Computational Approaches for Addressing Complexity In Biomedicine

by

Justin Reed Brown

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2012 by the
Graduate Supervisory Committee:

Valentin Dinu, Chair
Diana Petitti
William Johnson

ARIZONA STATE UNIVERSITY

May 2012

ABSTRACT

The living world we inhabit and observe is extraordinarily complex. From the perspective of a person analyzing data about the living world, complexity is most commonly encountered in two forms: 1) in the sheer size of the datasets that must be analyzed and the physical number of mathematical computations necessary to obtain an answer and 2) in the underlying structure of the data, which does not conform to classical normal theory statistical assumptions and includes clustering and unobserved latent constructs. Until recently, the methods and tools necessary to effectively address the complexity of biomedical data were not ordinarily available. The utility of four methods--High Performance Computing, Monte Carlo Simulations, Multi-Level Modeling and Structural Equation Modeling—designed to help make sense of complex biomedical data are presented here.

TABLE OF CONTENTS

Page

CHAPTER PAGE

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Introduction:

Berman defined biomedical informatics as, “the branch of medicine that combines biology and computer science” (Berman 2007). Musen & Van Bemmel suggested that the pervasive cognition that computers are the predominant focus has contributed to a “tendency to define any activities with heavy computer focuses as informatics” (Musen and Van Bemmel 2004). For years there was a strong push to see computers in biomedicine as a predominant focus in biomedical informatics (Shortliffe and Cimino 2006).

While computers are highly useful tools, they are not solutions to biomedical problems in and of themselves. Having large well curated collections of information that can be easily exchanged and shared is useful only insofar as we can analyze the data and make sense of what it means. Absent the ability to effectively analyze biomedical data, the potential for making advancements in medicine is unrealized. We need to be able to decode an understand what the vast amounts of biomedical data mean in order to develop new diagnostic screening tools for disease, clinical decision support systems or guidelines for addressing specific public health problems.

Despite the ever growing need to analyze biomedical data, the task is especially challenging in biomedicine because of the size and complexity of the data. Cios and Moore specifically discuss the uniqueness of biomedical data compared to other fields and note that “researchers in other fields may not be aware of the particular constraints and difficulties of the privacy-sensitive, heterogeneous, but voluminous data of medicine” (Cios and Moore 2002). Additionally, Cios and Moore also note that the “mathematical understanding of estimation and hypothesis formation in medical data may be fundamentally different than those from other data” (Cios and Moore 2002). National Institute of Heath’s Common Fund which is part of the Division of Program Coordination, Planning and Strategic Initiatives (DPCPSI) suggests that while there are many new challenges in dealing with biomedical data that “at the core of the challenge [today] is one of “big data” where handling and working with complex data at a large scale is both quantitatively and qualitatively different than at a smaller scale” (NIH 2011).

As a result, there is a need to develop, adapt and disseminate methods to help address challenges and complexity inherent in modern biomedical data. In this dissertation, four methods, High Performance Computing, Monte Carlo Simulations, Multi-Level Modeling and Structural Equation Modeling that help address challenges posed by complexity in biomedical data are described in the context of their application to specific real-life problems.

Origins and Nature of Complexity:

Complexity in modern biomedicine is a product of 1) millions of years of complex evolutionary change and 2) advancements in technology which enable the generation and storage of massive quantities of data (Crick 1988). (Cios and Moore 2002) (NIH 2011).

The notion that biomedical data has become increasingly complex has become a major theme among premier scientists in the past few decades. Dr. Francis Crick, a co-discovered the double helix structure of DNA and among the most influential scientists of the 20th century, was one of the first to articulate this theme. The preface for his 1988 book, “What Mad Pursuit” specifically states that “science in [the 20th] century has become a complex endeavor” (Crick 1988). In the book, Dr. Crick expands further on the notion of biomedical research being complex with the recurrent argument that, “while Occam's razor is a useful tool in the physical sciences, it can be a very dangerous implement in biology. It is thus very rash to use simplicity and elegance as a guide in biological research” (Crick 1988). Dr. Crick suggests that part of the reason for the inherent complexity in the biological sciences and its lack of grand theories is a product of evolution. While biologists have ‘laws’ such as Mendelian inheritance, these are not nearly as precise or accurate as, for example, the theory of relativity. In physics, the laws were set from the start and are not the product of incremental change over time. In biology, what we observe is the product of millions of years of sequential evolutionary changes with each change or adaptation built upon the previous (Crick 1988). Dr. Crick urges us to consider that while the end product is not always the most efficient system possible, nature had to build sequentially on what was already there to obtain the complex end result we see today.

Numerous noteworthy scientists have directly commented positively on Dr. Crick's views. For example, Phillip Anderson, who won the 1977 Nobel Prize in physics for his work on the electronic structure of magnetic and disordered systems which paved the way for many modern computing systems, specifically said that as a general principle for "learning the truth about the world around us, Crick's words are as good a guide to that end as I have seen" (Anderson 1990).

Adding to the inherent complexity of the natural world that Dr. Crick and others discuss are the explosion in the physical amount of data about the natural world and its structure that is spurred on by advances in computers and technology (Cios and Moore 2002) (Chen, et al. 2010) (Shortliffe and Cimino 2006). For example, the microarray is a relatively new technology used to measure genomic and proteomic properties of an organism which has massively increased the size of data. It is common for current generation microarrays to have tens of thousands to over a million data points on each array or for each subject or individual organism present in the study design (Brown, et al. 2011) (Affymetrix 2012). Secondly, database systems, information transfer technologies and the availability of inexpensive storage solutions are some factors which have contributed to an increasing number of dedicated data repositories which collect, integrate and when agreed upon share data or results. A few examples includes the Center for Health Information and Research which contains information on millions of patients and tens thousands of doctors in Arizona or the National Cancer Institute's Cancer Biomedical Informatics Grid known as caBIG which collects and shares data from over 700 cancer research institutions (Johnson, et al. 2011) (Nationanl Cancer Institute 2012). The massive quantity of data produces a significant computational challenge to be able to efficiently analyze the data.

Large datasets, including those from microarray experiments and data repositories can lead to further complexities in the data such as the presence of multi-level structures. Also, many modern statistical methods are iterative and take many cycles to complete which compound the computational burden associated with voluminous amount of data. (Muthen and Muthen 2011).

Overview of Methods:

The dissertation is divided into four chapters: high performance computing methods, Monte Carlo Simulations, multi-level modeling and finally structural equation modeling. The first two chapters describe near universal methods which help facilitate any quantitative analysis in biomedicine. The third chapter serves as a transition and shows that newer and more complex modeling approaches are necessary to obtain accurate results; such as when clustering and multi-level structures exists in the data. In addition, the multi-level modeling chapter also shows that newer modeling approaches to address complexity in biomedical data also allow researchers the ability to investigate relationships not possible with simpler models. The fourth chapter on structural equation modeling extends the notion that newer methods are often necessary to fully understand complex data and shows how structural equation modeling allows researchers new opportunities to develop diagnostic tests and to model unobservable latent constructs.

High Performance Computing Overview:

The high performance computing chapter lays out a number of methods and tools which are highly and nearly universally applicable to address problems of large data in biomedicine. The high performance computing methods presented in chapter 2 help address the informatics challenges of managing large data as well as physically running the large number of computations that are often necessary in an efficient fashion. This chapter details the use and benefits of database connectivity, pipeline parallelism, multi-core processing and distributed multi-core grid processing. This chapter is not meant to be an exhaustive treatise on computer science solutions to big data problems. Not only does the chapter not address every computer science challenge to big data, the scope in which they are examined is tailored towards individuals performing quantitative analysis of biomedical data. The topics presented were also used to facilitate the utilization of the other 3 method chapters.

Secondly, much of the literature on topics such as parallel processing is not tailored towards analysts and much of the relevant information is scattered throughout a number of different sources (ie. SAS manuals for example). The goal here is to present a number of tools, explain their utility and provide scripts and examples to help facilitate their implementation. The concrete examples and scripts that are included are focused at quantitative analysts and designed to make using such methods easier for many analysts.

Monte Carlo Simulations Overview:

This chapter advocates for broader use of Monte Carlo simulations in order to assess whether the models being used give accurate results as well as to pick the best alternative when one model fails. Many statistical assumptions such as homogeneity of variance (equal variances between two groups (Cohen, et al. 2002) (Keppel and Wickens 2007)) are commonly, and sometimes by design, egregiously violated in biomedical data (Quackenbush, Causton and Brazma 2003). It is impossible to make sense of biomedical data and to make scientific advances if the statistical models a researcher is using do not giving accurate answers. The problem of inaccurate models is exacerbated by big data. A Monte Carlo simulation is a method for testing mathematical models in which a large number of datasets are randomly generated to mimic a given set of properties. A model is then tested on each random dataset and the results are recorded to gain insight into model behavior. Given the complex nature of biomedical data, Monte Carlo Simulations are an indispensable method to help confirm that on average over repeated sampling the results obtained from an analysis are accurate.

Although Monte Carlo simulations have been a defacto standard for validating new statistical models for decades, the computational demands have largely kept them from being more widely used in routine statistical analyses (Fan 2002). Since many modern methods such as maximum likelihood are iterative there is no analytic solution to derive and simulations are the only way to determine model performance (Fan 2002) (Rubinstein and Kroese 2007). While it is advocated for a wider more general use of Monte Carlo simulations, even in smaller biomedical data sets when assumptions are not met, the even larger challenge today is comes from big data. Given the size of many modern biomedical datasets the question is not just restricted to which method will perform well under one specific set of conditions (ie. one pairwise contrast), but rather which will perform best over multiple comparisons across the entire range of data. With modern microarrays or databases containing population health information where tens of thousands or millions of observations are available, there may likely be a large spread in variances or adherence to model assumptions across the range of statistical contrasts. As a result, the question with big data becomes not which model is most accurate for any single contrast but which model is most accurate across the entire surface and range of responses in the data that will be taken into account.

One major challenge with big data is physically running the tens of thousands or millions of statistical tests so advocating running equality of variance tests on every observation is not realistic; even with the HPC methods that are presented in chapter 2 to help. Rather, sampling methodologies are suggested as an alternative. Depending on the structure of the experimental design, simple random sampling, stratified random samples, cluster randomized or even a multi-stage probability sample of a small portion of the data can be drawn. From this smaller and manageable subsample, screening tests on the data can be run and a Monte Carlo simulation can be created to determine the best method given the structure and surface of the dataset.

While sampling methodology helps reduce computational demands, HPC implementations for Monte Carlo simulations are presented to help make Monte Carlo simulations themselves possible as a part of general research routine and process flow. Single core and parallel processing SAS macro programs and templates are provided to help users more easily perform Monte Carlo simulations.

Multi-Level Modeling Overview:

This chapter demonstrates the proper use of Multi-Level modeling to address the complex problem of clustering in biomedical data with the real world example of modeling adherence to treatment. In addition to general correction for clustering, the chapter also illustrates the effective use of centering and how variance components estimated at the unit of observation and at the unit of clustering can be analyzed to help unravel the complexity in biomedical data.

Unlike violations of homogeneity of variance discussed extensively in the Monte Carlo simulation chapter, it is well understood and agreed that multi-level modeling is the correct approach to take to deal with violations of independence of observations. (Raudenbush and Bryk 2001) (Cohen, et al. 2002) (Tabachnick and Fidell 2006). The assumption of independence assumes that each observation is uncorrelated with or independent of every other observation in the dataset. The primary way in which this assumption is violated is when clusters or nesting exist within the dataset. This sort of clustering or nesting is pervasive in modern biomedical datasets; especially those derived from databanks. A few of the more commonly encountered biomedical clustering problems are: doctors are nested within hospitals, patients are nested within doctors, tissue samples might be nested within lab if a research group obtains samples from multiple tissue banks or repeated measures within the same patient such as blood tests over the course of time. Even very small violations of independence can lead to dramatically inflated type 1 error rates (Raudenbush and Bryk 2001).

Stulberg et al. 2010 used a multi-level model to control for clustering created by taking measurements across multiple hospitals and

Mills et al. 2006 used multi-level modeling to account for regional clustering in the sampling design across geographic regions (Stulberg, Delaney, et al. 2010) (Mills, et al. 2006). Wile there are numerous examples of multi-level modeling being used in premier journals such as JAMA, there are also examples in which no metion of efforts taken to control for possible clustering. Parker et al. 2009 propose a possible gene signature for breast cancer samples but they use samples from two distinct sample types fresh frozen and parrifin fixed formalin embedded (FFPE) which come from 5 different cohorts (Parker, et al. 2009). Across the many different comparisons performed, there is no mention of clustering in the paper. While the probablity does exist that the effect was neglegible, this is highly unlikely. Personal research not discussed in detail in this dissertation found that with similar breast cancer samples from multiple cohorts exhibited intraclass correlations (a measure of clusering) in excess of 0.6 (Seliegman and Brown 2011).

Additionally, although multi-level modeling is sometime used when necessary, many papers in biomedicine such as Stulberg, Delaney, et al. 2010 or Mills, et al. 2006 are uninterpretable because they omit key information about the model estimation and specification such as how the variables were centered. This is because the complex relationship between an independent predictor and dependent variable is comprised of variability at the unit level of observation as well as at the cluster or grouping level. Variability can be partitioned in different ways based on the research question of interest. The researchers decisoin on how to partition variance determines the interpreatation of the paramter estimates (Raudenbush and Bryk 2001) (Enders and Tofigi 2007). Without knowing critical model specifications such as centering, it is impossible to accurately interpret the paramete estimates and to know if the model estimated is confounded (Raudenbush and Bryk 2001) (Enders and Tofigi 2007).

While using multi-level modeling to correct for clustering is a useful and necessary procedure, multi-level modeling provides a plethora of additional information beyond simply correcting for the structure in the data which does not meet certain statistical assumptions. Most authors in biomedicine, Stulberg and Mills included do not make use of the additional information and estimates provided generated when running a multi-level modeling. The extra complexity of multiple levels of clustering also provides additional sources of information and is something that helps us better unravel complexity. In multi-level modeling the equations and variance components are estimated separately for each level of a given cluster. One example of the utility of this which is presented later is that this approach allows us to estimate the relative amount of variability in adherence to medical treatment for doctors and patients separately. Multi-level modeling is generally underutilized in biomedicine and when it is, it is almost exclusively used to correct for clustering rather than using the additional complexity to help better explain the world around us.

