging me to perform to my best abilities, despite my lackluster performance. Later in life as my educational interests awoke, yet I struggled to develop intellectually, I found strength

in a principle my father taught: "persistence will prevail." A valuable lesson that persistence can overcome any obstacle.

I offer special recognition of my dear wife Cheri, who has demonstrated exceptional love, patience, and devotion as I have struggled to prepare myself – requiring long hours of study. During the most difficult moments, when I questioned my own resolve, she showed complete support and confidence that I would, in fact, succeed. Her faith in me has given me the courage to confront every obstacle.

Finally, I express gratitude to God whom has expanded my mind in times of need and given me everything I have – including the friends and family that have supported me.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Breast cancer has captured the attention of medical professionals, researchers, women, and their loved ones around the world because the disease's cure has been elusive, due to its complexity. Breast cancer's complexity largely stems from its heterogeneity on the levels of molecular alterations, cellular composition, and clinical outcome – making proper diagnosis and treatment difficult [1]. Breast cancer is the most common cancer in women, amounting to nearly 1 in 3 cancer diagnoses, and is the second leading cause of cancer deaths among women, according to the American Cancer Society [2]. Furthermore, more than 230,000 new cases of invasive breast cancer as well as more than 57,000 *in situ* cases were expected to be diagnosed during 2011. Approximately 40,000 breast cancer deaths were also expected during 2011 [2]. Clearly, breast cancer dramatically changes the lives of millions of individuals annually, in the United States alone, when also considering the friends and loved ones for each newfound patient. However, researchers and physicians are making progress in efforts to eradicate breast cancer. For example, the breast cancer mortality rate decreased by approximately 24% to 37% between 1990 and 2007, which was attributed to early detection by screening and adjuvant treatment [2,3]. Such a dramatic decrease in breast cancer mortality is encouraging, but certainly not sufficient. Further efforts to understand the biological nature of breast cancer will ameliorate treatment for all cancers at any stage.

In an effort to better understand the biological nature of breast cancer, researchers and physicians have worked together to develop several multivariate assays (MVA) designed to elucidate the biological nature of breast tumors. MVAs are assays that exploit multiple biological measurements, referred to as contributing factors, to ascertain a clinical result. Example MVAs include the PAM50 Breast Cancer Intrinsic

Classifier, Oncotype DX®, BreastOncPx™, and MammaPrint® – each of which are both prognostic and predictive of a tumor's drug and chemotherapy sensitivity [1, 4–13]. Although these MVAs have proven useful, little is known about the effects of intrinsic technical uncertainty on the MVA's results. Each of the aforementioned MVAs measure expression of several genes, ranging from 14 to 70, making them highly complex. In fact, each gene measured in the MVA, although necessary to understand the biological nature of the tumor, is a separate dimension that introduces its own natural variation, or error, which may make results more unpredictable. As MVAs become more commonplace in medicine, a method to characterize the intrinsic technical uncertainty of such complex tests generally, as well as how to measure the confidence of individual results, will be essential.

Technical uncertainty may be separated into intrinsic error from measuring devices and *in vitro* biological processes, as well as intrinsic and explicit classification uncertainties, among others. There are also extrinsic factors such as local climate and human interactions that will affect MVAs. Essentially, technical uncertainty involves most aspects from the time a sample is received by the reference laboratory until a clinical report is generated. To be widely applicable, a method capable of characterizing intrinsic technical uncertainty must be scalable and generalizable. By developing a scalable and generalizable method to characterize intrinsic technical uncertainty, researchers will be able to assess and validate the technical aspect of current and future MVAs, as well as minimize the effects of technical uncertainty on the MVA's final result – making patient diagnosis and treatment more accurate and predictable.

1.2 Main Objectives

Several MVAs are already used in medicine and, in general, MVAs are likely become a standard of care for many complex diseases; however, three main shortcomings regarding MVAs to address, which are the primary objectives of this research, are as follows:

1. There is currently no standard method to characterize the intrinsic technical uncertainty of MVAs

2. There is currently no method to measure the confidence of individual results

3. Little is known regarding the effectiveness of using a parametric measurement within a single platform (e.g., microarray or RT-qPCR) versus using a nonparametric measurement.

A standard method to assess and validate the technical aspect of MVAs is essential because it will allow researchers to characterize and minimize the effects of intrinsic technical uncertainty (error) on an MVA's final result. Little is known about the effects of intrinsic technical error in measuring an MVA's contributing factors on the MVA's results. Likewise, little is known about how to measure the effects. Given the complexity of MVAs, results may be unpredictable, or easily corrupted by minor changes or mistakes in the process. For example, if a mistake is made in the process that causes a single contributing factor to be mismeasured, there is currently no method to assess the potential ramifications. Furthermore, there is no method to assess how much variation, or error, can exist within the process without compromising the results. Developing a generalizable method to characterize the intrinsic technical uncertainty is the first objective of this research.

A method to measure the confidence of individual results will advance MVAs toward personalized healthcare by providing personalized confidence measurements on the MVA's result. Clinical assays routinely report a general summary statistic measuring how well the assay performed on an independent test set when compared to the results of a reference standard method; however, such summary statistics do not provide personalized information on the confidence that the result for a specific sample was accurate. Since summary statistics, by nature, are meant to represent an entire distribution of results, a physician has no choice but to assume that all results provided by an MVA have equal confidence to the confidence portrayed in the summary statistic. There will undoubtedly be samples whose confidence is poor, but the physician will be uninformed that he or she should give special attention to the situation. Developing a method to measure the confidence of individual results is the second objective of this research.

An essential aspect of developing and assessing centroid-based MVAs (cbMVA), specifically, is identifying the best distance metric to determine classifications, however, little is known regarding when to use parametric versus nonparametric measurements. Given the limitless variations possible when designing cbMVAs, defining whether a parametric or nonparametric measurement is most appropriate is difficult at best. Nevertheless, a specific situation of interest is which measurement type is most appropriate when using a single gene-expression platform such as microarray or Real-time, quantitative PCR (RT-qPCR). The PAM50 clinical assay uses RT-qPCR exclusively and will be used to explore the potential ramifications. Since there are several possible measurements within both the parametric and nonparametric categories, two specific measurements of interest will be considered: euclidean, which is parametric, and Spearman's Rho (ρ), which is nonparametric. Exploring the effects of using euclidean versus ρ as a distance metric when using RT-qPCR exclusively is the third and final objective of this research.

CHAPTER 2

BACKGROUND

2.1 Uncertainty in Centroid-based Multivariate Assays

MVAs using gene expression as their contributing factors are becoming commonplace in assisting treatment decisions in medicine, especially in oncology. Examples of MVAs available for planning breast cancer treatment include the 55-gene subtype classifier (PAM50 Breast Cancer Intrinsic Classifier) [1], the 21-gene prognosis assay (Oncotype DX®) [14], the 14-gene prognosis assay (BreastOncPx™) [15] and the 70-gene prognosis assay (MammaPrint®) [16]. Although physicians may rely on these MVAs for planning treatment, little is known about the effects of intrinsic technical error in measuring an MVA's contributing factors on the MVA's results – in this case, all laboratory steps required for preparing a sample (post RNA extraction), preparing the assay, and the instrumental errors for measuring gene expression. While we expect that classification of samples in proximity to one of the centroids defining the tumor classes, referred to as archetypal samples here, will be stable with respect to experimental errors in the gene expression measurements, it is unknown what happens to the samples not in proximity to a single centroid. For example, if a sample lies in a "gray" area where the intrinsic errors in the gene expression measurements may result in a change of its classification each time the sample is run, the results reported to the attending physician may be misleading because he or she is getting the results from only one measurement and no information about the probability for sample misclassification. Given the serious consequences due to ambiguous results in clinical classifications, methods to measure the effects of an MVA's intrinsic errors need to be established and communicated to attending physicians.

The complexity of MVAs demonstrates the challenge of identifying and understanding error sources from the moment a sample is received by the reference labora-

tory until a clinical report is generated. In short, each contributing factor measured within an MVA is a new dimension with its own associated measurement error. Further error sources to be considered include heterogeneity (e.g., heterogeneity due to molecular and cellular composition) and sample preparation, as well as technical variability, which may be separated into error from measuring devices and *in vitro* biological processes, as well as intrinsic and explicit classification uncertainties. Potential classification uncertainties include differing metrics to define class boundaries, such as euclidean distance versus Spearman's Rho (ρ), as well as differing algorithms to determine the classification. The focus of this research was to characterize technical uncertainty within MVAs, in general, and more specifically how to estimate and measure the effect that intrinsic gene expression measurement errors, including those associated with sample and assay preparation, have on final MVA results overall (i.e., across large data sets), as well as for individual samples.

2.2 PAM50 Breast Cancer Intrinsic Classifier

In 2009 Parker et al. proposed a 50-gene expression signature, named the PAM50, as a method to standardize breast cancer subtype classification through gene expression profiling – in contrast to the conventional immunohistochemistry (IHC) and fluorescence in situ hybridization (FISH) methods. Both IHC and FISH have been criticized as being subjective or insufficient, or both, due to differing techniques, visual interpretation, and other inconsistencies [1, 17–24]. Conversely, quantitative values allow for more objective analysis, such as those produced through gene expression profiling techniques like microarray and RT-qPCR [1]. As such, the PAM50 is well positioned to provide objective, reproducible subtype classification for breast cancer, which is commonly classified into one of four recognized biological subtypes known as Luminal A, Luminal B, HER2-enriched, and Basal-like [1, 25–28].

The PAM50, later released as a clinical assay in 2011 at ARUP Laboratories as the PAM50 Breast Cancer Intrinsic Classifier (PAM50 assay), is a complex MVA based on gene expression for several genes. PAM50 measures the expression level of 55 genes (50 classifier genes and 5 housekeepers) creating a "signature" that is compared, using Spearman's Rho (ρ) as a distance metric, to each of five centroids [1, 29] representing

the Luminal A, Luminal B, HER2-enriched and Basal-like subtypes [1,26,27], as well as Normal-like tissue. The tumor sample is then classified as the subtype to which it is most similar by ρ.

Although the PAM50 is young, numerous studies have already shown the PAM50 subtype classification is both prognostic as well as predictive of a tumor's drug and chemotherapy sensitivity [1, 4–6]. Moreover, a recent study has also shown that the PAM50 may better identify low-risk ER+ tumors, when compared to Oncotype DX®, demonstrating PAM50's potential value in breast cancer treatment [30].

Despite all of the benefits the PAM50 and other prognostic MVAs provide, little is known about how to assess the overall and individual reproducibility of an MVA's final result, given the intrinsic technical uncertainty within the MVA system (e.g., instrument measurement variability, RT-qPCR variability, etc.). As such, it is essential to develop a method to characterize the uncertainty within a complex MVA and its effects on the MVA's final result.

2.3 Distance Metrics – Spearman's Rho (ρ) vs. Euclidean

Distance metrics are a key element of interest in the characterization of error within cbMVAs, specifically. Numerous distance metrics could be used and were explored during development of the PAM50; however, Spearman's Rho (ρ) was identified as the ideal distance metric to maintain consistency across gene expression platforms (e.g., RT-qPCR and microarray) [1]. Spearman's nonparametric approach involving a value's rank, when compared to its sister values, rather than the raw value itself makes the metric more robust across platforms because nonparametric statistics are more capable of assessing data that do not belong to any particular distribution (or different distributions) [31]. However, nonparametric statistics have several disadvantages such as being less specific (i.e., more generalized) because information is lost by using ranks instead of raw values, and nonparametric tests have less power to reject a false null hypothesis [31, 32]. Given clinical MVAs are likely to use a single platform (e.g., RT-qPCR), the benefits of a nonparametric test like Spearman's Rho (ρ) come into question. Parametric measurements, on the other hand, such as Euclidean distance, although not ideal across platforms, may be more accurate and consistent within a

single platform, and thus be a better choice for the clinical assays like the clinical PAM50 assay.

2.4 General and Individual Classification Confidence

Perhaps the most important element of characterizing intrinsic technical uncertainty in the classification of MVAs is to measure uncertainty's effect on the MVA's final results – in other words, measuring the overall confidence of the MVA's final results. However, this research also seeks to take an MVA's accountability one step further and provide a confidence measurement on the results of individual patient samples, based on the premise that knowing the overall reproducibility of an MVA provides little information for the confidence of a given sample's result. Essentially, the overall confidence of any assay is an average confidence, which suggests there is an unknown distribution of confidence values for individual samples. For example, clinical assays commonly report sensitivity and specificity of the assay, which represents the assay's overall ability to distinguish between classes; however, sensitivity and specificity are merely an overall measurement of confidence from a test set where truth is "known" (assumed) based on some reference standard. Such information provides little information for a sample lacking knowledge of the reference standard, except to say the assay "overall" provides sensitivity x and specificity y. A more specific measurement is essential to providing personalized confidence measurements.

The PAM50 uses Spearman's Rho (ρ) as a distance metric to determine the subtype of a tumor. Several statistical tests exist for determining statistical significance of ρ [33–35], but each test is insufficient for one of two reasons: (1) the test is not specific for individual samples; and (2) the test does not relate multiple ρ values. Statistical tests, such as those in the first case, are generally tailored for sample means rather than individual measurements and are not suited for calculating a statistical measurement for an individual breast tumor sample classification. Regarding the second case, Spearman's Rho (ρ) is capable of determining whether any two variables (e.g., tumors) are significantly correlated by ρ by testing whether ρ is significantly different from 0; however, this test is incapable of relating multiple ρ values. In other words, when the tumor is compared to each of the five centroids, the tumor may

be significantly correlated to more than one centroid, especially since the expression profiles for all centroids and the test sample are based on breast tissue. In fact, preliminary analyses suggest that $> 99\%$ of all test samples are significantly correlated with more than one centroid, and $> 85\%$ are significantly correlated to all five. Currently, there is no way to determine whether one classification is statistically the "best" classification. A new method to measure confidence of individual samples is clearly needed, and if medicine is, indeed, to become personalized, then clinical assays must provide personalized results, including personalized confidence measurements.

2.5 The Monte Carlo Method

A method that is capable of characterizing the technical uncertainty of MVAs must be thorough, scalable, and generalizable because of the complexity of MVAs – an ideal situation to employ the Monte Carlo method [36–39]. The Monte Carlo method was developed by Stanislaw (Stan) Ulam, John von Neumann, Nicholas Metropolis, and others at the Los Alamos Laboratory in the 1940's while trying to characterize neutron diffusion in fissionable materials, which is a complex, multidimensional problem [36–39] similar to characterizing technical uncertainty in MVAs. The Monte Carlo method is particularly amenable to complex, multidimensional problems because it uses random sampling from predefined statistical distributions for each dimension (contributing factor) to perform what have become known as "mathematical experiments" [38]. Furthermore, the upper limits on performing the "mathematical experiments" are determined only by available time and computational resources – both of which are readily available to researchers today. Allowing each contributing factor to be randomly sampled from its own distribution makes the Monte Carlo method highly scalable and generalizable, while being able to perform a large number of "mathematical experiments" allows the method to be highly thorough.

Ulam's Monte Carlo method has been used to solve numerous problems empirically. Ulam first conceived the idea in 1946 while playing the card game Solitaire, as he later stated in 1983:

> The first thoughts and attempts I made to practice [the Monte Carlo method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires [sic]. The

> question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays. This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later... [in 1946, I] described the idea to John von Neumann and we began to plan actual calculations [37].

As Ulam stated, he immediately applied the idea to understanding neutron diffusion, and more generally how to apply the method to problems pertaining to differential equations. In fact, the Monte Carlo method has since been shown to be useful in problems where conventional analytics are not feasible [38,40,41]. After some lengthy preparation, Ulam et al. later computed nine problems specific to material configurations, neutron distributions, and running times [38]. The Monte Carlo method has since been used to solve problems in fluids, thermodynamics, neuroscience, and other disciplines, demonstrating how the method is amenable to complex, multidimensional problems [38, 42–45].

CHAPTER 3

METHODS

A comprehensive experimental study to estimate the effect that the intrinsic gene expression measurement errors have on the classification of tumors and gene-score classifications requires, in principle, repeated testing of a significant number of samples from each subtype and thorough analysis of the misclassifications observed. Such a comprehensive approach is unfeasible in terms of cost and sample availability. Here we have adopted a hybrid approach in which we perform repeated experimental measurements on one sample from each subtype (i.e., Luminal A, Luminal B, HER2-enriched and Basal-like) to determine the experimental variability of the measured gene expression for each of the 50 genes included in the PAM50 assay. Using this experimental information we proceed to generate a Gaussian error distribution that can be used to generate multiple data sets by way of Monte Carlo simulations. These simulations impose random errors, given by the Gaussian distribution, on the set of experimental measured samples. Monte Carlo simulations are well suited for this hybrid analysis because there is extensive literature for using this approach to estimate errors in high-dimensional problems, such as fluids and thermodynamics [38, 42, 43], where conventional analytics are not feasible [38, 40, 41]. The simulated data sets are then classified by the standard PAM50 algorithm, and the misclassifications encountered in the synthetic data sets are used as a proxy for the PAM50's misclassification rate based on the assay's intrinsic error.

Specifically, in this study we followed five major steps: (1) collect and prepare four archetypal samples representative of each cancer subtype; (2) characterize the intrinsic error for each gene's expression values in the assay by making 12 measurements for each gene's expression on each archetypal sample and determine the distribution type that best models the experimental errors; (3) using Monte Carlo simulations

generate a sufficient number of simulated test samples, based on a defined confidence interval width, by imposing the errors generated using the distribution from (2) onto the archetypal samples; (4) determine the effect of the variability imposed in the simulated samples on their classification; and (5) repeat steps (3) and (4) on an independent set of samples from the GEICAM 9906 clinical trial (GEICAM) [46].