Structural Equation Modeling Overview:

The final chapter focuses on structural equation modeling in biomedicine. The modeling approach described here addresses another situation where simpler modeling approaches do not fully capture the richness of the data and complexity of the natural world. Multi-level modeling is a mathematical subset of structural equation modeling. The origins of structural equation modeling date back to path analysis in the 1920's. Innovations in computer processing combined with the development of algorithms, such as estimation maximization (EM), that help efficiently perform maximum likelihood estimation have led to an increased popularity of structural equation modeling in many areas of research especially in the social sciences. Structural equation modeling is, however, relatively underutilized in biomedicine.

In the few cases where structural equation modeling is used, the authors often specifically advocate for wider use of the method (Dahly, Adair and Bollen 2009). Tu 2009 argues more generally for expanded use of structural equation modeling as a potentially highly useful tool for advancing epidemiology and biomedicine (Tu 2009).

One aspect of structural equation modeling which is especially useful in biomedicine is latent variable modeling. Often times the specific objects or theory of interest are measured indirectly. For example, peptide microarrays use random 20mer peptides to indirectly measure the presence and activity of antibodies in a sera sample when a part of the antibody binds to one or more of the peptides on the array. Simple analytic methods such as a t-test, analysis of variance or logistic regression can determine if there are differences in expression of a peptide between groups but cannot tell us anything about the entire antibody of interest. Structural equation modeling provides a framework to not only infer which peptides might represent binding of a single antibody, but also provide detailed information about the latent antibody itself; such as how the antibody might correlate with disease status. This chapter describes latent class modeling of peptide data. Additionally, it demonstrates how latent class modeling can be useful as a medical diagnostic algorithm.

Discussion:

In summary, millions of years of sequential evolutionary pressure and random variation, each building upon the last, has helped to create the complex living world we inhabit and observe today (Crick 1988). Additionally, modern technological advances, primarily via computers, have created many additional complex challenges stemming from big data. Four methods, high performance computing, Monte Carlo simulations, multi-level modeling and structural equation modeling are presented to help address challenges posed by complex modern biomedical data. Furthermore, while simpler analytic methods can be a good starting points, there are many occasions in which more complex modeling techniques are necessary to obtain accurate results. The more complex models explored estimate new parameters and allow researchers to investigate relationships between variables and better unravel the complexity in modern biomedical data in ways not possible with more simplistic models. The hope is that the methods demonstrated in this dissertation will help accelerate the pace of biomedical discoveries; which will in the end help to ameliorate the quality of life for countless individuals.

CHAPTER 2

HIGH PERFORMANCE COMPUTING METHODS IN SAS

Chapter Overview:

This chapter presents four high performance computing methods: database connectivity, pipeline parallelism, multi-core parallel processing and distributed grid parallel processing. These methods are presented to help analysts cope with 1) the voluminous nature in physical size of modern biomedical datasets and 2) the computational increases associated with iterative algorithms which are common in many modern statistical models. Program code and example syntax is provided in SAS because it is the most widely used statistical analysis program in the world. Analytic results are presented that demonstrate the dramatic decrease in computational time from using these methods.

Problem Abstract:

Modern technological advancements such as microarrays and database technology have led to an explosion in the amount of data that must be or is available for analysis. This explosion of data has made it difficult to physically do all of the computations necessary to process large datasets or to run interactive models in a reasonable time frame. A number of computer science methods exist to help deal with the physical processing constraints. Unfortunately, many of these methods, such as parallel processing and distributed grid computing, require highly specialized computer science skillsets which many average analysts lack.

The goal of this chapter is to provide a set of templates and tools that allow average analysts the ability to easily and more efficiently make use of high performance computing methods. The use of database connectivity, pipeline parallelism, multi-core parallel processing, and distributed grid computing are described as tools to help process the large quantities of data in modern biomedical datasets. It is shown how using such methods reduce dramatically the storage needs and the time needed to analyze complex and voluminous biomedical data.

The methods are demonstrated using SAS. SAS was chosen because it is the world leader in analytic software. SAS analytic software is used by more than 55,000 sites including businesses, governments and universities in 129 countries. Additionally more than 90% of the Fortune Global 500 companies use SAS (SAS Institute 2012). Presenting these methods in SAS provides a common language and platform through which many analysts are already familiar; thus making the methods accessible to as large a group of analysts as possible.

Background:

The use of relational databases to store data has been well documented (Elmasri and Navathe 2006). Some of the reasons for storing data in relational databases are: improved retrieval of information, elimination of redundancy and reduction in storage space needs, and the potential for integrating multiple forms of data quickly and easily. Many text books such as those published by Elmasri & Navathe and Shortliffe &

Cimino specifically discuss these benefits of storing biomedical data (including genomic data) in databases (Elmasri and Navathe 2006) (Shortliffe and Cimino 2006). Beyond the many general discussions of the benefits and general strategies, authors such as Corwin, Siliberschatz, Miller & Marenco propose very specific database solutions such as the use of dynamic tables for storing biomedical data in relational databases (Corwin, et al. 2007).

The one thing that is usually missing from the discussion of relational databases in biomedicine is commentary on integrating databases with analytic software such as SAS. Having large well curated databases that are optimally normalized, can be easily exchanged and shared seamlessly across platforms is useful only insofar as the data can be analyzed and interpreted. An external program is almost always used to analyze the data and if a researcher cannot easily integrate the data with a database, many of the efficiencies and gains from a database are lost. Database connectivity is a fairly simple technology implemented in many programs such as SAS, SPSS, MATLAB and others. Connecting an analytic program directly to a database speeds up the process by not requiring the output of a flat file (such as a .csv or excel file) to the be loaded into an analytic program, reduces potential errors in exporting and importing as well as reducing storage size on disk necessary for replicating database information in flat files. Unfortunately, the benefits and process of database connectivity is not always explained in a straightforward way to end analytic users (Shamlin 2009) (Stokes, Bradstreet and Hill 2002).

Another highly useful tool is parallel processing and distributed grid processing. There has been a large focus in the biomedical literature on high performance computing (HPC) focuses on supercomputing clusters. In addition to technical papers, companies such as Cray and IBM extensively market to the biomedical community while many research institutions such as Arizona State University offer training on how to use HPC environments. Unfortunately HPC clusters are not available to many researchers or institutions and programming in an HPC environment requires an additional highly specialized skill set (Hager and Wellein 2010). While many analytic programs offer some level of parallelization, this dissertation specifically focuses on SAS.

Part of the reason for focusing on SAS or an analytic package generally is because the complex mathematical operators of modern statistical methods are already implemented. While one could theoretically program his or her own maximum likelihood and estimation maximization algorithm to perform structural equation models in a language understandable to a HPC system, the difficulty and time necessary to do so would be quite high. SAS offers a number of modules to help implement parallel processing in an easy and efficient fashion. Integrating them into a workflow can dramatically reduce the time needed to perform computationally intensive analyses.

Modern analytic packages such as SAS have become so large that even individuals who have used such programs for years are unaware of many of the HPC related features which have shown up in recent years. In addition, even when many researchers are aware they exist, some do not immediately recognize the benefits they provide to the biomedical community and many often assume that they are so complex that their use will be difficult. Most of the literature on the application and use of HPC parts of SAS are contained in their technical manuals and SAS User Group International (SUGI) publications.

The technical manuals are complex to understand and not well organized. HPC applications are spread across multiple different SAS applications requiring the user to often consult multiple manuals. Also, within each manual, key details needed to perform simple processes are scattered about and not well documented; even for those who are experienced in using such applications. SAS User Group International (SUGI)/SAS Global Forum publications are somewhat more useful. However, they do not always focus specifically on applications in biomedicine. There are some excellent SUGI publications such as “Threads Unraveled: A Parallel Processing Primer” which gives an excellent description of parallel processing (Shamlin, Threads Unraveled: A Parallel Processing Primer. 2004). However, such publications give no details on how to implement such processes in SAS. Conversely, there are other HPC related papers which do give technically correct descriptions, are sometimes brief , leave out many key features and do not make concrete conceptual links (Stokes, Bradstreet and Hill 2002). For example, good parallel processing primers rarely are extended to include distributed grid computing and grid computing primers (or technical manuals) almost completely omit any discussion of multi-core parallel processing. Also, most grid computing or papers which address scaling out are focused on sending processes to remote machines focuses almost entirely on dedicated servers.

The goal is to provide a context within which the processes presented can be useful to biomedical researchers as well as design templates to make it easier for analysts to more easily implement such features. Furthermore, with the prevalence of multi-core chips in almost every computer hardware application today, it is believed that a much more informative discussion of how to layer multi-core parallel processing with distributed grid computing is needed. This is because remotely submitting a job to a modern platform will almost always be going to a multi-core system. Taking advantage of the ability to scale out to a grid cannot be fully utilized without also being able to ensure the process will be processed in parallel once it reaches the grid or other machine. These computer tools, while not solutions to unraveling the complex world around us in and of themselves dramatically increase the performance and efficiency of quantitative analyses of biomedical data.

A few key HPC related processes which are highly useful for addressing current computational challenges in biomedicine are selected. The goal is also to provide a context within which the processes we present can be useful to biomedical researchers as well as how to simply implement them. Furthermore, with the prevalence of multi-core chips in almost every computer hardware application today, we believe that a much more informative discussion of how to layer multi-core parallel processing with distributed grid computing is needed. This is because remotely submitting a job to a modern platform will almost always be going to a multi-core system. Taking advantage of the ability to scale out to a grid cannot be fully utilized without also being able to ensure the process will be processed in parallel once it reaches the grid or other machine.

Methods:

Database Integration and Interfaces:

The concept of database integration ostensibly deals with connecting an analytics package directly to a database management system (DBMS) such as MS SQL, MySQL, Oracle or others. Database systems provide a number of benefits for the efficient storage and management of data – efficient use of storage space, data security, streamlined information persistence and retrieval workflow, rapid data retrieval through the use of indexes, etc.

The SAS package that enables users to make database connections is SAS/ACCESS. SAS/ACCESS is comprised of a collection of interfaces. Although all interfaces operate in a similar way, each interface is specifically designed to connect to or work with a specific DBMS. Ostensibly the way SAS/ACCESS works is by translating SAS commands into the language of the specified DBMS. The request is then sent in the appropriate language, usually a variant of Structured Query Language (SQL), to the DBMS and the data is returned to the SAS system. There are two primary ways of connecting to a database. One method is by using the SAS/ACCESS library engine and libname statement. The second method to access a DBMS is via the pass-through facility. Each of these methods has its own advantages and disadvantages.

When utilizing the library engine in SAS to interface with the database, performing operations on data is much more straightforward and usually requires less code to access data. A single libname statement is sufficient to access data. In this respect, when specified properly, accessing a table in a DBMS is the same as accessing any other file in a SAS library. The benefit of this is that an end user does not need to know SQL or anything about databases to work with the information they need. The advantage to using the pass through facility is that it more robust for optimizing queries when joining multiple tables or using summary functions. The pass through facility can also make use of indexes placed on columns in the DBMS in order to process queries faster. Furthermore, unlike, the library access method, pass through is able to accept more than just ANSI standard SQL.

Database connectivity technology can be useful is for various types of biomedical data and studies, e.g., microarray studies or in public health. Microarrays are widely used in genomic and proteomic research. Depending on the array type and study being conducted, a microarray can take anywhere from hundreds to millions of measurements from a biological sample. Within a specific study, measurements are usually taken from multiple individuals and thus produce large amounts of data; sometimes multiple gigabytes or even in excess of a terabyte. These data are commonly stored in databases, which provide an efficient mechanism for managing the large data sets. Without using a database interface, in order to extract meaning from the microarray data, usually by performing statistical analysis or data mining, data would have to be exported from the DBMS. This creates two sources of inefficiency. First, it takes up hard drive storage space by creating flat text files (usually the type used by statistical analysis programs) with redundant information. Secondly, it takes time to export data and then import it into an analysis program such as SAS. The export/import process creates a significant bottleneck in the workflow process. Furthermore, exporting and importing data multiple times creates more possibilities for data integrity to be compromised.

An additional benefit is that database interfaces reduce the burden on IT and database maintenance staff. This is because once an IT person sets up a database connection with a library reference engine, subsequently an end user or analyst is able to access information straight from the database without having to wait for IT to export for them a file from a database. Beyond this, there is no recurring need for an IT person to export a new dataset every time an update is made to the database. The analyst can simply access the database via a library in SAS and perform an analysis; thus streamlining workflow. Although there will probably be relatively little updates to the data made with microarray studies, an insurance company, hospital or public health entity may have updates to their database multiple times a day. Directly interfacing with a DBMS will help approach real time data analysis.

The process of physically integrating SAS/ACCESS with a database is quite simple. Once the ACCESS module for a specific database is installed a user can set up a database connection with syntax or point and click. To point and click, right click in the library window and click “new”.

In the engine pull down box, select the SAS/ACCESS engine for the database that a connection will be established with. Then fill in the necessary fields (in this case user name, password, database, server, port as well as other SAS specifiable options). Finally, click ok to create a new library containing the database information. The enable at startup box will automatically load the specified DBMS connection every time a user starts SAS.

The following is example syntax and reference code for syntactically connecting to a MySQL database:

```
Libname user_specified_library_name  mysql user=user_name-
password=user_password database=user_database_name_to-
_connect_to server=mysql_server_name
port=port_to_connect_to_mysql_database;
run;
quit;
```

The italicized portions are generic placeholders which a user would fill in to syntactically make a connection. The; and run and quit commands are parts of the SAS programming language. Tables in a database can be pulled simply using the library and file name syntax. PROC SQL can also be used to join tables and extract complex sets of data.

High Performance Computing Methods:

In addition to more robust data management procedures such as using databases, there are a number of simple coding and programming methods that can dramatically reduce the time needed to perform computationally intensive analyses. With terabytes of biomedical data being generated, the length of time required to effectively analyze this amount of data with traditional single thread processing can be very large. Additionally, many processes in biomedicine lend themselves nicely to parallel processing. For example, peptide, genotype and gene expression microarray data often necessitate repeating analyses tens of thousands or millions of times for each measurement on the chip. Public health data often requires repeating the same analyses on different datasets or separately for different subpopulations. Biomedical data across multiple domains may require running similar analyses under different assumptions or model specifications. Beyond repetition of tasks, any program in which individual pieces can be processed simultaneously or all of the data from one step is not needed for the next step to start (ability to overlap). High performance computing (HPC) methods presented can be of use.