3.1 Archetypal Sample Collection and Preparation

In order to characterize the error in gene expression measurements, four archetypal samples representative of each cancer subtype with sufficient genetic material were constructed – since most single samples do not have enough genetic material to be tested more than twice. Cell lines representative of Basal-like (ME16C) and Luminal B (MCF7) subtypes were grown in the Reagent Lab at ARUP Laboratories. Luminal A and HER2-enriched subtypes were not readily available as a cell lines. As such, 20 patient tumor samples previously identified as archetypal Luminal A (10 samples) and HER2-enriched (10 samples), based on PAM50 gene scores and classification, were collected under IRB approved protocols at the University of Utah to be combined and treated as single tumor samples.

RNA was extracted from tumor-enriched areas of formalin-fixed, paraffin-embedded (FFPE) tissue blocks, containing more than 70% tumor cells, as determined during review by a board-certified pathologist. Samples were deparaffinized using Citrus Clearing Solvent (Richard-Allen Scientific, Kalamazoo, MI, `http://www.thermofisher.com`) followed by dehydration in absolute ethanol. RNA extraction was completed on a Biomek NX Laboratory Automation Workstation (Beckman Coulter, Beverly, MA, `http://www.beckmancoulter.com`) using the AgenCourt FormaPure Kit (Beckman Coulter, Beverly, MA, `http://www.beckmancoulter.com`) according to the manufacturer's instructions and including a DNase I step. RNA quantification was done on a Paradigm Detection Platform (Beckman Coulter, Beverly, MA, `http://www.beckmancoulter.com`) using the Quant-iT RiboGreen Assay Kit (Invitrogen, Carlsbad, CA, `http://www.invitrogen.com`). cDNA synthesis was performed on the Biomek FX Laboratory Automation Workstation (Beckman Coulter, Beverly, MA, `http://www.beckmancoulter.com`) using 600 ng of RNA,

uracil containing dNTPs (Invitrogen, Carlsbad, CA, http://www.invitrogen.com), random primers (Invitrogen, Carlsbad, CA, http://www.invitrogen.com), gene-specific, downstream PCR primers (Idaho Technology, Salt Lake City, UT, http://www.idahotech.com), and SuperScript III Reverse Transcriptase (Invitrogen, Carlsbad, CA, http://www.invitrogen.com).

Each 5 µL reaction contained 1X LightCycler 480 SYBR Green I Master Mix (Roche Applied Sciences, Indianapolis, IN, http://www.roche-applied-science.com) and 1.67 ng cDNA were added to the experimental sample wells. Sample cDNA was incubated with LightCycler Uracil-DNA Glycosylase (Roche Applied Sciences, Indianapolis, IN, http://www.roche-applied-science.com) at 40 °C for 10 min and inactivated at 95 °C for 10 min prior to performing RT-qPCR. RT-qPCR was performed on the LightCycler (LC) 480 (Roche Applied Sciences, Indianapolis, IN, http://www.roche-applied-science.com) as follows: 45 cycles at 95 °C for 4 sec, 58 °C for 6 sec and 72 °C for 6 sec. To assure target specificity, RT-qPCR was followed by a melting curve analysis: 95 °C for 15 sec, 65 °C for 1 min followed by raising the temperature to 99 °C while taking 10 fluorescence acquisitions/°C. We then classified the RT-qPCR data from each run. One run from the Luminal A sample failed quality control and was not included in further analysis.

3.2 Error Characterization

In order to estimate the intrinsic experimental error in gene expression measurements within our laboratory, we performed 12 measurements for each gene within each archetypal sample. Specifically, each of the four archetypal samples was separated into, and treated as 12 individual samples (after extraction), and measured by RT-qPCR on the Roche LightCycler (LC) 480 (Roche Applied Sciences, Indianapolis, IN, http://www.roche-applied-science.com). The error distribution function type could not be estimated using only the 12 measurements for each gene within a given sample subtype, therefore to determine the error distribution function type for each gene, all four sample subtypes were median-centered by gene and combined, giving 47 data points per gene, since one of the archetypal Luminal A samples failed quality control. As depicted in Figure 3.1, the resultant error distributions for each gene

can be reasonably approximated by Gaussian distributions. Therefore 200 Gaussian distributions were generated, one per gene within each archetypal sample, using the mean and standard deviation of the 12 data points within the given gene and archetypal sample. Note that only the 12 data points available for each gene were used to determine the mean and standard deviation because the mean and standard deviation must be specific to the gene and subtype, whereas all 47 data points per gene were necessary to form a recognizable distribution. Gaussian distributions were generated using the "rnorm" function within The R Statistical Package (R) [47].

3.3 Sample Size Justification for Archetypal Sample Monte Carlo Simulation

Before performing the Monte Carlo simulations an analysis to justify sample size was performed to ensure sufficient confidence for the analysis when using our target of 100,000 simulated samples. We calculated the 95% confidence interval width around the percentage of correct classifications using 100,000 simulated samples for each archetypal sample. For a dichotomous variable (i.e., misclassified or not), a confidence interval width (W) can be calculated (Equation 3.1) given an alpha

$$W = \pm Z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{3.1}$$

level ($Z_{0.05} = 1.96$), expected proportion of misclassifications ($\hat{p} = 0.02$) and the sample size, or in this case, the number of simulations (n = 100,000). The calculated confidence interval width is ± 0.00087 for each simulation of 100,000 samples, which is an acceptable value.

3.4 Monte Carlo Simulation Using Archetypal Samples

Monte Carlo simulations were performed using the mean (μ) value and standard deviation (σ) for the expression of each gene within the archetypal samples, as described above. This procedure created a total of 200 independent distributions, i.e., 50 Gaussian distributions (one for each gene) for each of the four archetypal samples. For example, the mean expression value (μ) for ACTR3B from the 12 Luminal B values was 1.94 and the corresponding standard deviation (σ) was 0.085. Therefore

the ACTR3B expression value for each of the 100,000 simulated Luminal B samples was randomly selected from a Gaussian distribution centered on a mean (μ) of 1.94 with a standard deviation (σ) of 0.085. Randomly selecting a value from the Gaussian distribution, as described, does not assume gene expression values are independent of one another, rather the method assumes that the measurement error for each gene is independent. Specifically, since each Gaussian distribution is centered on the gene's mean expression value for the given sample, any genes within the sample that are generally upregulated will generally be upregulated in the simulated samples, while allowing the error to deviate independently. The 100,000 simulated samples for each archetypal sample were classified using the standard PAM50 process. The effect of intrinsic gene expression measurement error on the tumor classification was assessed by determining the percentage of simulated samples that were classified identically to the original sample. This value provides an estimate of the reproducibility of the results for archetypal samples.

3.5 Evaluation Using the GEICAM Independent Test Set

Testing archetypal samples is valuable for determining how the PAM50 assay will perform under ideal circumstances, but these results may not be informative when the samples are not as well characterized as the archetypal samples. Thus, the method described above for the archetypal samples was adapted and applied to the larger and more diverse set of independent samples from the GEICAM 9906 clinical trial. A total of 911 breast tumors collected by the GEICAM group for the GEICAM 9906 clinical trial were run and classified by the PAM50. Tumor samples were prepared following the same methods described above for the archetypal samples and those with insufficient tumor content to be classified were excluded from further analyses.

3.6 Monte Carlo Simulation using GEICAM Samples

As depicted in Figure 3.2, the data from the multiple measurements in the archetypal samples show that standard deviation depends on the relative average gene expression value and on the sample subtype. To understand the sample subtype

dependence one should understand that the expression values of the genes defining the expression pattern for each of the cancer subtype are quite different, e.g., the Luminal A subtype expresses all 50 genes at a level that is more easily quantified, producing a lower standard deviation; however, the HER2-enriched subtype expresses some genes at lower levels such that they are less easily quantified, producing a higher standard deviation. Therefore as depicted in Figure 3.2, the relative errors in the gene expression measurements in a Luminal A sample are smaller than those in a HER2-enriched sample. Accordingly our methods to produce simulated samples have to be modified to take into account these dependencies when applied to a set of nonarchetypal samples. Using locally weighted scatter plot smoothing (loess), based on the PAM50's characterized error functions depicted in Figure 3.2, we developed error distributions that can be used to impose the error on individual test samples dynamically. The loess model was fit using the R function "loess" (span = 0.75, degree = 1, surface = "direct" and family = "symmetric") and graphed using "panel.smoother" from the R lattice graphics package. A 95% confidence interval for the fitted line was also calculated to test "best-case" (lower limit of the 95% confidence interval), "worst-case" (upper limit of the 95% confidence interval) and "average-case" (the fitted line) scenarios for subtype reproducibility. The "worst-case" scenario is considered as such because it uses the highest estimated standard deviations, or error. Once the loess models were developed, we used these models to predict the standard deviation (σ) to generate Gaussian distributions for the Monte Carlo simulation. Specifically, given the test sample's original subtype and the expression value for a given gene, we used the loess model to predict the standard deviation (σ), or error, to be used in the Gaussian distribution for said gene and sample. The expression value was used as the mean of the Gaussian distribution. We repeated this process for all genes within each sample.