Since most new computers today make use of multi-core architecture, using the methods presented will help to make full use of the resources available to researchers without requiring a large financial investment in hardware. The sections below outline some of the HPC implementations available in SAS. In SAS, pipeline parallelism, multi-core parallel processing and distributed grid processing are all features of the SAS/Connect package. Much of the syntax between these three processes is structurally quite similar. All three can be mixed and matched to best suit the needs of a given project. Additionally, all of these methods can be implemented when reading data from a database as described above. Although this dissertation focuses on SAS, again, other analytic programs are capable of doing similar things. The focus and goal is to illustrate a number of methods that are becoming more commonplace in analytic packages that can be highly useful to biomedical researchers without the need for a highly technical programming background.

Pipeline Parallelism:

The first of the high performance computing related methods we will explore is pipeline parallelism. In its simplest form, pipeline parallelism uses TCP/IP ports on a computer to ‘pipe’ output from a process or step to a subsequent process or step in an analysis.

Pipeline parallelism provides a number of advantages. First, and most obvious, is that this reduces intermediary writes to a hard drive. This is significant because writing large quantities of data to a hard drive and having to subsequently read it, is often one of the slowest steps in an analysis. This is because hard disk input/output (I/O) is usually orders of magnitude slower than processing data in memory. A second benefit of pipeline parallelism is the ability to process sequential processes in a more parallel fashion. Often times subsequent steps in an analysis do not need all of the information from the previous step to start working. For example, on a peptide microarray, thousands of regression models may need to be run to analyze the significance of each peptide and a subsequent step may be to merge all of the results into a single data file. The merging or union of the individual results step does not need to wait for the thousands of individual regressions to finish before starting. By using pipeline parallelism, regression outputs as they finish can be piped directly into a data step merging the results.

Despite the benefits of pipeline parallelism, there is a cost associated with it. For each pipe, there is an associated signon and signoff process. Signon and signoff commands initiate and stop, respectively, links between local and a remote SAS sessions. Actions such as data steps piped to a sort procedure can be accomplished very efficiently with pipeline parallelism. Although looped regression equations and other more complex models can sometimes can benefit pipeline parallelism, the tradeoff between traditional read and write times need to be balanced between signon and signoff times. Generally, the larger the output form a given process, the greater the increase in overall performance will be. This is because the cost of a signon and signoff tasks relative to the overall process is reduced. With process producing large output, piping can often be more efficient than other methods of aggregating data; such as merging the data afterwards or a loop to run multiple SQL union procedures. Also, output from one process can be used as input in a subsequent process when building complex algorithms. A number of factors such as model complexity, amount of data being piped, overall memory utilization and overall computer load all likely play a role in determining the overall performance of piping; especially when piping output from one analytic procedure into another.

In the sample syntax below, three regression equations are performed and the output is piped into a single dataset as they are completed. The options autosignon=yes command is used to have SAS automatically open a TCP port for piping rather than having to specify signon task1, etc. The rsubmit statement is the basic SAS command for remotely submitting a command, e.g., to open a port, to a specific processor or another machine. Wait=no tells SAS to execute the command immediately. Sysrputsync=yes is used if one was to nest this inside of a macro. This ensures that the macro variables would be updated. The SASCMD is a command used to specify options related to the remote submission. Each rsubmit statement is closed with an endrsubmit statement. Everything nested between these two bocks is what is remotely submitted.

```
Optionsautosignon = yes;

        rsubmit task1 wait=no sysrputsync=yes SASCMD="!SASCMD";
                libname out1 sasesock":9001";
                libname simtemp "c:\simtemp";

                proc reg data=simtemp.ttest COV OUT
                OUTEST=out1.tstats1 tableout MSE;
                model class = score1;
                run;
                quit;
        endrsubmit;

        rsubmit task2 wait=no sysrputsync=yes SASCMD="!SASCMD";
                libname out2 sasesock":9002";
                libname simtemp "c:\simtemp";

                proc reg data=simtemp.ttest COV OUT
                OUTEST=out2.tstats2 tableout MSE;
                model class = score2;
                run;
                quit;
        endrsubmit;

        rsubmit task3 wait=no sysrputsync=yes SASCMD="!SASCMD";
                libname out3 sasesock":9003";
                libname simtemp "c:\simtemp";

                proc regdata=simtemp.ttest COV OUT
                OUTEST=out3.tstats3 tableout MSE;
                model class = score3;
                run;
                quit;
        endrsubmit;

        rsubmit task4 wait=no sysrputsync=yes SASCMD="!SASCMD";
                libname in1 sasesock":9001";
                libname in2 sasesock":9002";
                libname in3 sasesock":9003";
                libname simtemp "C:\simtemp";

                data simtemp.final_merged;
                set in1.tstats1 in2.tsats2 in3.tstats3;
                run;
                quit;

        endrsubmit;

signoff task1;
signoff task2;
signoff task3;
signoff task4;
```

Within the rsubmit block, we need a unique name for each remotely submitted process. In this case, task1 task2, andtask3 are used. The libname statements are ostensibly what define the syntax as pipeline parallelism. A libname needs to be set to specify the TCP port SAS will open. This is done with the sasesock “:xxxx” command. The name of the library is arbitrary but out1 out2 and out3 were used here because this is the step in which we were outputting results. When processing remote statements, SAS does not natively inherit libraries from the base SAS session so libraries used in a rsubmit block must be manually specified. This is done in the second libname command in each rsubmit block; although the ordering of libname statements is arbitrary. Output files need to be specified to go to the library associated with the TCP port. In this example, the regression coefficients are output via the outest command to out1.tstats1 and so on. In the rsubmit4 block, the aggregation takes place. This block works in a highly similar way to the others. The only major difference is that the output library ports from the other rsubmit blocks are defined as input libraries in this rsubmit block. The naming is arbitrary of the library but the ports used in above blocks must be used in the later block of SAS is to make a connection to use the incoming data. After the final rsubmit block, a signoff statement is needed to close the port. There is not autosignoff option. It is important to note that the number of rsubmit blocks should not dramatically exceed the number of cores a computer has. The role of cores in this syntax will be discussed more below in the section on multi-core parallel processing. Although we present sample code for a t-test, piping is often more efficient when processing large amounts of output. This is because the associated input from a pipe can become cumbersome to program when a large number of repetitive tasks are done. We present piping in the context of a statistical test to demonstrate the robustness of the procedure; especially since the use of piping in other contexts such as data sorting is well documented in the SAS manuals and literature.

Multi-Core Processing:

Multi-core parallel processing is perhaps one of the most powerful tools to aid in processing large amounts of biomedical data. Expanding the number of processors and local system resources is often referred to as scaling-up. Although new programming languages and tools are starting to place more of an emphasis on threaded and multi-core programming, current languages in wide use such as Java are often difficult and cumbersome to perform multi-core and threaded programming. SAS/Connect is a simple and intuitive package that allows for easily parallel programming. The SAS/Connect rsubmit statement makes the task of creating multiple threads to submit to a unique core very simple. Given the size of many biomedical datasets utilizing all of the cores in a computer can dramatically reduce processing time. The exact amount of reduction will depend on the amount of the overall analysis that is able to be parallelized. As the amount of the program which can be parallelized increases, the time reduction from parallelization increases. Regardless of how much of the program can be parallelized, the ability to have a 2-8x increase in number of available processors provides a substantial benefit; especially when considering the low cost of multi-core computers today.

Below is syntax which demonstrates the use of multi-core parallel processing. In the syntax below, two processors are used to perform 10,000 regression models with 5,000 per processor. A SAS macro and a do loop is used to iteratively loop through different regression equations. The syntax for creating multiple threads and making use of a multi-core PC is quite similar to that used in the pipeline parallelism. This is because each different task in piping is a separate thread and thus the reason why it is not recommended for the number of threads to exceed the number of processors. All of the syntax nested between a rsubmit and endrsubmt block will create a new thread that is sent to a specific processor. The names task1 and task2 are arbitrary but must be different for each rsubmit statement and an associated signoff of each rsubmit is necessary. In this example, the rsubmit statements will signoff as they finish. However, the command _waitfor_=all; can be used to have SAS wait for all of the processes to finish before continuing. Here, we specify the library of files being used for analyses overtly in the libname statement. There is an inheritlib command which will allow SAS to inherit libraries

More information can be found in the SAS Macro language manual about the specifics of macro programming. Macros are user defined mini programs or blocks of code that can be reused. However, in the example below, the block between %macro regs() and %regs(); is a single macro. The %mend ends the macro program. The block between %do and %end is the do loop. The ods output statement inside the do loop tells SAS to output parameter estimates to the file test.paramsx&i. In the do statement the counter variable i was used. Everywhere in the macro that &i appears, SAS substitutes the value of i for that pass through the loop.

```
Options autosignon=yes;

        rsubmit task1 wait=no sysrputsync=yes SASCMD="SAS";
                libname test 'c:\test';
                %macroregs();
                %do i = 1%to5000;
                PROC reg data=test.test1;
                model y=peptide&i;
                ods output ParameterEstimates = test.paramsx&i;
                run;
                quit;
                %end;
                %mend;
                %regs();
        endrsubmit;

        rsubmit task2 wait=no sysrputsync=yes SASCMD="SAS";
                libname test 'c:\test';
                %macro regs();
                %do i = 5001%to10000;
                Proc reg data=test.test1;
                model y=peptide&i;
                ods output ParameterEstimates = test.paramsx&i;
                run;
                quit;
                %end;
                %mend;
                %regs();
        endrsubmit;

signoff task1;
signoff task2;
```

Distributed and Grid Parallel Processing:

Beyond single multi-core parallel processing on a single PC, the syntax can easily be extended to remotely submit a multi-core parallel program to a remote machine. When utilizing other computing resources networked to a host computer, this is often referred to as scaling out. A simple and inexpensive way to scale out is to use other computers (remote machines) running SAS. The remote machine can be a server or another PC running SAS. High performance computing clusters (HPC) are a tremendous resource for analyzing biomedical data. However, there is a large cost often associated with HPC clusters and using them requires specialized knowledge . Many HPC clusters require the code to be submitted in languages such as C or Fortran and via a secure shell program. Given the complexity of many of today's statistical analyses, programming, for example, a mixture model using maximum likelihood estimation with robust standard errors is extremely complex. Beyond the difficulty, the time needed to program such a model in C or Fortran could be prohibitive. Furthermore, licensing for software on HPC is often very expensive. Often times, research institutions or labs will have SAS on multiple PCs. Being able to remotely submit programs to other PCs will allow users to approach the processing power traditionally only available to researchers with access to HPC clusters. For example, 3 Intel I7 PCs could theoretically provide more than 200,000 CPU hours (3 x 8 x 24 x 365 = 202,752) of processing time per year.

Below is syntax for submitting programs to another PC running SAS. In this program, the multi-core parallel program is submitted to another PC and the results are sent back to the host machine. In addition to the syntax provided, the SAS Object Spawner needs to be running on the remote machine.Spawner.exe is included with SAS/Connect.

```
Filename rlink "C:\Program
Files\SAS\SASFoundation\9.2\connect\saslink\tcpwin.scr";
%let node=XXX.XXX.XXX.XXX;
signon remote=node;
libname test "c:\test";
rsubmit remote=node wait=no;
libname test 'c:\test';

proc upload inlib=test outlib=test;
proc download inlib=test outlib=test;

options autosignon=yes;
```

```
    rsubmit task1 wait=no sysrputsync=yes SASCMD="SAS";
        libname test 'c:\test';
        %macro regs();
        %do i = 1%to5000;
        Proc reg data=test.test1;
        model y=peptide&i;
        ods output ParameterEstimates = test.paramsx&i;
        run;
        quit;
        %end;
        %mend;
        %regs();
    endrsubmit;

    rsubmit task2 wait=no sysrputsync=yes SASCMD="SAS";
        libname test 'c:\test';
        %macro regs();
        %do i = 5001%to10000;
        Proc reg data=test.test1;
        model y=peptide&i;
        ods output ParameterEstimates = test.paramsx&i;
        run;
        quit;
        %end;
        %mend;
        %regs();
    endrsubmit;

signoff task1;
signoff task2;

endrsubmit;
signoff node;
```

The filename rlink provides the location of the file tcpwin.scr. This is a script which tells SAS how to signon to a remote PC. The path to the file needs to be provided. The %let statement specifies the local IP address to the remote pc. The signon command remote=node is what tells SAS to signon to the remote PC. In SAS, a signon remote= command must have a SAS variable instead of an IP address. The %let command sets node equal to the IP address. The syntax node is arbitrary and anything could be used. For example, PC1, PC2, PC3 could be used if submitting multiple blocks to multiple PCs. The signon process is ended with and endrsubmit and signoff statement. The signoff statement must specify which PC to signoff of; in this case the PC named node.

As within an rsubmit statement, the remote SAS session on another PC will not inherit the host system's libraries. Therefore, the library needs to be defined. Here we specify the library test to be located in c:\test. The location of this file is on the remote PC not the host PC. A database connect could be used here. However, if a database is not used the files for analysis either need to be manually copied into the library on the remote PC or uploaded via syntax. PROC upload is specified after a signon statement and is used to upload a complete library (or specific files in a library by using a where clause) to a remote PC. The inlib is the library on the host pc and the outlib is the library on the remote PC. When processing files on the remote PC, SAS will save files on the remote PC. The files will often need to be brought back to the host PC for aggregating. This is accomplished via the proc download command. In proc download, the inlib is the library on the remote PC and the outlib is the library on the host PC that files will be copied to. Once complete, the files will be available for processing or viewing on the host pc.

We recommend using a local area network for processing for a few reasons. First, using local area network IP addresses reduces the complexity which can sometimes be associated with external firewalls. Secondly, local area networks often have much faster transfer rates. New PC's often have a gigabit per second transfer rates for hard wired local area networks. If a large amount of data is being transferred, having the fastest possible network connection between computers will minimize lag.

Results:

To test the gain in performance for multicore and multicore distributed processing we use a Monte Carlo simulation to study the classical Behrens Fisher problem. In a Monte Carlo simulation, data is generated and a specific test is conducted a large number of times. The Behrens Fisher problem is a statistical debate without an analytic solution relating to the effect of unequal variances on the T-Test. In this simulation we study 3 sample sizes each with 3 different variances for a total of 9 conditions per replication. We replicate the simulation 10,000 times. In addition, SAS runs an equal and Satterthwaite unequal variance T-Test in the standard proc ttest procedure. This results in 90,000 executions of the proc ttest procedure and 180,000 tests being conducted. The output from each run is saved and aggregated using the SQL union operator.

We test 3 different uses of multicore processing to illustrate the performance gains above the baseline of a single core process without parallelization. The machines used in this analysis were PCs each with one 2.5 GHz Xeon processor and 16gb of ram. PCs are connected on a gigabit Ethernet. First, a baseline test was run using a single thread and core to process the entire simulation. In the second study all 4 cores were used on a host PC to run the simulation. Both the T-Tests and SQL unions were split into 4 equal parts. The third approach used both a local PC host and a second remote machine to process the simulation resulting in a total of 8 processors being used and the task being split into 8 equal parts. The same dataset was used for these performance studies. The benchmark test results are summarized in Table 1.

Table 1 HPC Execution Times for the Simulation Benchmark Test

	Single Thread	4 Threads One PC	8 threads on 2 PC's
T-Test's Only Time	18 minutes 28 seconds	5 minutes 39 seconds	2 minutes 53 seconds
Total Time (including I/O)	87 minutes 6 seconds	28 minutes 14 seconds	12 minutes 8 seconds

In this simulation we notice that the total execution time for 8 threads on 2 PC is less than half the time of 4 threads on a single PC. However, this is one example of a common occurrence encountered when working with large datasets in biomedicine where breaking tasks down into smaller pieces can have an nonlinear and higher order increase on performance. In this simulation we merge the results using a loop and an SQL union procedure. By breaking the task down into smaller pieces, the time to process each join is substantially reduced. This is because the load and write time for smaller files is significantly faster. This effect is especially apparent towards the upper end as more results are added to the merging file and the size of this file grows. Although there are likely more efficient ways to integrate the simulation results than presented here, we chose this method to illustrate the fact that while parallel and distributed processing can dramatically improve performance. Careful attention to program design and thinking about the entire workflow process are often significant moderators of performance gains.

The advanced multi-core processors on the market today provide vast increases in the potential computing power available to researchers. Multi core processors and technologies such as Intel's Hyper threading (which ostensibly allows for 8 simultaneous threads on a quad core chip) can be procured relatively inexpensively. A competent quad core desktop as well as one with Hyper threading can be procured for at or under $1,000. By using the parallel processing and distributed computing infrastructure in SAS researchers can leverage the processing power of modern computers and expand their research with relatively little financial expenditure.

210,240 CPU hour per year are theoretically available from 3 new Intel based PCs with Hyper threading technology (8 cpu hours per pc x 24 hours/day x 365 days/year x 3 pcs = 210,240 cpu hours/year). Assuming a total cost of $5,000 for purchasing 3 such PC, including monitors and ancillary equipment such as networking supplies (a purposely high estimate), SAS licensing not included, results in approximately 0.024

cents per CPU hour if the total cost of the computers and equipment purchased is fully depreciated in a single year. However, it is likely that the computers and equipment would last considerably longer. Cloud based computing systems such as Amazon EC2 are at a minimum 5 times more expensive per CPU hour before factoring in the cost of data transfer and hosting. Additionally, SAS cannot be run on many cloud based clusters and the licensing to run SAS on a HPC supercomputing cluster is an additional cost. While SAS licensing costs are quite variable depending on a number of factors such as the packages chosen, we expect that the additional cost to license SAS on a second or third PC would be cumulatively less than the cost to purchase licensing to use on a HPC supercomputing cluster.

However, there are still clear benefits using HPC clusters which are undeniable. For one, the entire 210,240 CPU hours could in theory be run in much less time than one year on a large HPC cluster. Additionally, HPC clusters are also likely to have more access to memory for complex models which may not always be available to researchers using high end desktop computers. Nonetheless, with new computers able to support 24+ GB of memory, we suspect that for many applications the need for significantly more memory will be limited to highly complex and specialized cases; and cases requiring such capacity will be so large and complex as to already be beyond the scope of this dissertation.

Discussion:

The vast amounts of data being generated in biomedicine and the health care industry today are generating a number of analytic and technical challenges. One major problem is how to effectively manage and analyze the data in a timely fashion. The techniques and methods including database integration, pipeline parallelism, multi-core and distributed grid computing can help dramatically increase the speed of analyzing biomedical data. As the size of biomedical data sets grow, e.g., higher density microarrays or high-throughput genomic sequencing, understanding how to maximize the efficient use of available computing resources will only become more of a challenge. In addition, being able to increase the speed and performance of analysis of biomedical data, these processes will aid in increasing efficiency, streamlining workflows and promoting breakthroughs in biomedicine which help to ameliorate the quality of life for millions of individuals. The relatively inexpensive cost of new computers may for many researchers provide a distinct cost advantage over other HPC options or allow researchers to maximize the resources already at their disposal. Although the methods described here will not completely surpass traditional HPC cluster, we believe that these methods provide a significant advantage in terms of ease of programming and enabling inexpensive access to computing power that will be a valuable resource and appealing alternative for many researchers in the biomedical domain.

Conclusion:

This chapter demonstrates a set of useful computer science tools in a format (SAS) that is familiar to many analysts. To further illustrate with more concrete examples of how these methods are useful, each subsequent chapter gives an explanation of how these methods were used to facilitate the research. Utilizing these methods will help reduce storage needs and processing time for large biomedical datasets or of complex iterative models. Such performance gains will increase the speed of biomedical data analysis as well as the speed of biomedical research more generally.

CHAPTER 3

MONTE CARL SIMULATIONS

Chapter Overview:

This chapter presents the use of Monte Carlo simulation methodologies to help pick the most appropriate model which minimizes bias and error due to violation of model assumptions. While Monte Carlo methods are generally applicable to all violations and for testing new model estimation routines, this chapter specifically focuses on violations of homogeneity of variance. This is because homogeneity of variance is among the most common violations in biomedical data and can easily lead to dramatically incorrect type 1 and type 2 error rates. The real world example involving immunosignature data shows that this violation can easily result in type 1 error rates in excess of 60%; an order of magnitude higher than the standard 5%. Additionally, given the size of biomedical datasets, there can be a range of violations across the dataset stemming from no violation in some cases to massive violations in others. As a result, sampling methodologies are proposed as a computationally efficient tool for screening the extent of violations and to help inform the design of a Monte Carlo simulation that will maximally represent the structure of the dataset.

Problem Abstract:

Statistical models are based on a number of assumptions such as homogeneity of variance. When model assumptions are violated the models do not yield accurate or predictable results. Additionally, in the case of heterogeneity of variance or with complex iterative models such as those which utilize Maximum likelihood, there is no exact analytic solution that can be derived for how the violation will impact model performance. As a result, determining the most appropriate correction is not always clearly defined and is difficult even in the univariate case. This challenge is made significantly more complex by modern biomedical data such as microarrays where there are thousands or millions of comparisons need to be performed. As a result, Monte Carlo Simulations and sampling methods are advocated to help pick the model which that provides the most accurate results across the entire range of the data.

Background:

The results of the analysis of complex biological and biomedical data are not always correct. This is not because of any malicious intent by researchers to make the results of their analyses incorrect but rather because statistical and mathematical models used to conduct the analysis are sensitive to the assumptions underlying their design. Thus, just because a computer gives a researcher an answer does not mean it is the correct answer; in the same way a student punching numbers into a calculator is not guaranteed to come out with the correct answer just because he or she used advanced technology. In order to extract meaning from analyses of biomedical data, whether we are using simple or complex modeling techniques, we need to ensure that the most informative analyses are being done.

A Monte Carl simulation is a study in which a large number (usually thousands) of datasets are randomly generated with a given distributional property and a statistical test is performed on each dataset. The results of the test are aggregated and the performance of the test can be studied. Since there are many cases such as unequal variances and iterative models in which no exact analytic solution can be precisely derived, Monte Carlo simulations have become the defacto gold standard for understanding model performance (Fan 2002).

The challenge of big data has made the task of picking the correct model exponentially more difficult. Beyond the single test or comparison case, modern biomedical datasets often have tens of thousands or millions of comparisons which need to be investigated; specifically microarray datasets. Given the difficulty in making this decision for one comparison, it is vastly more complicated to pick a correct method for use across thousands or millions of contrasts. There is no single method which is ideally suited to all cases (Keppel and Wickens 2007) (Cohen, et al. 2002). Given the natural variability by chance alone across thousands or millions of contrasts, without looking at the data in a thorough way, it is ostensibly impossible to have any idea which method is the most accurate across the state space of the dataset. As a result, the use of sampling methodologies to gain an understanding across of the range and magnitude of violations across entire datasets combined with Monte Carlo simulations to pick the best method is strongly advocated.

Statisticians have been debating one such case of violation of statistical model assumptions known as the Behrens Fisher problem for more than 80 years with no definitive solution yet to emerge. A Behrens Fisher problem arises when trying to estimate the difference in two means when groups have unequal variances because point estimates, hypothesis tests and type 1 and type 2 error rates can become unreliable (Behrens 1929) (Seock-Ho and Cohen 1995). While there are many approaches which have been proposed, many of the classical methods such as Fisher's fiducial theory (R. Fisher 1935), Jersey Neyman and Egon Pearson's sampling proposal (Neyman and Pearson 1928) or a Bayesian method proposed by Harold Jeffreys (Jeffreys 1940), all of these solutions tend to give differing answers; especially with small sample sizes (Seock-Ho and Cohen 1995).

A similar problem arises in linear models such as Student's T-Test when violations of normality exist. In 1960 John Tukey noted that there are multiple cases in which normality can dramatically bias confidence intervals, effect size measures and reduce power (Tukey, A survey of sampling from contaminated distributions 1960). Tukey and McLaughlin proposed a method of trimmed means (Tukey and McLaughlin 1963). As with the Behrens Fisher problem, no single best method has arisen because a closed form analytic solution does not exist (Seock-Ho and Cohen 1995) (Wilcox 1995).

One method which has become widely used to estimate solutions when an exact analytic solution does not exist, the computation time for an exact solution is excessive and to understand the behavior of statistical models is a Monte Carlo simulation. In the simplest form of a Monte Carlo simulation, many datasets (often thousands with modern computer experiments) are randomly generated and a statistic is tested on each set. The results are then aggregated to obtain an approximate estimate of model behavior.

Although the first formal publication linking repeated random sampling to the term Monte Carlo was by Nicholas Metropolis and Stan Ulam in 1949 stemming from their work on the Manhattan project (Metropolis and Ulam 1949). One of the earliest uses of a Monte Carlo method was an 1872 report by Asaph Hall in the Journal Messenger of Mathematics (Hall 1872). Hall reports Captain O.C. Fox randomly throwing wire pins at a wooden board with equidistant parallel lines while recovering from battle wounds during the Civil War (Hall 1872). Captain Fox used the values from repeated tosses to calculate the approximate value of pi (Hall 1872).

William Sealey Gossett actually used a method ostensibly similar to a Monte Carlo method in much of his early work to validate his theoretical ideas about the T-Test and distributions of correlations. Many of Gossett’s works used random sampling methods in a similar way to which modern researchers use Monte Carlo simulations (W. S. Gossett 1908) (W. S. Gossett 1908) (W. S. Gossett 1921).

Again, the modern terminology for Monte Carlo experiments came out of work by Stan Ulam, Nicholas Metropolis and Jon Von Neuman from their work at Los Alamos National Laboratory in the 1940’s from their work on nuclear weapons development (Metropolis 1987) (Metropolis and Ulam 1949). During World War II,one of the earliest computers called ENIAC was originally built and used for nuclear research. Because of the new found ability to compute numbers more rapidly, Ulam suggested resurrecting older statistical ideas which had been brushed aside because of the computational time intensity and difficulty (Metropolis 1987).

Along with Metropolis and Von Neuman, the decision was made to emply ENIAC and statistical methods to model neutron multiplication and diffusion in fissionable material. This was important because, at the quantum level, there is inherent randomness and the complex geometry inhernt in the design of nuclear reactions makes modeling neutrons difficult. When a block of fissionable material is compressed to a sufficient state that it reaches critical mass, a nuclear chain reaction is started. As atoms are split or combined (depending on the type of reaction) neutrons are released which then split other atoms increasing the energy yield of nuclear reactions. The team selected a random distribution of neutrons surrounding a spherical core of fissile material with a random velocity and then tested the path and history of a neutron. This was repeated numerous times until a statistically valid model was generated (Metropolis 1987). When working to develop the Monte Carlo method Metropolis suggested the name Monte Carlo in part related to an uncle of Ulam who was always borrowing money to go to the grand casion's and Monte Carlo (Metropolis 1987). The casino games are ostensibly games of chance, which is related to the random sampling or generation of data, the name Monte Carlo stuck (Metropolis 1987).

In the half century since the formal development of Monte Carlo methods by Metropolis, Ulam and Von Neuman, the method has found wide spread use in many areas of statistics and mathematics. These range from estimating differential equations, entire statistical methods such as a Markov Chain Monte Carlo based on the Monte Carlo method as well as the enormous use to validate statistical methods (Fan 2002). There are literally thousands of statistical articles which use Monte Carlo simulations to test the effectiveness or validity of a given statistical method. In fact, the approach is so prevalent that entire Monte Carlo packages are intergrated into advanced statistical software such as Mplus (Muthen and Muthen 2011). While Monte Carlo simulations are often used to test the performance of statistical models in methodology papers, they are rarely if ever used by researchers to pick the best model for their given experiment.

Given the frequency, or almost certainty with which biomedical datasets violate major assumptions of classical linear models, performing a basic Monte Carlo to pick the best method or correction (ie. T-Test versus Satterwaithe correction etc) should be as commonplace as background subtraction, normalization and transformation in microarray processing.

Metropolis, Ulam and Von Neuman were able to resurrect ideas from Gossett and others about random sampling because of computers. In the decdes since their early work at Los Alamos, Monte Carlo simulations have become commonplace to estimate partial differential equations, as part of Markov Chain Monte Carlo methods and to test statistical methods. In statistics, Monte Carlo simulations have historically been intensive research problems that even with computers could take months or years. However, microprocessors and high performance computing methods, such as those advocated in chapter 2, including parallel procesing and distributed grid processing the time can now be measured in hours or minutes; if the question is sufficiently focused.

Ensuring the test conducted is given the expected results or performing under the expected parameters (ie. 5% type 1 error rate) is critical to extracting meaning from complexity in biomedical datasets. However, the process is made more difficult by big data. This is because the question is often one of thousands or millions of comparisons. As a result, the question is not simply of which model is best for a single comparison but rather which will perform most robustly across the range and surface of the entire dataset.