The model described above was used to generate random variants of the 847 GEICAM samples remaining after excluding those that could not be classified. The subtype classification reproducibility for each GEICAM sample was tested by generating 100,000 simulated samples using Monte Carlo simulations for each of the error models considered above, i.e., "best-," "average-" and "worst-case," for a total of

300,000 simulated samples per GEICAM sample. Based on the same sample justification analysis used for the archetypal samples, the calculated confidence interval width is 0.00087 for each simulation of 100,000 samples, which is an acceptable value.

3.7 Error Effect on PAM50 Results for GEICAM Samples

After simulating 300,000 samples for each GEICAM sample based on the error models described above, subtype reproducibility for the individual samples was summarized for each tumor subtype (based on the original sample's subtype) using two statistics: (1) the total percentage of simulated samples that did (or did not) change subtypes with respect to their parent sample and (2) the proportion of simulated samples, corresponding to a single original sample, that were classified (or misclassified) identically to the parent sample. These data serve as an estimate of PAM50's misclassification rate based on intrinsic error within the assay when samples are not in proximity to the PAM50 centroids and represents the error effect on PAM50 results.

3.8 Spearman's Rho (ρ) vs. Euclidean Distance

Simulations using Spearman's Rho (ρ) produced results to measure the effect of technical uncertainty on the clinical PAM50 assay's results, but does not address whether the PAM50 would benefit by using a parametric measurement when using only one gene expression platform, as is the case with the clinical PAM50 assay. We compared Spearman's Rho (ρ) to euclidean distance through two methods: (1) comparing the continuity of PAM50 subtype classifications in hierarchical clusters when using ρ as a distance metric versus using euclidean distance; and (2) comparing overall subtype reproducibility (error effect on PAM50 results) between when ρ is used versus euclidean distance, as described. In this case 1,000 simulated samples were generated for each parent sample. Statistical significance between using ρ and euclidean as the distance metric was tested for overall significance (not distinguishing between subtypes) and for each subtype using the Wilcoxon signed-rank test. The R function "wilcox.test" (paired=TRUE and alternative="less") was used.

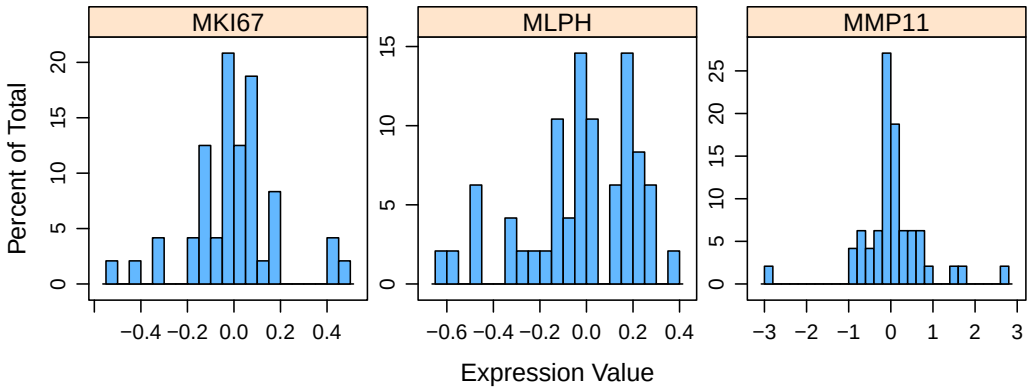

Figure 3.1: Error distributions of the expression values for genes MKI67, MLPH, and MMP11. These error distributions are representative of those for each of the 50 classifier genes and were used to determine the Gaussian distribution best represented the overall error distribution.

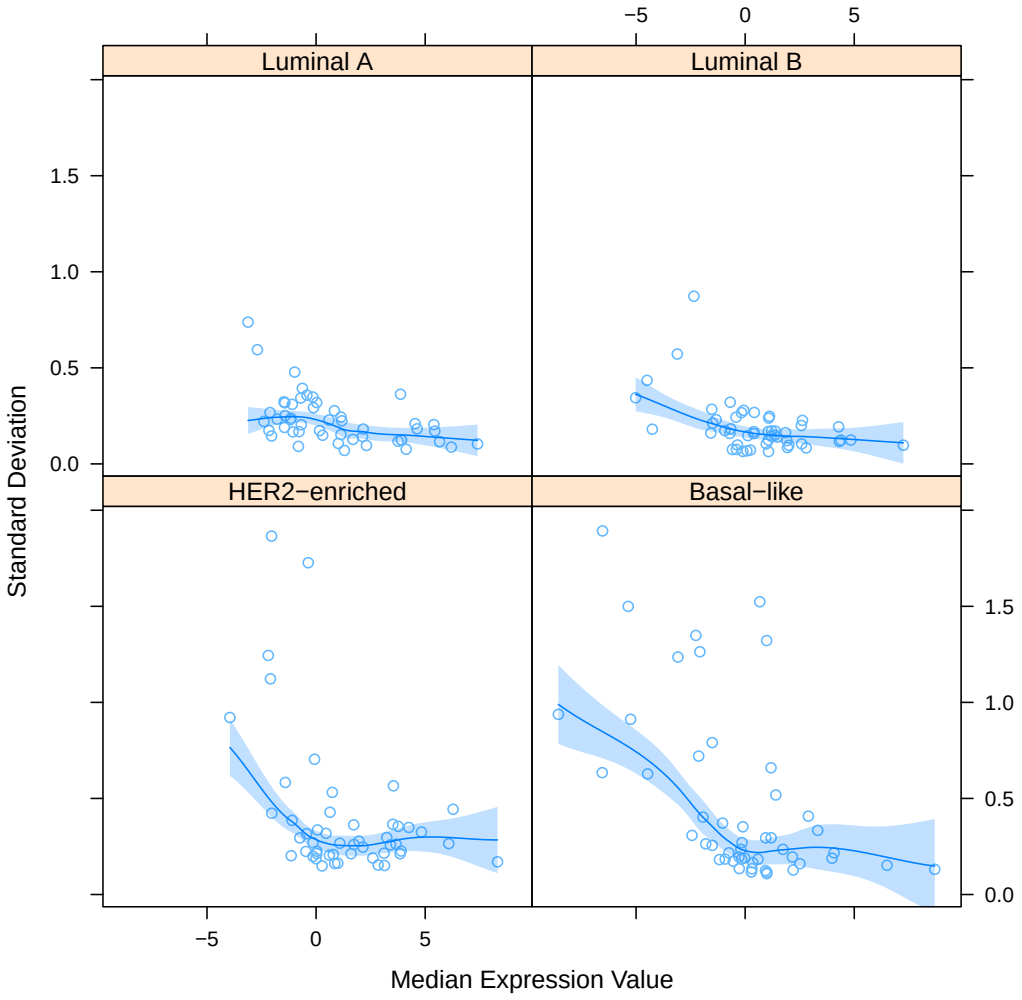

Figure 3.2: Standard deviation as function of expression values and tumor subtype and loess model predicting standard deviation based on subtype and expression.

CHAPTER 4

RESULTS

4.1 Archetypal Samples

All simulated samples derived from the archetypal samples were classified as the same subtype as the parent sample, indicating that the PAM50 subtype classification for samples with characteristics close to the PAM50 centroids is highly reproducible and resilient to experimental errors in gene expression measurements. Although this positive result is encouraging for PAM50 classifications, it is not generalizable beyond the four archetypal samples.

4.2 GEICAM 9906 Samples

Tables 4.1, 4.2, and 4.3 present the results for the classification of all samples produced by Monte Carlo simulation using the "best-" (Table 4.1), "average-" (Table 4.2) and "worst-case" (Table 4.3) scenarios for the error models described above. These scenarios correspond to using the lower values of the 95% confidence interval for the predicted error-model line, the actual predicted line and the upper values of the 95% confidence interval, respectively. The upper values of the 95% confidence interval are representative of the "worst-case" scenario because a greater amount of error is introduced into the simulation.

Results for the GEICAM sample simulations suggest that all subtype classes are highly reproducible. When considering the three different error models, the most reproducible subtype among all GEICAM samples was always the Basal-like for which 98.47% (average-case) of the simulated samples did not change classification, followed by Luminal B (96.63%, average-case), Luminal A (95.46%, average-case) and HER2-enriched (90.07%, average-case). The differences in reproducibility between "best-" and "worst-case" ("best-case" percentages minus "worst-case" percentages) in percentage points are 0.89, 1.31, 1.31, and 4.35 for Basal-like, Luminal B, Luminal

A, and HER2-enriched, respectively. Consequently, the selection of the error model is not a determining factor in assessing the robustness of the classification due to experimental errors in gene expression measurement.