It is common practice in text books to suggest that statistical texts advocate running tests of model assumptions as part of a standard workflow (Keppel and Wickens 2007) (Cohen, et al. 2002) (Tabachnick and Fidell 2006). SAS implements a number of test procedures to test equality of variances. In proc ttest, the Folded-F method is a default output. PROC GLM for running regression or Analysis of Variance models offers a number of tests including Levene’s and the Brown Forsyth test (SAS Institute 2011). Both Satterthwaite and Welch robust tests are implemented in proc ttest and proc glm respectively. In fact, the Satterthwaite correction is generated by default in proc ttest.

There are also many other ways beyond a t-test and a one way ANOVA to test whether there is a mean difference between two groups. A linear regression equation can use dummy codes (which will produce equivalent parameter estimates to ANOVA when grand mean effect coded since ANOVA is a special case of regression) as well as logistic regression. Weighted least squares and robust regression methods are also alternative ways of correcting heteroskedasticity (unequal variances) (Rao, et al. 2010). Additionally, there are multiple estimation methods available for logistic regression models in SAS (SAS Institute 2011).

The fundamental design of many biomedical analyses will logically produce results with heteroskedastic or unequal variances. For example, in a cancer study, if there is a common unerlying biomarker, whether it is an expressed gene, the presence of an antiboy or other marker, a group of cancer patients with the same type of cancer may be expected to show a similar response profile. Normal patients who do not have cancer or the condition in question might be expected to have a wider variation in the observed values of their responses if sampled from the population at random. This is because all of the population variability would be encapsulated in the normal samples whereas only a smaller subset of those who exhibit similar characteristics on a given trait would be observed in the cancer or condition samples. Conversely, it is also possible that normal patients will have a more similar response profile and those with a condition will have a wider variation. This could result in situations in which there is a relatively small homeostatic window and any response outside of that leads to a disease. Also, we may observe differences in variance structures based on sample size. If there is a large difference in variation it may simply be because a larger number of samples was obtained from a given group and the larger number of samples asymptotically led to a more normal distribution consistent with the central limits theorem.

Ostensibly, it is not always clear that we can or should expect equal variances in biomedical studies. Combine this with big data concerns such as testing thousands or millions of variables (genes, peptides etc) and the probability is that there will be a number of observations for which the assumptions are perfectly met and a number where they are egregiously violated; with everything inbetween. This might lead some researchers to suggest computing a different model for each observation or comparison. For example, one suggestion could be to run a basic ANOVA model with a test for equality of variances and if it is met then run the standard ANOVA otherwise run a correction or alternative method.

There are a few issue with running different models for each observation. First, this would pose a massive increase in computational difficulty and run time. Secondly, a p-value, F or T statistic are not measures of effect size and cannot be compared across models (Cohen, et al. 2002). Some researchers might suggest that an effect size measure might be a more approprite method since they can be more easily compared and are less sample dependent than p-values F or T statistics. However, effect size measures also have their own set of model assumptions; many of which are the same as for traditional statistical tests.

The original validation work by E.S. Pearson and others on the Pearson product moment correlation coefficient was done assuming that the correlatin was zero (Pearson 1929) (Pearson 1931) (Rider 1932). The work by E.S. Pearson and others early on demonstrated that the pearson prodcut moment correlation is highly robust when the correlation coefficient was zero or very nearly so, that unequal variances and other violations of normal theory (Pearson 1929) (Pearson, The Test of

Significance for the Correlation Coefficient 1931) (Rider 1932) (Haldene 1949). Later research by Kowalski and others showed, using Monte Carlo simulations in Kowalski's case that as the correlation coefficient increases, the bias of the correlation coefficient also increases (Kowalski 1972). Ostensibly, regardless of what method a researcher uses whether a classical statistical test, effect size measure or data mining technique, they are subject to some underlying assumptions and the premise that they are not likely to hold over the range of thousands or millions of observations is still a concern.

One technical challenge posed by suggesting that researchers check assumptions across entire datasets is one of computational intensity. Simply running all of the comparisons can be computationally intensive to begin with and suggesting more tests beyond a simulation be run multiplies the computational burden. To help alleviate this concern, beyond the high performance computing methods discussed earlier, two suggestions are proposed. First, there is no need to run rigorous equality of variance tests and secondly, the use of sampling methodologies are recommended.

The reason for not running a rigorous equality of variance test such as a Brown-Forsyth test is because it adds to the computational demand and really does not add much information. This is because such tests only tell a researcher if there is a violation and give no details about the magnitude of the violation (Keppel and Wickens 2007). A researcher will then have to run univariate statistics above and beyond the equality of variance tests to get the input parameters to the Monte Carlo simulation. As a result, since univariate tests are going to be performed anyways, it is not necessary to add the step of a formal equality of variance test. However, if the researcher runs such test and notices small differences in the variances between two groups, individual equality of variance tests could be run in the rare even that a big dataset has thousands or millions of comparisons with very tiny differences in variances across all comparisons.

Secondly, the use of sampling methodologies are highly recommended to take a subsample of the data. Sampling methods have been well developed in statistical literature and are easily implemented in many software packages such as SAS; including proc surveyselect, surveyreg, surveyfreq and surveylogistic (SAS Institute 2011). A number of different sampling procedures such as simple random sampling, stratified random sampling or cluster randomized samples can be taken from the dataset depending on the nature and structure of the data. For some complex studies, a multi-stage probability sample may be necessary; especially if the large study comes as part of a larger survey. The book Sampling: Design and Analysis 2nd Edition by Dr. Sharon Lohr provides an unsurpassed discussion and presentation of sampling methodologies and is an excellent resource for researchers (Lohr 2009).

Additionally, the book also emphasizes the use of SAS and provides extensive code examples.

Model:

Since much genetic and proteomic expression data in raw form is generally on a multiplicative scale rathre than a linear scale (assumed by most parametric statistical models), the first step in any genetic data analysis is background correction, normalization and log transformation to make the data as amenable as possible to general statistlcal models (Quackenbush, Causton and Brazma 2003). Because genomic data inherently violates these assumptions and no clear normalization process has been identified to always give a reasonable correction, it is questionable how often standard model assumptions are actually met and no formal meta analysis are known to have looked at the quesiton. As a result, it is imperative that researches test to ensure tha the models being used will produce accurate results.

The procedure of using Monte Carlo simulations to determine optimal statistical tests was used in the immunosignaturing chapter discussed in chapger 5. The simulation work described in this chapter was the genesis behind using a Satterthwaite corrected t-test in the immunosignaturing chapter. The introduction to this chapter primarily focuses on the assumption of homogeneity of variance or heteroskedasticity as is sometimes referred to and type 1 errors (false positives). This is because linear models are usually more robust to violations of normality than they are to violations of unequal variance and multi-level modeling corrections for violations of independence widely agreed upon (Keppel and Wickens 2007) (Raudenbush and Bryk 2001). Additionally, multi-level corrections for violations of independence are discussed in detail in the next chapter.

Immunosignaturing is described in chapter 5 in detail. As a basic summary, immunosignaturing is a microarray based technology for profiling humoral immune responses. Thousands of random 20 mer peptides were selected from a phage library to give broad coverage of human immune responses and are spotted onto a glass slide. Purified sera samples are applied to the array and antibodies, primarily IgG bind to the random peptides. When an antibody binds with a peptide it will flouesce when exposed to a laser; thus giving a measure of binding affinity (Johnson and Stafford 2009).

The study of interest using immunosignatures compared normal, single breast cancer tumors and second primary tumor samples with the goal of differentiating and the three groups as a diagnostic test (Brown, et al. 2011). Before running basic screening models across the more than 10,000 peptides on the array, basic descriptive statistics were estimated and a random sample of the peptides were taken. A simple random sample of 500 peptides was taken and differences in variances for each group was calculated. The sample was taken using proc surveyselect in

SAS. The largest magnitude difference was approximately 6 fold between normal samples and a second primary tumor peptide.

Monte Carlo simulations were run using equal variances, 1, 2, 4 and 6 fold differences between each of the three groups in univariate contrasts. Two datasets with sample sizes equal to those in the study (52 for normal, 98 for single primary tumor and 21 second primary tumor samples). A scaling factor was used to increase the variance of one group by the given factor. Since type 1 errors were the primary interest, the means were simulated to be equal. SAS IML random number generator was used to generate the data. 2,000 tests were run for a standard T-Test, Satterthwaite T-Test, least squares regression, logistic regression as well as maximum likelihood regression and logistic regression with sandwich estimators. With a standard alpha level of 0.5, it is expected that by chance alone 100 tests would be significant. As a result, more than 100 significant results would suggest a greater than 5% type 1 error rate. While not investigated in this chapter, power could also be studied. This would have been done by changing the mean difference of the two groups and then the percentage of significant tests would be equal to power; for the given magnitude difference.

Results:

Table 2 shows that the simulation results for logistic regression outperforms standard ordinary least squares regression and a standard pooled T-Test with respect to type 1 errors. However, both maximum likelihood regression and logistic regression with sandwich estimator robust standard errors as well as the Satterthwaite T-Test both significantly outperform the standard models; especially when the larger group has the smaller variance.

Table 2 Monte Carlo Simulation Results

Trial	Variance	Logistic Robust SE Error	Logistic Type1 Error	OLS Reg Type 1 Error	ML Reg Robust SE Type1 Error	Std T-Test Type1 Error	Satterthwaite Type1 Error
cancer vs normal	10	90	83	182	95	91	89
cancer vs normal	20	108	206	444	125	222	115
cancer vs normal	40	77	238	526	107	263	98
cancer vs normal	60	95	294	632	121	316	111
cancer vs second	10	111	86	210	131	105	104
cancer vs second	20	153	344	770	134	385	103
cancer vs second	40	159	567	1236	141	618	105
cancer vs second	60	141	643	1420	129	710	93
normal vs cancer	10	93	83	194	106	97	99

normal vs cancer	20	94	37	86	111	43	110
normal vs cancer	40	63	14	32	96	16	92
normal vs cancer	60	74	18	40	103	20	101
normal vs second	10	88	80	188	103	94	93
normal vs second	20	119	233	556	120	278	85
normal vs second	40	96	306	760	129	380	100
normal vs second	60	78	349	858	132	429	99
second vs cancer	10	105	91	206	129	103	104
second vs cancer	20	84	4	14	104	7	96
second vs cancer	40	67	1	2	87	1	85
second vs cancer	60	80	3	5	106	12	103
second vs normal	10	101	84	228	130	114	114
second vs normal	20	75	28	74	121	37	112
second vs normal	40	63	4	18	111	9	109
second vs normal	60	54	3	12	102	6	93

The three corrective models show consistently performance with respect to type 1 errors. The Satterthwaite corrected T-Test shows the best results with a highest type 1 error rate of 5.75% compared to 7.95% for the logistic regression with robust standard errors and 7.05% for linear regression with robust standard errors. Additionally, the Satterthwaite T-Test had the lowest range of type 1 error rates. Having a consistent range is desirable because it allows for more consistent interpretation across the range of the data. Also, as can be seen in the non-corrected tests, violations of assumptions can also deflate test statistics (Keppel and Wickens 2007). This shows up as dramatically lower error rates in this simulation. Having error rates as close to 0.05 as possible allows for maximal inference within the general linear framework researchers are used to.

Discussion:

Monte Carlo Simulations are a useful method to ensure the results obtained by a given statistical test are accurate and trustworthy; especially across the range of large biomedical datasets. While the thousands of calculations which must be performed can be computationally intensive and time consuming, the use of sampling methods combined with high performance computing methods will dramatically reduce the time necessary to perform Monte Carlo Simulations. Depending on the number of contrasts needed to be run (ie sample size and variance pairs for example) with the aid of high performance computing methods, simulations such as the one presented in this chapter can likely be perfomed on an average single computer in less than an hour.

While there is a cost associated with adding another step to standard research protocolsl, researchers should consider the cost in the context of a larger research project. At a minimum, excessive type 1 or type 2 errors could yield an uninterpretable dataset. However, corporations intending on commercializing a product based on experimental results could easily lose tens or hundreds of thousands of dollars when future testing fails to replicate inaccurate results. So, while there may be a small cost associated with performing Monte Carl simulations routintely, the cost is likely to be dramatically less than the cost of obtaining inaccurate results.

High Performance Computing Methods:

In this experiment since the data is randomly generated, there is no need to connect to a database because the data is not stored anywhere. However, given the thousands of models which need to be estimated, multi-core parallel processing and grid distributed processing are highly useful for minimizing the processing time necessary to complete such simulations.

There are a number of strategies that can be employed to parallelize a Monte Carlo simulation ad described in this chapter. One is to divide the thousands of replications within a specific model (ie

Satterthwaite T-Test) across multiple cores or machines. A second option is to run all of the replications on a single core but stripe the methods across cores or machines. For example, all of the runs in a T-Test could be run on one core while all of the runs for the Satterthwaite correction could be run on a second core. The choice on how to parallelize the task is largely process dependent. If a number of complex iterative models are being run with a number of simple methods, the simpler methods may finish much quicker than the complex iterative methods thus not making optimal use of available computing resources.

Conclusion:

The results of this simulation suggest that the Satterthwaite T-Test is the best and most consistent test given the structure and nature of the immunosignaturing data. While it is predictable that the normal theory tests would fail, the magnitude with which the someitmes did was not predictable. For example, ordinary least squares regression sometimes produced type 1 error rates in excess of 50%. Without running this simulation, the magnitude of the failure of OLS as well as the fact that logistic regression with robust standard errors had a nearly 40% higher maximum type 1 error rate than the Satterthwaite test would have been impossible to know.

This simulation is the underlying research which led to the use of the Satterthwaite test in chapter 5. Given the complexity and range of observations in modern biomedical data it is highly recommended that researchers make use of simulations to ensure the tests they are using are giving accurate results not only for a single comparison but across the entire range of data in the study.

CHAPTER 4

Multi-level Modeling

Chapter Overview:

This chapter presents the use of multi-level modeling in the real world example of modeling adherence to medical treatment. The ways in which clustering commonly exists in biomedical data is discussed along with the deleterious effects clustering has on model performance by violating assumptions of independence of observations. Multi-level modeling examples are presented to illustrate how to correct for clustering. In the process it is discussed why it is necessary to center variables and for researchers to document how they centered in order to obtain and accurately interpret multi-level modeling results. Additionally, multi-level models estimate new variance parameters. The estimation and interpretation of these new parameters is highlighted as a powerful method for helping to better understand the complex relationships underlying the data.