Figure 4.1 presents the histogram depicting the percentage of simulated samples, corresponding to a single parent sample, that change. Based on the "average-case" simulations and analyzing the classification of the 100,000 simulated samples generated for each sample we found that 80% (68 of 85) of Basal-like samples never change subtype during the Monte Carlo simulation, 69% (186 of 270) of Luminal B samples never change, 59% (178 of 303) of Luminal A samples never change and 22% (42 of 189) of HER2-enriched samples never change subtype. These values correspond approximately to the first "bucket" in the histogram. While the histograms confirm the results from Tables 4.1, 4.2, and 4.3 on the general robustness of the PAM50 classification, they also reveal that there are a nonnegligible number of samples for which a large number of simulated samples change classification. For instance the most variable sample for Basal-like, Luminal B, Luminal A and HER2-enriched changed in 42%, 86%, 67% and 75% of the simulated samples, respectively. As an example, the sample named "GEICAM_09-02639_UU" was originally classified as HER2-enriched, but 38.7% of its simulated samples were classified as something else. Specifically, the simulated samples were classified as HER2-enriched 61.3% of the time, Luminal A 21.4% of the time and Normal-like 17.3% of the time. These percentages could be translated into probabilities that can be reported to clinicians using a scorecard like the one depicted in Figure 4.2.

4.3 Distance Metrics

Figure 4.3 presents a hierarchical cluster of all samples from the GEICAM 9906 clinical trial comparing the continuity of PAM50 subtype classifications when using Spearman's Rho (ρ) as a distance metric versus using euclidean distance. Subtype classifications are colored according to Luminal A (dark blue), Luminal B (light blue), HER2-enriched (pink), Basal-like (red), and Normal-like (green); and are separated by euclidean (top row) and Spearman's Rho (denoted "Clinical"). The continuity of the euclidean classifications is noticeably superior within the HER2-enriched and

Luminal B groups, and moderately better within the Basal and Luminal A groups.

Table 4.4 presents the results for the classification of all samples produced by Monte Carlo simulation using the "average-case" scenario for the error models described, with the exception of using of using euclidean as the distance metric. Results suggest that euclidean provides a more reproducible classification for each PAM50 subtype when using data from within one gene expression platform (RT-qPCR in this case), which may also be generalizable to other cbMVAs. As with Spearman's Rho (ρ), Basal-like was most reproducible with 98.81% of the simulated samples not changing classification, followed by Luminal B (97.76%), Luminal A (97.45%), and HER2-enriched (93.79%) – though the gap between Luminal A and Luminal B is markedly smaller. The differences in reproducibility between the "average-case" for euclidean and Spearman's Rho (ρ) (euclidean percentages minus Spearman's Rho percentages) in percentage points are 0.34, 1.13, 1.99, and 3.72 for Basal-like, Luminal B, Luminal A, and HER2-enriched, respectively.

Furthermore, Figure 4.4, in contrast to Figure 4.1, presents the histogram depicting the percentage of simulated samples, corresponding to a single parent sample, that change when using euclidean as the distance metric. There is a markedly strong left-shift towards zero, including a dramatic decrease in the maximum value within each subtype. Results from the Wilcoxon signed-rank test demonstrate a statistically significant increase for the overall change ($p < 6.87 \times 10^{-8}$), i.e., treating all subtypes together. Also, when treating subtypes independently, Luminal A ($p < 0.002$), Luminal B ($p < 2.90 \times 10^{-6}$), and Basal-like ($p < 0.049$) were statistically significant, however, HER2-enriched ($p < 0.106$) was not.

Table 4.1: Original GEICAM Sample Subtype vs Simulated Sample Subtype (as Percentage) - Best-Case Scenario

Original Subtype (simulated samples in thousands)	**Classified as Subtype (%)**				
	Luminal A	**Luminal B**	**HER2-enriched**	**Basal-like**	**Normal-like**
Luminal A (30300)	**96.12**	1.67	1.55	0	0.66
Luminal B (27000)	1.93	**97.27**	0.80	0	0
HER2-enriched (18900)	3.80	2.99	**92.25**	0.24	0.73
Basal-like (8500)	0	0	1.14	**98.86**	0

Table 4.2: Original GEICAM Sample Subtype vs Simulated Sample Subtype (as Percentage) - Average-Case Scenario

Original Subtype (simulated samples in thousands)	**Classified as Subtype (%)**				
	Luminal A	**Luminal B**	**HER2-enriched**	**Basal-like**	**Normal-like**
Luminal A (30300)	**95.46**	1.98	1.74	0	0.81
Luminal B (27000)	2.34	**96.63**	1.03	0	0
HER2-enriched (18900)	4.69	3.76	**90.07**	0.45	1.03
Basal-like (8500)	0	0	1.51	**98.47**	0.02

Table 4.3: Original GEICAM Sample Subtype vs Simulated Sample Subtype (as Percentage) - Worst-Case Scenario

Original Subtype (simulated samples in thousands)	**Classified as Subtype (%)**				
	Luminal A	**Luminal B**	**HER2-enriched**	**Basal-like**	**Normal-like**
Luminal A (30300)	**94.81**	2.29	1.93	0	0.97
Luminal B (27000)	2.75	**95.96**	1.29	0	0
HER2-enriched (18900)	5.65	4.45	**87.90**	0.67	1.33
Basal-like (8500)	0	0.01	1.98	**97.97**	0.04

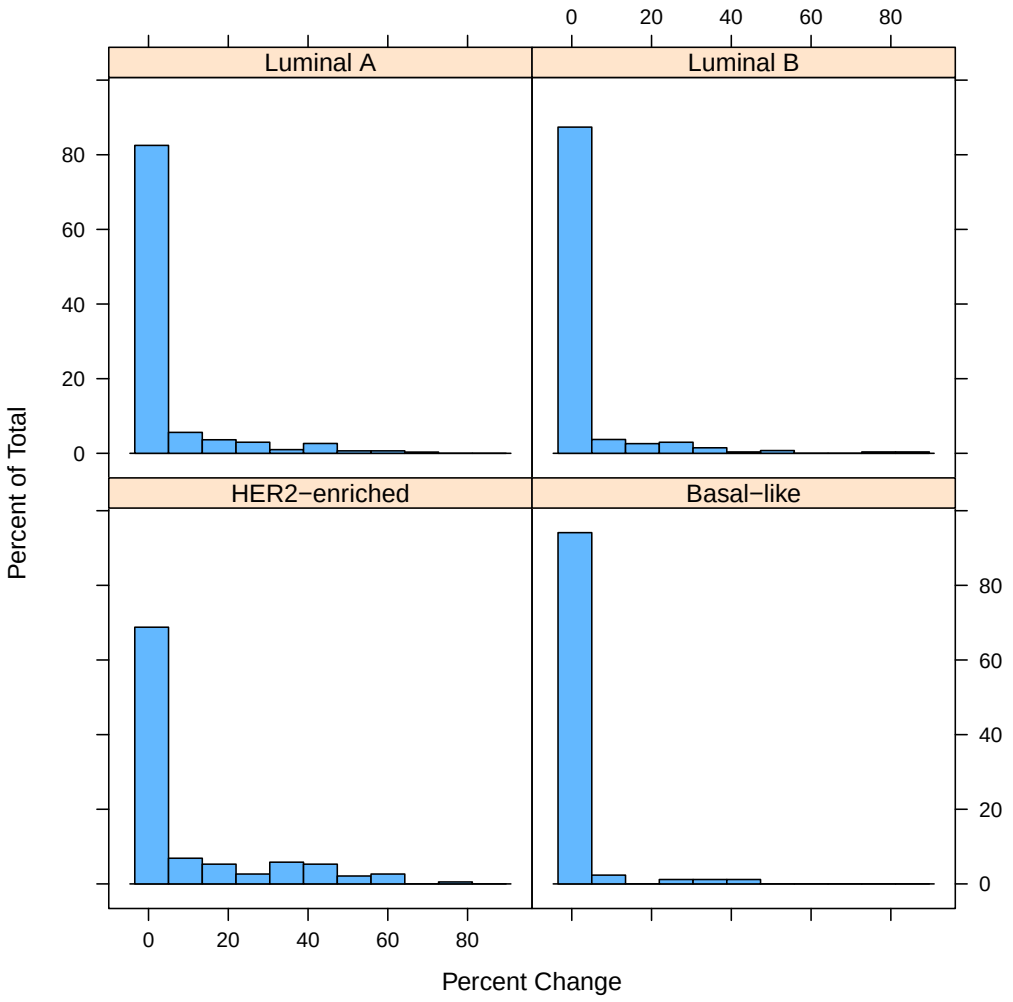

Figure 4.1: Distribution of subtype reproducibility for replicas of individual GE-ICAM samples. Each of the four histograms shows on the y-axis the percentage of parent samples for which there are any number of replicas that changed classification.

Sample ID: GEICAM_09-02639_UU		**Sample ID:** GEICAM_09-02594_UU		**Sample ID:** GEICAM_09-02588_UU	
PAM50 Classification: HER2-enriched		**PAM50 Classification:** Luminal A		**PAM50 Classification:** Luminal B	
Probability[1] to be classified as:		**Probability[1] to be classified as:**		**Probability[1] to be classified as:**	
Luminal A	21.4%	**Luminal A**	94.3%	**Luminal A**	0.0%
Luminal B	0.0%	**Luminal B**	0.0%	**Luminal B**	100.0%
HER2-enriched	61.3%	**HER2-enriched**	0.0%	**HER2-enriched**	0.0%
Basal-like	0.0%	**Basal-like**	0.0%	**Basal-like**	0.0%
Normal-like	17.3%	**Normal-like**	5.7%	**Normal-like**	0.0%

Figure 4.2: Prototype of scorecard to report uncertainty in PAM50 classification due to intrinsic experimental errors in measuring gene expression factors using the example samples GEICAM_09-02639_UU, GEICAM_09-02594_UU, and GEICAM_09-02588_UU.