Problem Abstract:

Clustering is a common occurrence in biomedical data that dramatically increases its complexity. If left uncorrected, clustering can cause a number of problems. These include parameter estimates that can be incorrect in both their sign and magnitude. This often artificially reduces the standard errors and thus leads to inflated type 1 error rates. Multi-level modeling is a widely accepted method for addressing the problems created by clustering. Unfortunately, multi-level modeling is not as widely used in biomedicine as it should be. When multi-level models are used in biomedicine they often omit critical pieces of information necessary to fully interpret their results, such as how the variance was partitioned.

Additionally, beyond simply correcting for a clustering problem, multi-level models are also necessary to truly make sense of the complexity inherent in the data. Multi-level models provide estimates of a number of new parameters, such as variances at multiple levels, which allow researchers to answer questions about the complex structure and nature of the data that cannot be answered with classical models.

Methodological Background:

While T-Tests and classical linear models are highly useful, they cannot answer every question. Computer implementations of Maximum likelihood and the EM algorithm have given rise to an entirely new set of methods in recent years which allow researchers to ask new and fundamentally different questions than they could in the past. Although T-Tests and other generally computationally simple methods are highly useful in many situations, given the increased complexity of the data and our need to ask more intricate questions, newer more advanced models can help researchers extract more information from biomedical datasets.

As with so many other statistics, the history of multi-level modeling can be traced back to R.A. Fisher in his paper, “The Correlation of Relatives on the Supposition of Mendelian Inheritance” (R. Fisher 1918).

While the mathematical foundations have existed for some time, the use of random effect and multi-level models was not highly utilized for years because one problem with Fisher’s early work was that the model would not work with unbalanced designs. Fisher did also invent maximum likelihood but it was the advent of the EM algorithm and computers that made the iterative process for estimating general forms of these models possible (Goldstein 1986) (Longford 1987).

Multi-level models started to gain more widespread acknowledgement in the mid 1990’s and have become fairly common; especially in social sciences and epidemiology. Stulberg et al. 2010 used a multi-level model to control for clustering created by taking measurements across multiple hospitals and Mills et al. 2006 used multi-level modeling to account for regional clustering in the sampling design across geographic regions (Stulberg, Delaney, et al. 2010) (Mills, et al. 2006). Wile there are numerous examples of multi-level modeling being used in premier journals such as JAMA, there are also examples in which no metion of efforts taken to control for possible clustering. Parker et al. 2009 propose a possible gene signature for breast cancer samples but they use samples from two distinct sample types fresh frozen and parrifin fixed formalin embedded (FFPE) which come from 5 different cohorts (Parker, et al. 2009). Across the many different comparisons performed, there is no mention of clustering in the paper. While the probablity does exist that the effect was neglegible, this is highly unlikely. Personal research not discussed in detail in this dissertation found that with similar breast cancer samples from multiple cohorts exhibited intraclass correlations (a measure of clusering) in excess of 0.6 (Seliegman and Brown 2011).

Even though many papers have used multi-level modeling in bionmedicine, one error or omission which often exists is in centering the variables. Enders and Tofigi 2007 note that even in the social sciences where multi-level modeling is heavily used, that “the issue of centering has been discussed in the literature, but it is still widely misunderstood” (Enders and Tofigi 2007). This is equally valid for biomedical research. Neither Stulberg, Delaney, et al. 2010 or Mills, et al. 2006 who take the first step and recognize a multi-level model is necessary, make any mention of centering; which is critical to obtaining accurate parameter estimates in multi-level models.

In a standard regression model, centering is the process of subtracting a constant, often the mean, from all observations (Cohen, et al. 2002). This has the result of making the intercept the expected value at the mean of the data rater than when x=0 (Cohen, et al. 2002). This process only changes the interepretations of the coefficients but in no way changes the significance of the model.

However, in multi-level models, centering is necessary to obtain accurate non-biased parameter estimates (Raudenbush and Bryk 2001). This is because there is a complex relationship between the independent and dependent variables at two or more levels; the individual unit level of observation as well as the grouping or cluster level (Raudenbush and Bryk 2001) (Enders and Tofigi 2007). This leads to a significant problem because, as in standard ordinary least squres regression, the relationship between an independent and depenedent variable is captured by a single variable.

The interpreation depends on how the multiple variances are partitioned. There are two main types of centering: centering within cluster where the cluster mean is subtracted from each observation within a cluster or grand mean centering where the grand mean for the entire sample is subtracted from each score. Both methods produce dramatically different interpretations and parameter estimates. For example, if the question of interest is a level 1(or unit of observation level) such as number of comorbid conditions a patient has, the recommended approach is to center within cluster because this removes all of the variaibility due to the level 2 variable or unit of clustering such as the doctor he/she sees or the hosptial he/she is admitted to. Grand mean centering the level 1 variable such as number of comorbid conditions would yield an estimate confounded with level 2 or cluster level variability (Enders and Tofigi 2007). Conversely, if the question of interest was a level 2 cluster level variable or a cross level interaction, such whether the effect of number of comorbitidies (level 1 unit level observation) of a patient on some outcome depends on the number of physician years of experience (level 2 cluster observation), centering within cluster would make it impossible to estimate because all of the cluster level variability would be gone. Rather a researcher would want to grand mean center in these types of cases. Ostensilby, without knowing how, if at all, Stulberg, Delaney, et al. 2010 as well as Mills, et al. 2006 centered their variables, it is impossible to interpret their model results, parameter estimates and thus conclusions.

While using multi-level modeling to correct for clustering is a useful and necessary procedure, multi-level modeling provides a plethora of additional information beyond simply correcting for the structure in the data which does not meet certain statistical assumptions. Most authors in biomedicine, Stulberg and Mills included do not make use of the additional information and estimates provided generated when running a multi-level modeling. One example of the utility of this which is presented is that this approach allows researchers the ability to estimate the relative amount of variability in adherence to medical treatment for doctors and patients separately. Multi-level modeling is generally underutilized in biomedicine and when it is, it is almost exclusively used to correct for clustering rather than using the additional complexity to help better explain the world around us.

The following section which investigates adherence to treatment illustrates not only the classical use of multi-level modeling for correcting for clustering within the data but also demonstrates how additional sources of information in the model such as the new variance estimates can be very useful in helping untangle the complexity underlying adherence to treatment.

Experimental Study Background:

Adherence to standards of care has been studied for decades across a variety of populations and settings with mixed results. Patient's characteristics (age, sex, ethnicity, marital status) are important in some studies but not in others (DiMatteo, et al. 2002) (Vermeire, et al. 2001) (Martin, et al. 2005). The differences among the studies may reflect unobserved interactions between patient characteristics and different health conditions, omitted effects such as insurance coverage and geographic differences in practice, or may be simply the artifacts of different methods.

Case studies find that patient – physician communication is an important influence on adherence. The odds of a patient adhering have been found to be 2.16 times greater if his or her physician is a good communicator (Zolnierek and & DiMatteo 2009) (DiMatteo, et al. 2002) (Vermeire, et al. 2001). The advantages of the communication studies are, however, achieved at the cost of limiting inferences to small groups of physicians and patients, often in experimental settings. The results presented here are complementary, gathering information on day to day care in non-experimental settings for large numbers of physicians and patients.

We use a large, community wide, multi-payer data set to estimate the extent to which variations in adherence rates among patients within each of 17 groups of health conditions reflect differences in the patient-constant characteristics of primary care physicians or physician-constant differences among their patients. The community is Maricopa County, Arizona which includes Phoenix, the sixth largest city in the United States. The focus on one county, albeit a very large area, minimizes geographic variations in customary care**.** The data are a subset of the data supplied by three commercial insurers, namely: Cigna, Humana and Health Net of Arizona; and the Arizona Medicaid (AHCCCS) system as part of the Phoenix Healthcare Values Measurement Initiative (PHVMI) (Johnson, et al. 2011). The complete data set include rates of adherence to more than 300 guidelines, 58 health conditions and 38 million claims for 918,370 patients. The analysis data include 52,895 patients, 17 chronic conditions and 3,037 primary care physicians who treated the patients. The 17 conditions, which are described in Table 3, are illnesses for which adherence to recommended care can yield significant benefits. Primary care physicians were selected rather than specialists because we assume that PCPs are more likely to have ongoing contacts with their patients.

Table 3 List of Conditions Studied

Condition
Diabetes Care (NS)
CAD (NS)
Asthma (NS)

Cholesterol Management (NS)
CHF (NS)
DMARD Therapy in RA (NS)
ADHD (NS)
LBP Imaging (NS)
Pharyngitis (NS)
URI (NS)
Bronchitis, Acute (NS)
Depression Med Management (NS)
COPD (NS)
Cardiac Surgery (NS)
Alcohol Treatment (NS)
Emergency Medicine (NS)
COPD Exacerbation (NS)

Adherence has been defined as the “active, voluntary, and collaborative involvement of the patient in a mutually acceptable course of behavior to produce a therapeutic result”. (Meichenbaum & Turk, 1987) The comparison of a large number of standards of care for a large population requires a complex array of assumptions and procedures. We selected Symmetry EBM Connect® 7.6 as our software of choice. EBM Connect®, a product of the Ingenix Corporation, identifies gaps between clinical evidence and health care practice with applications for a variety of health care organizations. (Ingenix, Inc., 2008) EBM Connect® compares actual, observed patient care with care indicated by research-based guidelines

Methods:

A two-level hierarchical model is used to estimate the results. As in previous studies the hierarchical model controls for the effect of clustered variables, namely clustering of patients by physician (Stulberg, Delaney, et al. 2010) (Mills, et al. 2006). Correlations between groups of clustered variables reduce variance estimates, leading to inflated test statistics; type 1 (false positive) errors.

We are also interested in estimating the extent of clustering to understand how much of the variance in adherence rates can be attributed to patients versus physicians.

The model includes two parts, the first of which (level 1) is used to estimate the variation in adherence rates among patients, controlling for differences among physicians. Level 2 of the model estimates the influence of differences among physicians on adherence rates controlling for differences among their respective groups of patients. Separate estimates are prepared for each of the 17 health conditions.

Two different specifications of the model are estimated, namely: a model without covariates (the random effects model) and the model with covariates.

Adherence rates are known to vary among different conditions but the variance in individual adherence rates within a specific condition is not well established. We begin our modeling within specific condition groups such as asthma and diabetes. We select physicians who see at least 5 patients for a given condition. The selection process removes outliers, providing a more representative estimate of physician level variances.

Patients are defined as adherent when 80% or more of rule measures were met (Halpern, et al. 2006); (Mallion, et al. 1998); (Lee, et al. 1996).

Random Effects Multi-Level Model:

We use multi-level modeling to study adherence. The first step in any multi-level model is to assess how much, if any, clustering exists and whether or not that amount of clustering warrants the use of a multi-level modeling correction. In our model, factors relating to patient adherence will comprise the level 1 variables while factors relating to physicians will be the level 2 variables. Since patients are nested or clustered within physicians, patients seeing the same physician will likely be more similar in adherence because of factors such as physician-patient communication patterns. Clustering reduces within class error rates because adding one additional case to a study does not add one full piece of information as a result of the correlation between cases nested within a class.

In other words, because individuals seeing a similar physician are likely to be have at least some correlation, (a degree of similarity) knowing something about one patient provides some information about other patients and how they are likely to adhere; part of which may be due to seeing same physician. Subsequently, this phenomenon known as an intra-class correlation reduces the denominator of many regression based statistics and thus false positive rates associated with statistical hypothesis testing.

The first phase of this study is a random effects multi-level model as specified by the following set of equations:

$$Y_{ij} = B0_j + r_{ij} \qquad \text{Equation 1}$$

$$B0_j = Y00 + \mu_j \qquad \text{Equation 2}$$

where: the outcome variable Y_{ij} is a measure of adherence to treatment for an individual i in class j. $B0_j$ is the mean adherence value for class j and r_{ij} is the deviation between and individual's score i and their respective class mean j. Y00 measures the grand mean across all groups and μ_j is the deviation between the grand mean and the mean for class j. This notation shows level 1 and level 2 equations separately. However, the level 2 equation can be substituted into the $B0_j$ term of the level 1 equation to create a combined equation as follows:

$$Y_{ij} = Y00 + \mu_j + r_{ij} \qquad \text{Equation 3}$$

Equation (3) implies that an individual's level of adherence can be accounted for by the physician/prescriber they see as well as some unique individual variance. This model partitions the variance into independent orthogonal level 1 and level 2 components.

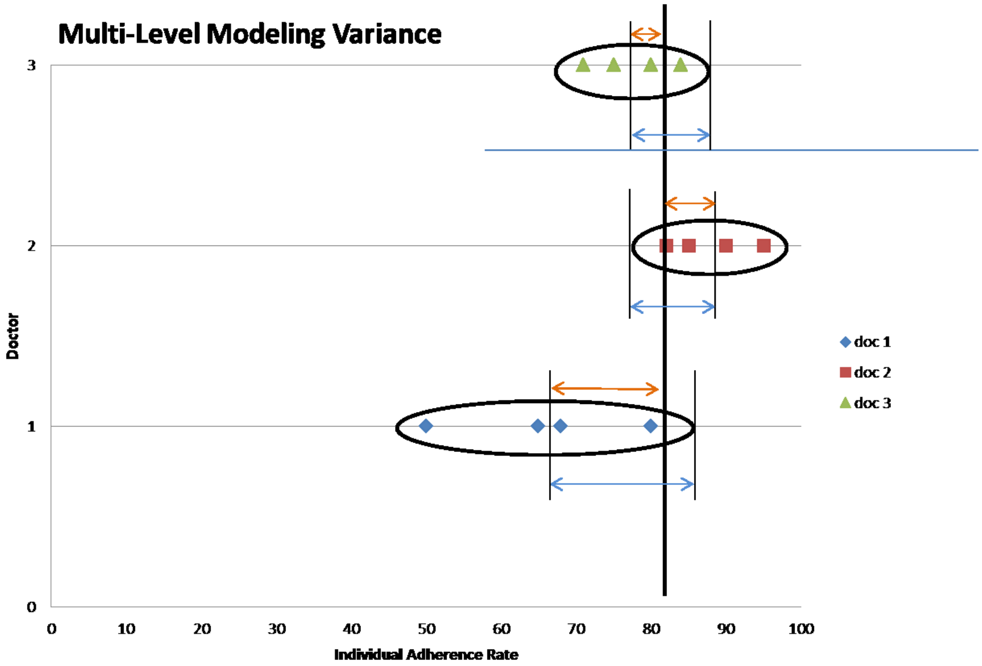
This model can be used to estimate a number of informative factors. One is the intra-class correlation (ICC). The ICC measure quantifies the effect of level 2 clustering on the data. In other words, the ICC provides the expected correlation between two patients' scores from the same level 2 cluster. In this study the ICC measures correlation between the adherence scores of two patients who see the same physician. The ICC is calculated as follows:

$$\text{ICC}= \frac{\tau 00}{\tau 00+ \sigma^2} \qquad \text{Equation 4}$$

The model also estimates parameters τ00 and σ^2. τ00 is the level 2 variance between class means. This variance estimate is between class means because the average value for individual scores i form the input for class j at level 2. σ^2 is the average within class variance at level 1. Taking the square root of the variance estimates τ00 and σ^2 will yield standard deviations and an estimate of the size of the difference between an individual's score and their group mean (the average for others who see the same physician, σ^2) as well as how much difference there is between physicians on mean adherence (τ00). Table 4 shows the ICC, level 1 variance, level 1 standard deviation, level2 variance and the ICC. Notice in table 4 that the square root of the variance estimates give the standard deviations.