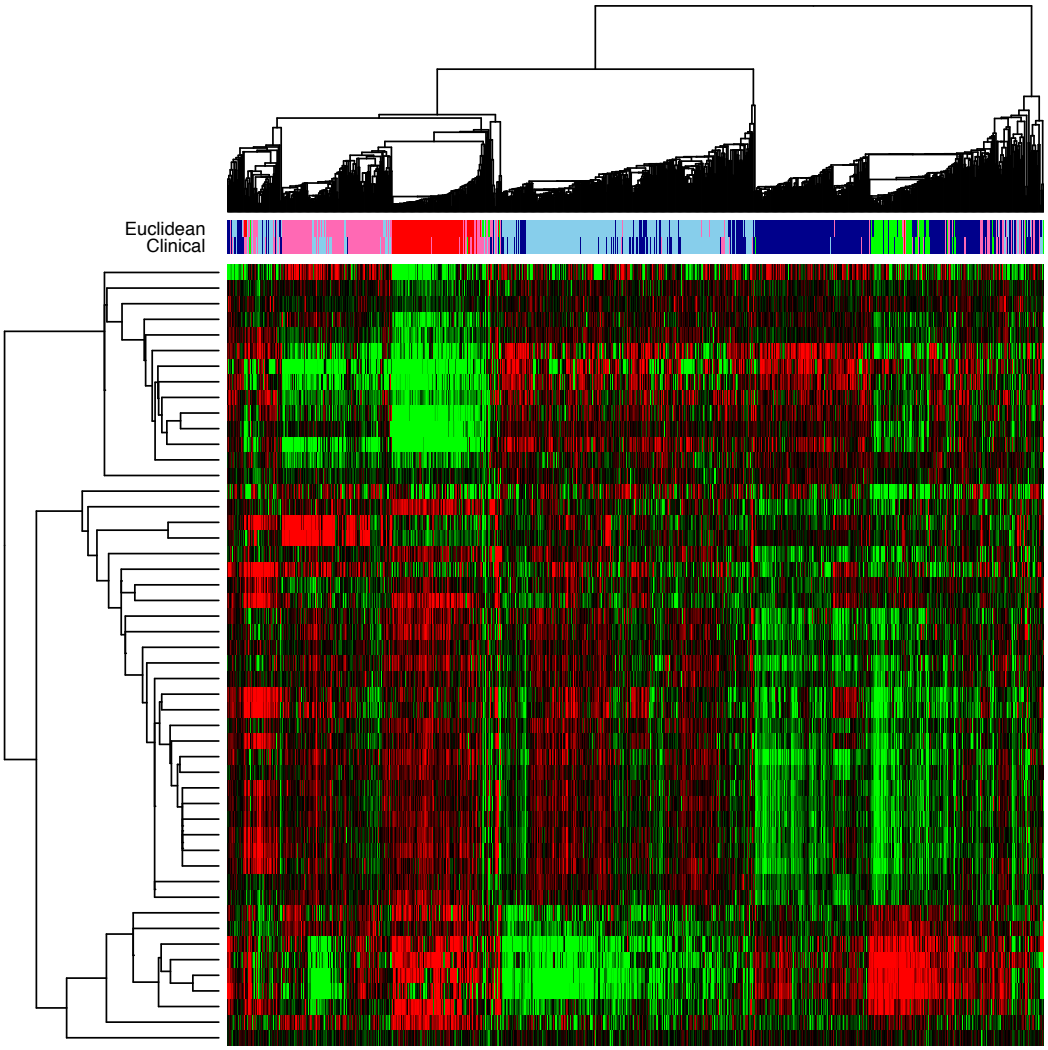

Figure 4.3: Hierarchical clustering of RT-qPCR data for the PAM50 classifier genes normalized to the 5 control genes using samples from the GEICAM 9906 clinical trial. Continuity of PAM50 subtype classifications when using Spearman's Rho (ρ) as a distance metric versus using euclidean distance. Subtype classifications are colored according to Luminal A (dark blue), Luminal B (light blue), HER2-enriched (pink), Basal-like (red), and Normal-like (green).

Table 4.4: GEICAM Sample Subtype Using Euclidean as Distance Metric (as Percentage) - Average-Case Scenario

Original Subtype (simulated samples)	**Classified as Subtype (%)**				
	Luminal A	**Luminal B**	**HER2-enriched**	**Basal-like**	**Normal-like**
Luminal A (303000)	**97.45**	1.23	0.99	0	0.33
Luminal B (270000)	1.13	**97.76**	1.11	0	0
HER2-enriched (189000)	2.12	3.25	**93.79**	0.03	0.81
Basal-like (85000)	0	0	0.93	**98.81**	0.25

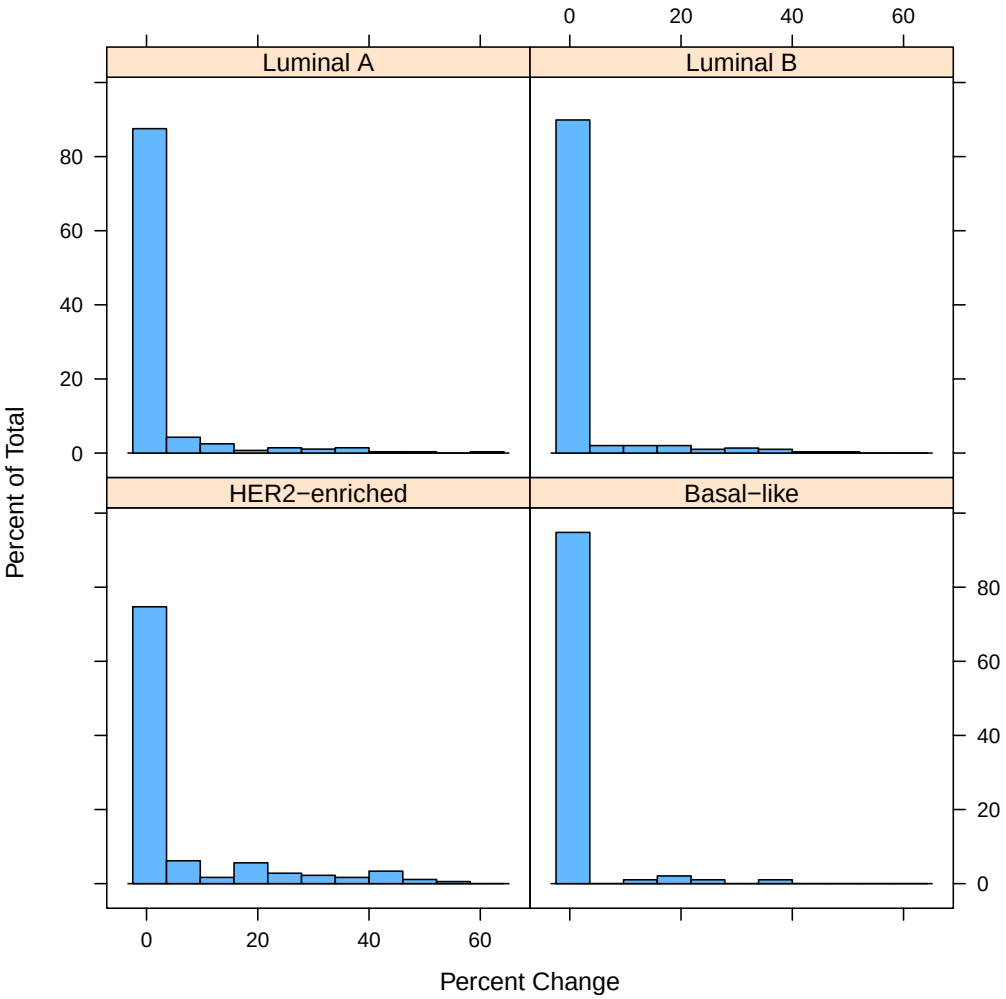

Figure 4.4: Distribution of subtype reproducibility for replicas of individual GE-ICAM samples when using euclidean as the distance metric. Each of the four histograms shows on the y-axis the percentage of parent samples for which there are any number of replicas that changed classification.

CHAPTER 5

DISCUSSION

5.1 Major Contributions

MVAs are becoming commonplace, particularly in oncology, to assist treatment decisions; however, in combination with the advantages of MVAs, there are issues that must be addressed to ensure MVAs are reliable and personalized. As secrets of the human genome continue to be unlocked, MVAs are likely to become more common, still. Thus, measures must be taken to assess the intrinsic technical uncertainty of MVAs, yet there is currently no standard method for this purpose. The research presented has made valuable contributions to fill some of the aforementioned knowledge gaps; specifically, contributions made are as follows:

1. A method to characterize the intrinsic technical uncertainty of MVAs and its effects on the final result

2. A method that advances MVAs towards personalized healthcare by providing personalized confidence measurements on the MVA's result.