From here, assuming we find evidence of clustering and differences between physicians, a number of other variables can be added to the model. Age, severe comorbidities and type of insurance (public or private) can be added as level 1 variables. Physician specialization can be added as a level 2 variable. After adding variables to a model, both the level 1 variance σ^2 and the level 2 variance τ00 can be recomputed. The recomputed values will tell us how much variance at each level was accounted for by adding a given set of predictors. In addition, we can estimate how much variation in level 1 and level 2 is not due to the added predictors in our model.

Figure 1: Multi-Level Modeling Variance Partitioning



In graph 1 represents sample data for the purpose of helping to make the multi-level conceptions of variance more clear. In this graph individual compliance rates are plotted by different doctors. The thick black line is the grand mean or the average compliance rate across all patients. The circles identify more clearly patients seeing a single doctor and the thin black lines at the center of the circles are the mean compliance rate for physicians. The blue lines between the physician mean (thin black line in the middle of the circles) to the line at the end of the circle is indicating the spread of patient adherence rates. This is the variation for patients seeing a single doctor. The average spread or variance across all doctors is the level 1 variance. If there was no level 1 or individual level variance, there would be no circle and all of the patients would have the same adherence rate. The orange line from the physician mean to the grand mean represents the level 2 or physician level variance.

Although there are a number of factors relating to individual adherence, at a more fundamental level, graph 1 shows that an adherence rate for an individual has two components. One part is the natural propensity of an individual to adhere due to intrapersonal factors such as age, sex, number of times they visit a doctor. The second part is the physician they see. Factors such as physician communication and experience may serve to shift an individual's likelihood of adherence up or down. The magnitude of this shift can be conceptualized as the overall width of the orange lines. As physicians pay a larger effect on an individual's likelihood of adhering, the average width of the orange lines (distance between average compliance of all patients a physician sees and the grand mean) gets larger. Factors such as physician communication patterns and years of experience may influence how their patients adhere. Unfortunately because we are using public health data, we do not have data on communication patterns. However, we are still able to measure the magnitude of the effect that physician level (level 2) and individual level (level 1) variables have on the adherence rates of individuals.

Results Random Effects Multi-level Model:

Table 4 represents the findings from the random effects multi-level model. We will start by considering diabetes as an example to illustrate the interpretation of the results. For coronary artery disease (CAD), τ00 has a value of 0.006148 and is the level 2 variance which quantifies between cluster (physician variability). The standard deviation for the τ00 (square root of the variance) at level 2 (within physician) is approximately 7.8% (0.07841). In other words, the mean compliance rate of patients being treated by a given physician for CAD should over repeated sampling on average range between 40.96% and 71.7% (95% confidence interval). Within each cluster (physician) the variance (level 1 variance / sigma squared) is 0.0951. This says that within each cluster (physician) an individual patients compliance will on average differ from that doctors mean compliance by approximately 30.84% (note the square root of a variance is a standard error and the $\sqrt{0.0951}$ = .3084).

Is what this illustrates is that the within class variation is much greater than the between class variation or that there is much more variation between individual patients seeing a physician than there is between the average compliance rates across physicians. Said differently, in the case of diabetes, although the physician bears some responsibility, individuals are far more responsible for adherence; or lack thereof than physicians.

An ICC of .05 is often used to determine whether or not there is a significant effect of clustering. Having an ICC greater than .05 suggests that there is a significant effect of physician or that adherence differs significantly based on which physician an individual goes to.

By looking at the ICC and variance components across multiple conditions (Table 4) a number of interesting trends emerge. First, the level 1 or individual level variance is always much larger than the level 2 or physician level variance. This suggests that like with CAD, factors relating to the individual seem to be a larger driving force behind adherence than factors associated with the physician. The level 1 variances also replicates the findings from previous research showing that individual adherence differs across condition.

As with the level 1 variances, the level 2 variances differ across condition. Interestingly, the level 2 physician level variances and ICC's suggest that difference among physicians in the treatment of conditions such as coronary artery disease, ADHD, cardiac surgery and depression medication management physicians, have little effect on adherence rates. With ICC's less than .05 in these conditions, variations in adherence rates among patients account for such a disproportionately large amount of variance, that the role of the physician is negligible; if even existent.

In the case of ADHD, this makes some intuitive sense. This is because a primary treatment for ADHD today is amphetamine based stimulants. Drugs such as Adderall and Ritalin are controlled substances for which there are very strict treatment guidelines. As a result, it would be unlikely to see large differences between physicians. In the case of cardiac surgery, since heart related conditions are among the leading causes of death in America, there is substantial amount of literature and emphasis on best practices for treatment. Therefore, like with ADHD, the message physicians are giving their patients is not likely to vary too much. However, depression medication is interesting because in many respects it is exactly the opposite of ADHD and coronary artery disease because best practice treatment guidelines are not well defined. Perceptions about treatment and the theoretical construct of depression likely varies a significant amount among both physician and individuals. It is initially unclear as to why there is almost no discernible difference at the physician level. One hypothesis might be that because the individual level variance is large, the physician level variance has little room to play any role. The problem with this hypothesis is that DMARD therapy also has a large individual level variance while simultaneously having one of the largest physician level variances. Furthermore, the physician level variance for ADHD is ostensibly zero. There are other conditions such Pharyngitis and COPD in which the physician level plays a large role. Understanding the causal factors underlying why the physician level variances differ so precipitously across condition is a novel question for future research.

Table 4 Physician Level Variances

Condition	**Level 2 Variance**	**Level 1 Variance**	**Level 2 Standard Deviation**	**Level 1 Standard Deviation**	**ICC**
Diabetes Care (NS)	0.002875	0.02666	0.05362	0.16327	0.09737
CAD (NS)	0.006148	0.0951	0.07841	0.30838	0.06072
Asthma (NS)	0.002845	0.1008	0.05334	0.31752	0.02745
Cholesterol Management (NS)	0.004328	0.02941	0.06579	0.1715	0.12828
CHF (NS)	0.08466	0.1422	0.29096	0.3771	0.37318
DMARD Therapy in RA (NS)	0.04707	0.1992	0.21695	0.44631	0.19113
ADHD (NS)	>.0001	0.2344	>.0001	0.48412	>.0001
LBP Imaging (NS)	0.007291	0.1678	0.08539	0.40961	0.04165
Pharyngitis (NS)	0.085	0.142	0.29154	0.37678	0.3745
URI (NS)	0.04487	0.1141	0.21184	0.33772	0.28235
Bronchitis, Acute (NS)	0.01069	0.1943	0.10339	0.44078	0.05215
Depression Med Management (NS)	>.0001	0.2018	>.0001	0.44921	>.0001
COPD (NS)	0.05198	0.1742	0.228	0.41733	0.22987
Cardiac Surgery (NS)	0.003214	0.06865	0.05669	0.26201	0.04472

Alcohol Treatment (NS)	0.03799	0.05936	0.1949	0.24363	0.39023
Emergency Medicine (NS)	0.03544	0.08143	0.18827	0.28536	0.30327
COPD Exacerbation (NS)	0.03156	0.1298	0.17766	0.36033	0.19556

Beyond the ICC which tells us which conditions have significant clustering, the interesting and noteworthy finding is that while physician's do bear some level of responsibility for lack of adherence, the patients always account for a greater proportion of variability. This finding would not have been possible with simpler models such as a t-test or standard ordinary least squares regression. This is because neither model allows researchers to estimate the parameters necessary to partition the variance between patients and physicians in an interpretable fashion. The next section shows how multi-level models can be used with covariates to understand predictor variables which help explain adherence to treatment. Additionally, the section demonstrates proper centering of variables to obtain accurate and un-confounded parameter estimates.

Multi-Level Model with Covariates:

A random effect multi-level model is a starting place to determine the feasibility of further exploring more complex multi-level models. For example, one reason for not continuing to build a more complex model is if random effect results do not show a significant ICC or level 2 variance component. This is because if there is no intra-class correlation and significant level 2 variance component then there is no need or validity to using a multi-level model. As a result, a subset of conditions were selected for further analyses.

The variables explored were public versus private insurance (insurer type), gender, age, number of patient visits, ethnicity and physician years of experience. Physician experience was centered at the grand mean while patient age and number of patient visits were centered within cluster in order to give an unbiased estimate of the level 1 or individual level effect. Centering patient age or number of patient visits at the grand mean would confound the estimates with level 2 variability. Physician experience was centered at the grand mean because it is a level 2 variable predicting physician mean adherence and therefore centering within cluster is not an option. Insurer type, gender and ethnicity were dummy coded. For insurer type, private was coded zero, for gender male was coded zero and for ethnicity Caucasian was always coded zero. In a dummy coded model, the regression coefficients represent the expected change in the mean from the group coded zero to the group coded 1. For example, if there was a positive coefficient for a gender dummy code, this would mean that on average, females are expected comply more than males by a given amount. The model estimated is as follows:

$$Y_{ij} = Y00 + Y01(\text{physican years experience cgm}) + Y10(\text{patient age cwc}) + Y20(\text{number of visits cwc}) + Y30(\text{patient gender}) + Y40(\text{insurer}) + Y50(\text{Asian}) + Y60(\text{Black}) + Y70(\text{Hispanic}) + Y80(\text{Native American}) + Y90(\text{Other}) + U0_j + r_{ij} \qquad \text{Equation 5}$$

Or

$$Y_{ij} = B0_j + Y10(\text{patient age cwc}) + Y20(\text{number of visits cwc}) + Y30(\text{patient gender}) + Y40(\text{insurer}) + Y50(\text{Asian}) + Y60(\text{Black}) + Y70(\text{Hispanic}) + Y80(\text{Native American}) + Y90(\text{Other}) + rij$$

$$B0_j = Y00 + Y01(\text{physican years experience cgm}) + U0_j \qquad \text{Equation 6}$$

Y00 is the grand mean, Y01, Y10 – Y90 are regression coefficients $U0_j$ are random slope for level 2 and r_{ij} is the level 1 residual. Patient age, number of visits and gender are level 1 variables while physician years of experience is a level 2 variable. The results of the multi-level models do at some point find a significant effect for all of the variables. However, variables are not significant across all conditions. Furthermore, the effect size and trend is equally inconsistent. A Bonferroni correction was made for alpha inflation or multiple testing and our nominal alpha or p-value for significance was set at 0.005. This was done because with p=.05, by chance alone 1 out of 20 tests would be significant. Since we had 10 variables we were testing 10 variables per condition, we reduced the alpha level or nominal p-value for significance to keep the probability of a type 1 error or false positive what it would have been if we only ran one test for one condition. Table 5 shows the statistically significant factors after this correction factor was made.

Table 5 Significant Predictors

Description	**Effect**	**Estimate**	**Standard Error**	**Degrees of Freedom**	**T-Value**	**P-Value**
Alcohol Treatment (NS)	Gender	-0.04814	0.01043	2411	-4.61	<.0001
CAD (NS)	Gender	-0.04927	0.007967	6021	-6.18	<.0001
Cholesterol Mgmt (NS)	Gender	-0.01838	0.004686	6371	-3.92	<.0001
COPD Exacerbation (NS)	Gender	0.06915	0.01607	2576	4.3	<.0001
CAD (NS)	Insurance Type	-0.1204	0.01026	2141	-11.73	<.0001
COPD Exacerbation (NS)	Insurance Type	-0.3468	0.02915	2574	-11.9	<.0001
Diabetes Care (NS)	Insurance Type	-0.02175	0.002145	5.30E+04	-10.14	<.0001
CAD (NS)	Patient Age	-0.00444	0.000306	5860	-	<.0001

					14.53	
Cholesterol Mgmt (NS)	Patient Age	-0.00216	0.000204	6294	-10.57	<.0001
COPD Exacerbation (NS)	Patient Age	-0.01226	0.000778	2552	-15.75	<.0001
Diabetes Care (NS)	Patient Age	0.000923	0.000045	7.50E+04	20.57	<.0001
DMARD Therapy in RA (NS)	Patient Age	-0.00709	0.001354	638	-5.24	<.0001
Alcohol Treatment (NS)	Number of Visits CW	0.01069	0.002016	2348	5.3	<.0001
CAD (NS)	Number of Visits CWC	0.01905	0.002953	5792	6.45	<.0001
COPD Exacerbation (NS)	Number of Visits CWC	-0.00952	0.002106	2535	-4.52	<.0001
Diabetes Care (NS)	Number of Visits CWC	0.00223	0.000175	7.50E+04	12.71	<.0001
LBP Imaging (NS)	Number of Visits CWC	-0.1867	0.01751	2317	-10.66	<.0001

COPD Exacerbation (NS)	Hispanic	1.0194	0.08615	2559	11.83	<.0001
Diabetes Care (NS)	Hispanic	0.09725	0.0119	7.50E+04	8.17	<.0001
Diabetes Care (NS)	Native American	-0.078	0.003468	7.40E+04	-22.49	<.0001
Diabetes Care (NS)	Other Ethnicity	0.006962	0.00139	7.50E+04	5.01	<.0001
URI (NS)	Years Practiced GMC	-0.0032	0.000741	187	-4.32	<.0001
Diabetes Care (NS)	Asian	-0.01431	0.00505	7.60E+04	-2.83	0.0046
CAD (NS)	Hispanic	-0.1812	0.0622	5995	-2.91	0.0036
Alcohol Treatment (NS)	Native American	-0.09684	0.02878	2402	-3.36	0.0008
Alcohol Treatment (NS)	Years Practiced GMC	-0.00263	0.000742	195	-3.54	0.0005
Cholesterol Mgmt (NS)	Insurance Type	-0.02726	0.007537	5339	-3.62	0.0003

Bronchitis, Acute (NS)	Patient Age	0.003572	0.00099	1655	3.61	0.0003

A number of more complex models with higher order terms, such as quadratic trends, interactions, cross level interactions and additional random slopes were studied. However, trying to force more exotic model specification onto all conditions led to problems such as non-convergence need a highly tailored model specification for each individual condition to yield meaningful results. As a result of the highly specific nature of each model, such results are not readily comparable across conditions and are not presented.