3. Explored effectiveness of using a parametric measurement within platform versus using a nonparametric measurement.

Perhaps the paramount issue to address with MVAs is how to characterize the effects of intrinsic technical uncertainty on an MVA's final result; or, in other words, to have a method capable of grasping how variation (error), no matter the size, affects the MVA's final result given the complexity of the MVA. In the PAM50, for example, the expression of 50 classifier genes are measured in 50 separate RT-qPCR reactions (in addition to the five housekeeper genes), essentially creating a 50 dimensional space with each dimension bringing its own associated error, with added potential to make results unpredictable. The method described in this research can be applied in

at least two ways, to characterize the reproducibility of PAM50 classifications: (1) the method can be used to measure the effect of error on the final results, given a known error; and (2) error limits can be defined, given an acceptable amount of classification reproducibility. Additionally, this method is believed to be generalizable to any problem that is based on continuous data values, which provides immediate relevance to numerous disciplines.

In addition to providing a generalizable method to assess the effects of technical uncertainty, another consequential issue was to develop a method to provide personalized confidence measurements on the assay result. At present, there is a strong effort by worldwide government agencies, hospitals, physicians, and researchers to bring personalized medicine to fruition. If medicine is indeed to become personalized, there must also be a method to describe a result's confidence for a single patient sample rather than a summarizing statistic of the overall accuracy of an assay. This research broached the barrier to provide a personalized confidence measurement and will promote improved patient care by informing physicians of the MVA's confidence in the final result.

Similarly, defining under what circumstances a parametric or nonparametric measurement, and which specific measurement, should be employed is critical to designing cbMVAs. Which measurement to use will undoubtedly depend on the cbMVA design; however, based on the research presented, there is evidence that parametric may be the best choice within a single platform. Although the research presented needs to be expanded, it laid ground work to better understand the effects of parametric versus nonparametric statistical measurements within single gene expression platforms.

5.2 Limitations

The methods presented in this research are thorough, necessary, and well-founded; however, there are inherent limitations that must either be taken into consideration or may necessitate further research. Known limitations are as follows:

1. The general method to characterize technical uncertainty in MVAs requires generating a model to generalize variation. Models cannot perfectly represent the true nature of a system.

   (a) The repeated measures experiment is an estimate of the error

   (b) It is assumed that error varies independently (Note: it is not assumed that genes vary independently)

2. The method does not account for uncertainty beyond what is technical (e.g., tumor heterogeneity)

3. The confidence measurement only measures the confidence that a sample is properly classified according to the defined classes (i.e., assumes the class definitions accurately represent the biology)

4. Euclidean distance shows greater reproducibility, however, the work presented does not quantify accuracy

Despite meticulous efforts researchers may make to develop a model, no model can fully represent the intricate details of a system or process. According to the New Oxford American Dictionary, a model is "a simplified description... of a system or process, to assist calculations and predictions" [48]. In the case of the model described in this research, there are two specific limitations: (1) the repeated measures experiment is only an estimate of the error; and (2) error is assumed to vary independently between genes. Both of the aforementioned limitations have minimal implications when addressed appropriately. In the first case, performing a sufficient number of repeated measures will produce an accurate estimate within a small confidence interval. In the second case, a correlation test can be performed using the repeated measures data to determine if there is any significant correlation of the error between genes. If significant correlation is discovered, error could then be varied dependently, according to the associations discovered.

Perhaps the most apparent limitation of the method presented is that it only pertains to technical uncertainty and it does not account for outside factors such as, in the case of the PAM50, tumor heterogeneity, tumor isolation (i.e., directed punch versus full-face cut), and RNA extraction, among others. Accounting for the factors specified will require an entirely different study design and cannot be addressed with the data prepared for this research.

A perhaps more obscure limitation is that the confidence measurement developed only measures the confidence that a sample is properly classified according to the defined classes – meaning the measure assumes class definitions accurately represent the biology. While knowing the confidence of the classification according to the defined classes is important, it should be clear to anyone interpreting the results that the defined classes have limitations of their own. Specifically, anytime classes are imposed on continuous data values, there will be boundaries that engenders ambiguity, by definition.

A final limitation worth mentioning is that the research contrasting euclidean and Spearman's Rho (ρ) as distance metrics does not quantify accuracy, though euclidean distance is demonstrated to have superior reproducibility. Reproducibility is essential to any MVA, at which euclidean distance is clearly superior to ρ, but without a comprehensive comparison of accuracy, judgement cannot be fully executed. On the other hand, there is preliminary evidence that euclidean distance may also have superior accuracy because of the superior classification continuity within the hierarchical cluster – though that is assuming class definitions and hierarchical clustering accurately represent the biology. Further research is necessary to determine whether euclidean, and parametric measurements alike, are more accurate than ρ and other nonparametric measurements within a single gene expression platform.

5.3 Opportunities for Future Work

Scientific research is a complex, never-ending process that often rouses more questions to be explored than were answered to begin with. The complexity of individual disciplines such as biology, medicine, and informatics as separate entities (or silos) is enough to occupy a researcher for the duration of his or her career. Yet, trying to utilize biology, medicine, and informatics together, such as is the case with biomedical informatics, seems to increase the complexity in a nonlinear fashion. Given the methods presented in this research traverse all of the specified disciplines, there are diverse opportunities for future work to clarify and ameliorate the characterization of intrinsic technical uncertainty in MVAs. Two general areas, with broad implications, that need further work are to : (1) explore nontechnical uncertainty; and (2) explore

the implications of parametric versus nonparametric measures.

Exploring nontechnical uncertainty alone requires extensive knowledge and experimentation within biology, medicine, and informatics. Within biology there are questions regarding genetics, heritability, genomics, and proteomics. For example, it is possible the disease was genetic (directly predisposed), or on the other hand there may have been a genetic weakness that made the patient more susceptible to the disease (indirectly predisposed); there are important treatment implications based on the genetics. Likewise, heritability, which accounts for relative contributions of genetic and nongenetic (e.g., environmental) differences, has important implications. Furthermore, deeper knowledge of genomics and proteomics will provide direct insight for drug development. Combining the biology with medicinal expertise (e.g., pathology, internal medicine, etc.) and informatics (computer science, statistics, data storage, data mining, etc.) presents numerous paths to not only improve treatment, but the MVAs used to help determine the ideal treatment. Areas of immediate interest, specific to the PAM50, are understanding the effects of tumor heterogeneity, tumor isolation (i.e., directed punch versus full-face cut), and RNA extraction methods. There are many opportunities to improve this work.

Exploring the implications of parametric versus nonparametric measures, though perhaps less complex, is a formidable challenge. Not only exploring parametric versus nonparametric measures, but there are numerous measures within each class. Additionally, the results are likely to change based on the MVA design, whether there are one or multiple platforms employed, and the underlying natural value distributions. It may, in fact, be that there are few generalizations to be made, and that the various options available will need to be tested individually for each MVA.

5.4 Relevance to Biomedical Informatics

Biomedical informatics is a multidisciplinary field and, as defined by Shortliffe and Blois, is "the scientific field that deals with biomedical information, data, and knowledge – their storage, retrieval, and optimal use for problem solving and decision making" [49]. Essentially, biomedical informatics will play a critical role in improving

patient care by improving physician access to, and understanding of biomedical information. Among the many topics within the realm of biomedical informatics, improving MVAs is a critical piece to realize improved patient care, since physicians will be able to make more informed decisions. This research has improved MVAs in the following fashions: (1) developed a generalizable method to characterize intrinsic technical uncertainty in MVAs; (2) developed a method to provide personalized confidence measurements for individual patient samples; and (3) laid ground work to better understand the effects of parametric versus nonparametric statistical measurements within single gene expression platforms.

A generalizable method to characterize intrinsic technical uncertainty in MVAs will become increasingly important as scientific and medical breakthroughs allow for more complex MVAs to provide personalized treatment guidance. Moreover, MVAs will likely become more complex over time and there must be methods to characterize the intrinsic technical uncertainty within the MVA to ensure the MVAs are reproducible and accurate. The method described was developed for and tested on the PAM50, but the principles employed are generalizable to any assay based on continuous data values; and training in biomedical informatics facilitated the method's development.

At present, there is a strong effort by worldwide government agencies, hospitals, physicians, and researchers to bring personalized medicine to fruition. If medicine is indeed to become personalized, there must also be a method to describe a result's confidence for a single patient sample rather than a summarizing statistic of the overall accuracy of an assay. This research broached the barrier to provide a personalized confidence measurement and will promote improved patient care by informing physicians of the MVA's confidence in the final result.

Although there is much work to be done to understand when parametric and nonparametric statistical measurements should be employed, not to mention which of the many available measurements within either class, there is evidence that parametric may be the best choice within a single platform. A deeper understanding of the intricacies of statistical measurements for use in MVAs will allow for more reliable and accurate MVAs – an important task within biomedical informatics.

Each contribution from this research was facilitated by training in biomedical informatics. Many researchers have spoken of, and endeavored to overcome, the silo-like nature of historical disciplines [50–53] such as biology, medicine, and statistics – and the biomedical informatics discipline is a direct result of their efforts. By seeking to bring together the knowledge from several "silos," this research achieved tasks that otherwise may have been overlooked.

CHAPTER 6

CONCLUSION

A method was developed based on Monte Carlo simulations and limited experimental measurements to estimate the effect of the intrinsic experimental errors in the measured factors contributing to MVAs. While the specifics of the error distribution functions given are not universal functions, and recalculation for each lab or experimental setting must be considered, the proposed method is generalizable and adaptable to any MVA.

Furthermore, the method presented advances MVAs towards personalized healthcare by providing personalized confidence measurements on the assay result. Providing personalized confidence measurements on the assay result will allow physicians to make better treatment decisions, although the confidence measure assumes that classes accurately represent the disease's biology.

While the effectiveness of using a parametric measurement within platform versus using a nonparametric measurement is inconclusive, the data presented suggest that parametric measurements may be better suited for use within a single platform. Further understanding of the effects of parametric versus nonparametric under various circumstances will be invaluable for the design of future MVAs.

Finally, using the proposed method based on Monte Carlo simulations and the error model described here, we have presented data that suggests PAM50's subtype classifications are highly reproducible on a large, independent sample set from the GEICAM 9906 clinical trial. We also show that there are a nonnegligible number of samples for which a significant number of the Monte Carlo simulated samples classify differently than the parent sample, indicating that the classification of the original sample may not be reliable – thus highlighting the need for the new score card that can inform clinicians on the probability that a particular sample could be classified as a different tumor subtype.

REFERENCES

[1] Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology.* 2009;27:1160 –1167.

[2] DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. *CA: A Cancer Journal for Clinicians.* 2011;61:408–418.

[3] Berry DA, Cronin KA, Plevritis SK, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine.* 2005;353:1784-1792.

[4] Cheang MCU, Voduc KD, Tu D, et al. Responsiveness of intrinsic subtypes to adjuvant anthracycline substitution in the ncic.ctg ma.5 randomized trial. *Clin Cancer Res.* 2012.

[5] Esserman LJ, Berry DA, Cheang MCU, et al. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the i-spy 1 trial (calgb 150007/150012; acrin 6657). *Breast Cancer Res Treat.* 2011.

[6] Prat A, Parker JS, Karginova O, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* 2010;12:R68.

[7] Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med.* 2004;351:2817-26.

[8] Habel LA, Shak S, Jacobs MK, et al. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast Cancer Res.* 2006;8:R25.

[9] Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol.* 2006;24:3726-34.

[10] Dowsett M, Cuzick J, Wale C, et al. Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a transatac study. *J Clin Oncol.* 2010;28:1829-34.

[11] Albain KS, Barlow WE, Shak S, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive,

oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncol.* 2010;11:55-65.

[12] Sotiriou C, Neo SY, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A.* 2003;100:10393-8.

[13] Buyse M, Loi S, Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst.* 2006;98:1183-92.

[14] Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of Tamoxifen-Treated, Node-Negative breast cancer. *New England Journal of Medicine.* 2004;351:2817–2826.

[15] Tutt A, Wang A, Rowland C, et al. Risk estimation of distant metastasis in node-negative, estrogen receptor-positive breast cancer patients using an RT-PCR based prognostic expression signature. *BMC Cancer.* 2008;8:339.

[16] Vijver MJ, He YD, Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002;347:1999–2009.

[17] Gancberg D, Järvinen T, Leo Ad, et al. Evaluation of HER-2/NEU protein expression in breast cancer by immunohistochemistry: An interlaboratory study assessing the reproducibility of HER-2/NEU testing. *Breast Cancer Research and Treatment.* 2002;74:113–120.

[18] Hoang MP, Sahin AA, Ordòñez NG, Sneige N. HER-2/neu gene amplification compared with HER-2/neu protein overexpression and interobserver reproducibility in invasive breast carcinoma. *American Journal of Clinical Pathology.* 2000;113:852–859.

[19] Jacobs TW, Gown AM, Yaziji H, Barnes MJ, Schnitt SJ. Specificity of HercepTest in determining HER-2/neu status of breast cancers using the united states food and drug administration-approved scoring system. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology.* 1999;17:1983–1987. PMID: 10561248.

[20] Nielsen TO, Parker JS, Leung S, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin Cancer Res.* 2010;16:5222–32.

[21] McShane LM, Aamodt R, Cordon-Cardo C, et al. Reproducibility of p53 immunohistochemistry in bladder tumors. *Clinical Cancer Research.* 2000;6:1854–1864.

[22] Tubbs RR, Pettay JD, Roche PC, Stoler MH, Jenkins RB, Grogan TM. Discrepancies in clinical laboratory testing of eligibility for trastuzumab therapy: apparent immunohistochemical false-positives do not get the message. *Journal of*

Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2001;19:2714–2721. PMID: 11352964.

[23] Rhijn BWG, Vis AN, Kwast TH, et al. Molecular grading of urothelial cell carcinoma with fibroblast growth factor receptor 3 and MIB-1 is superior to pathologic grade for the prediction of clinical outcome. *Journal of Clinical Oncology.* 2003;21:1912–1921.

[24] Wiley EL. High-Quality HER-2 testing: Setting a standard for oncologic biomarker assessment. *JAMA: The Journal of the American Medical Association.* 2004;291:2019–2020.

[25] Hu Z, Fan C, Oh DS, et al. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics.* 2006;7:96.

[26] Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature.* 2000;406:747–52.

[27] Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A.* 2001;98:10869–74.

[28] Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A.* 2003;100:8418–23.

[29] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A.* 2002;99:6567–72.

[30] Kelly CM, Bernard PS, Krishnamurthy S, et al. Agreement in risk prediction between the 21-gene recurrence score assay (oncotypedx®) and the pam50 molecular classifier in early stage er-positive breast cancer. In Press 2012.

[31] Siegel S. Nonparametric statistics. *The American Statistician.* 1957;11:pp. 13-19.

[32] Pagano M, Gauvreau K. *Principles of Biostatistics*;2. Duxbury 2000.

[33] CHOI SC. Tests of equality of dependent correlation coefficients. *Biometrika.* 1977;64:645 –647.

[34] Pearson ES, Snow BAS. Tests for rank correlation coefficients III. distribution of the transformed kendall coefficient. *Biometrika.* 1962;49:185–191. ArticleType: research-article / Full publication date: Jun., 1962 / Copyright © 1962 Biometrika Trust.

[35] Page EB. Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association.* 1963;58:216–230. ArticleType: research-article / Full publication date: Mar., 1963 / Copyright © 1963 American Statistical Association.

[36] Doolen GD, Hendricks J. Monte carlo at work. *Los Alamos Science.* 1987;15:142-143.

[37] Eckhardt R. Stan ulam, john von neumann, and the monte carlo method. *Los Alamos Science.* 1987;15:131–137.

[38] Metropolis N. The beginning of the monte carlo method. *Los Alamos Science.* 1987:125–130.

[39] Metropolis N, Ulam S. The monte carlo method. *Journal of the American Statistical Association.* 1949;44:335–341. ArticleType: research-article / Full publication date: Sep., 1949 / Copyright © 1949 American Statistical Association.

[40] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics.* 1953;21:1087.

[41] HASTINGS WK. Monte carlo sampling methods using markov chains and their applications. *Biometrika.* 1970;57:97 –109.

[42] Panagiotopoulos A. Direct determination of phase coexistence properties of fluids by monte carlo simulation in a new ensemble. *Molecular Physics.* 1987;61:813–826.

[43] Ehrman JR, Fosdick LD, Handscomb DC. Computation of order parameters in an ising lattice by the monte carlo method. *Journal of Mathematical Physics.* 1960;1:547.

[44] Boulesteix AL, Porzelius C, Daumer M. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics.* 2008;24:1698-706.

[45] Nakagawa S, Hauber ME. Great challenges with few subjects: statistical strategies for neuroscientists. *Neurosci Biobehav Rev.* 2011;35:462-73.

[46] Martin M, Rodriguez-Lescure A, Ruiz A, et al. Randomized phase 3 trial of fluorouracil, epirubicin, and cyclophosphamide alone or followed by paclitaxel for early breast cancer. *J Natl Cancer Inst.* 2008;100:805–14.

[47] Team RDC. R: A language and environment for statistical computing. 2011.

[48] *New Oxford American Dictionary.* Oxford University Press2nd ed. 2012.

[49] Shortliffe Ee, Cimino Jae. *Biomedical informatics: Computer applications in health care and biomedicine.* New York: Springer;3rd ed. 2006.

[50] Warden R. Impact of cabig on the european cancer community. *Ecancermedicalscience.* 2011;5:225.

[51] Amalberti R, Benhamou D, Auroy Y, Degos L. Adverse events in medicine: easy

to count, complicated to understand, and complex to prevent. *J Biomed Inform.* 2011;44:390-4.

[52] Valenta AL, Brooks I, Laureto RA, Ramaprasad A. Breaking the silo. using informatics to support clinical and translational science. *J Healthc Inf Manag.* 2007;21:15-7.

[53] Arp R, Romagnoli C, Chhem RK, Overton JA. Radiological and biomedical knowledge integration: The ontological way.