Results Multi-Level Model with Covariates:

Beyond the global trends and partitioning variance in adherence between patients and physicians, our models provide a wealth of information about how common covariates differ across condition. Age is the single most prevalent influence on adherence rates. Age significantly influences adherence rates for 9 of the 17 conditions studied. The effect of age on adherence rates is similar to the other significant influences in that it is positive or negative, depending on the health condition being considered. An additional year of age, all else equal, increases adherence rates among patients with diabetes, ADHD and Bronchitis but reduces adherence rates among patients with CAD, Asthma, Cholesterol Management, DMARD therapy, COPD and COPD Exacerbation.

Statistical significance tests indicate that an effect is not random, but very small effects can be statistically significant. It is useful, therefore, to also consider the size of the significant effects. The effect sizes associated with age are generally small, ranging from -0.1 percentage points to +3.7. Except for the maximum estimate (+3.7) which applies to a very restricted age range, the estimates would increase substantially if applied to the multi-year age groups typically used in research on health or health care. Being female has a significant influence on 5 of the 17 health conditions with the direction (positive or negative) varying among different conditions. All else equal, female patients are less adherent than males if they are being treated for CAD, Acute Bronchitis or problems with Alcohol. The estimates range from -2.5 to -4.1 percentage points relative to the adherence rates, all else equal, of males with the same conditions. Females are likely to be more adherent than males if the criteria refer to avoidance of imaging for acute low back pain or appropriate treatment for COPD Exacerbation. The estimates are 3.9 percentage points for lower back pain imaging and 6.3 percentage points for COPD.

The effects of ethnicity are not uniform among the health conditions. Being Hispanic is a significant influence on adherence for only 2 of the 17 conditions, namely diabetes or COPD exacerbation. Hispanic (H) patients with diabetes are, all else equal, more likely to be adherent than White, Non-Hispanic patients (WNH), but Native American (NA) patients are less likely to be adherent than either Hispanics or White Non-

Hispanic (WNH) patients with diabetes. Hispanic patients, all else equal, have adherence rates that are 3.7 percentage points higher than those of WNH. NA patients have rates that are 7.3 percentage points lower than those of WNH. The only significant effect of being NA, other than for diabetes, is for patients with upper respiratory infections, and the effect is positive rather than the negative effect for diabetes. Adherence rates for NA URI patients are 9.9 percentage points higher than for WNH patients.

The variations among different diseases in direction, significance and size of the association between patients' characteristics and adherence rates echoes the lack of agreement in previous research studies concerning the importance of demographics on adherence rates (cited in the Background section of this portion of the report). Our results show that demographic characteristics have a significant effect on adherence, but those relationships are very different for different conditions.

The remaining results refer to one measure of public versus commercial insurance coverage and two measures of physician characteristics. The results for insurers do not allow for interactions between insurer type and adherence rates. That would require estimating separate results for each type of insurer which was beyond the scope of this part of the report. The inclusion of a one-zero variable in which public insurer=1 captures shifts in the intercept of the equation, but not interactions between insurer type and all the other variables in the model.

There is no significant difference between coverage by public versus commercial insurance for 13 of the 17 health conditions. The exceptions are for CAD, Cholesterol Management, COPD Exacerbation and Diabetes for which adherence rates are, in each case, reduced if, all else equal, a patient is covered by public insurance. The estimates range from -3.9 percentage points for Cholesterol Management to -18.1 percentage points for COPD Exacerbation. The results for public insurance may include a selection bias since persons with mental health problems are more likely to be on public insurance.

Two variables, namely years in practice and number of visits for each patient seen by a physician were included in the model to provide some controls for differences among physicians. The interpretation of the association between years of practice and adherence rates, controlling for patient characteristics, is straight forward. The effect of number of visits is, however, subject to uncertainty concerning causal direction, although it does indicate the extent of contact between a physician and a patient. We will, therefore, restrict our discussion of the results to the years of practice variable and treat the visits variable as a pure control.

An additional year of practice by a physician is significantly associated, all else equal, with differences in adherence rates for only 3 of the 17 conditions, namely: CAD, URI, and Alcohol Treatment. The effect of a single year is quite small, less than one to three tenths of one percent, indicating that substantial differences only occur over ten to twenty year differences in physician practices. Thus, one expects the large differences to exist between relatively young physicians and physicians in their sixties or older.

Discussion:

Adherence to treatment has been studied in numerous different contexts across a plethora of settings. Our study takes a slightly different approach in that we ask the question of how much of the variance in adherence is due to physician related factors and how much is due to patient factors. Our use of hierarchical (multi-level) modeling to partition variance between groups in clustered datasets is by no means technically new. However, this method is less commonly used in the study of adherence. Stemming from past literature, in a second set of analyses we also look at the role of previously identified factors relating to adherence. While in the first part of our study, hierarchical (multi-level) modeling was used as a tool to partition variance, in the second part modeling factors on adherence, the use of this method is necessary to obtain statistically valid estimates. This is because clustering in datasets, if not correct for through methods such as hierarchical (multi-level) modeling will lead to dramatically inflated type 1 error rates (Raudenbush and Bryk 2001).

In contrast, many studies tend to focus more on the specific factors relating to adherence. For example, McKinlay et al. examine sources of variation in physician adherence with clinical guidelines and Cabana et al. similarly study why physicians do not follow clinical practice guidelines. In acknowledgement of the fact that physicians often do not follow guidelines, the New England Health Care Institute conducted a study to identify barriers and develop strategies for improving physician adherence to clinical guidelines. The New England Health Care Institute identifies the medical payment system, IT infrastructure, physician culture and the current guideline development process as barriers to physician adherence to guidelines. The Institutes report also identifies a number of potential interventions to improving physician adherence to guidelines. Although we do not explicitly test many of these things such as IT infrastructure or medical payments, we believe that many of these factors would be subsumed within the physician variance component. This is because to the extent that IT infrastructure or medical payments are a factor, differences in adherence related to them would show up in different mean adherence rates between physicians. While these may be important factors which show statistically significant results in studies on their own, our research suggests that the overall magnitude of their effect in the broader context of adherence is likely small because physician related factors globally contribute to a small fraction of the variance in adherence rates. We believe that interventions focusing more on patient centric aspects of adherence would be a better use of finite resources.

In addition to physician factors on adherence, a vast literature spanning decades exists examining patient factors relating to adherence. Dimatteo et al. and Vermeire et al. both provide excellent meta-analyses and reviews of the literature regarding patient factors on adherence (DiMatteo, et al. 2002) (Vermeire, et al. 2001). Our study, like many meta-analyses does not find a clear and consistent effect for many demographic variables such as age, gender, race or even physician years of experience on adherence. Dimatteo et al. and Richard et al. examine the effects of patient non-adherence in health outcomes. Choudhry, Fletcher and Soumerai find that physicians with more years of experience are sometimes at risk for providing lower quality care and may need quality improvement interventions (Choudhry, Fletcher and Soumeral 2005). We find that in 3 of the 17 conditions studied, physician years of experience is correlated with decreased adherence rates. Although we only observe this in a small subset of conditions, it is tangentially consistent with the general findings of physician performance degradation over time.

Many researchers have noted that adherence is also likely a factor of the interaction between physicians and patients (Zolnierek and & DiMatteo 2009). Zolnierek & DiMateo specifically conduct an extensive meta-analysis regarding the effects of physician patient communication on adherence. Decades of research identifies a number of factors on both the patient and physicians sides of adherence as well as factors relating to the interaction between patients and physicians. Our contribution is that we are able to show that while there is clearly a non-ignorable portion of variance associated with adherence is due to physicians, in every condition we studied, a much larger portion can be attributed to the patients. Given the consistency of the finding that more variance can be attributed to patients than physicians in all 17 conditions studied, we suggest placing more emphasis on focusing interventions more on patients than on physicians. While physicians are an easier group to reach for interventions, without a strong intervention that produces large effects, in the bigger picture of overall adherence, such programs may not have the largest overall impact.

The characteristics included in the model (insurance type, age, gender, ethnicity and number of visits, years of practice) account for some of the variance in adherence rates but the amount of overall variance explained is low. In the full models, a pseudo r-squared calculation to measure the proportion reduction in variance never reaches the threshold for even a medium or moderate effect size for the individual level variables tested. The results, calculated for each condition, reflect the fact that much of the difference in adherence rates is determined by the type of condition being treated rather than variations among patients with the condition or among the physicians who are treating them.

Adherence rates, within type of condition are influenced in greater part by differences among patients than by differences among physicians, recognizing that our measures of physician characteristics are limited. Many conditions exist where a specific physician does have a noticeable effect, but even in those circumstances the physician does not appear to be the predominant influence.

High Performance Computing Methods:

The Center for Health Information Research stores archival data in a SAS database. As a result the analytic program is natively connect to the analytic program and eliminates the need for database connectivity programs. Nonetheless, the same inherent benefits are realized in that the redundant write to disks and increased size of excess output flat files is minimized.

Multi-level models are iterative models which utilize maximum likelihood and the EM algorithm. Given the complexity of the data these models sometimes take a large amount of time and number of iterations to converge on a solution. Since the model is iterative each run cannot be parallelized but each different condition can be. Parallel processing was used in a dual core fashion to run the 17 conditions in parallel thus substantially reducing the necessary processing time.

Conclusion:

This chapter demonstrates the effective use of multi-level modeling. Proper use of the model and centering is illustrated along with how new parameter estimates such as level 1 and level 2 variances can be used to address the complexity in modern biomedical datasets. Without using multi-level modeling the parameter estimates from the model would have been dramatically inflated and the researcher would have likely made incorrect conclusions about the relationship of many variables.

Additionally, without the use of multi-level modeling it would not have been possible to partition the complexly tangled variability in adherence to treatment accurately between patients and physicians. This result gives significant insight into a decades old question and suggests that more resources should be allocated to interventions targeted at patients to increase adherence to treatment.

CHAPTER 5

Structural Equation Modeling

Chapter Overview:

This chapter presents the use of structural equation modeling to better address the complexity associated with the presence of unobserved latent variables in a research project. This is illustrated through the real world example of an immunosignaturing study to develop diagnostic tests to screen for breast cancer. The first section shows that while classical tests such as a t-test provide useful information, they are very limited in their ability to fully address the complexity inherent in the data and have no ability to model underlying latent constructs. The second section expands by show how structural equation modeling can be used to better make sense of complex data and directly model latent constructs (in this case antibodies). Additionally, it is shown how not only can we make inferences about the existence of latent constructs but it is also possible to model their relationship to other variables such as how they predict disease outcomes. This chapter also discusses how the measurement model portion of a structural equation model may be used as a diagnostic tool for medical devices.

Problem Abstract:

One final complexity in biomedical datasets is the presence of latent factors where the outcome of interest is not measured directly but rather via a number of proxy observations. This occurs commonly in microarray technologies where single nucleotide polymorphisms are used to represent the effect of an entire gene, single messenger RNA fragments are used to study the expression of an entire gene or peptides are used to study the effects immune response and antibodies. Simple statistical models are only capable of studying the marker variables and have no way to make any inferences about the larger underlying latent or unobserved construct. Structural equation modeling is specifically designed to help answer questions about hypothesized but unobserved latent constructs. Additionally, beyond investigating the presence of latent factors, structural equation models are capable of estimating much more complex models in which hypothesized latent factors are treated as unique variables to help understand their relationship in the broader context of the research study.

It is demonstrated how structural equation modeling can make sense of complex datasets even when the outcome of interest is not directly measured. Applications for diagnostic screening using latent factors are also presented.

Methodological Background:

Structural equation modeling is commonly used in the social sciences today is comparatively underutilized in biomedical data analysis. Despite the relative underutilization of structural equation modeling in Biomedicine today, the historical roots can be traced back to work by the noted population geneticist Sewall Wright’s work on path analysis. In his

1921 paper entitled Correlation and Causation, Wright lays out a covariance modeling approach to path analysis as a method for studying plant and animal biology (Wright 1921). Path analysis is the basic foundation for the structural portion of structural equation modeling models. Later, measurement models based largely off of factor analytic methods were added to what has become the modern day structural equation modeling framework (Bollen, Structural Equations with Latent Variables 1989).

Structural equation modeling provides a robust modeling frameworks that is applicable across a wide range of biomedical data from genomic and proteomic to public health data. Both Dahly and Yu-Kang argue that structural equation modeling is underutilized in many biomedical fields; especially epidemiology (Tu 2009) (Dahly, Adair and Bollen 2009). The general structure of structural equation modeling makes it well suited to addressing many analytic challenges across a wide range of biomedical domains. This is because structural equation modeling allows users to specify anything from a t-test or simple linear regression to highly complex models with latent factor variables and multiple covariance structures (Bollen, Structural Equations with Latent Variables 1989). The structural equation modeling framework also provides many useful attributes for estimation of missing data problems (Enders, Applied Missing Data Analysis 2010).

Tu suggests that part of the underutilization of structural equation modeling in biomedicine may be related to a lack of familiarity with the models and the need to learn how to specify more complex models that structural equation modeling is capable of addressing (Tu 2009). There are examples of structural equation modeling in biomedicine though such as the study by Dahly, Adair and Bollen which investigate blood pressure (Dahly, Adair and Bollen 2009). Given that Kenneth Bollen is a premier researcher on SEM, this in part explains the reason for this study.

In genetics, most of the structural equation modeling based literature focuses on older data and research designs such as twin and adoption studies. These studies also tend to focus on psychological questions and psychology is a field where structural equation modeling is commonplace. Bartels, Cacioppo, Hudziak and Boomsma were interested in studying genetic and environmental contributions to stability in loneliness throughout childhood while D'Onofrio, Hulle, Waldman, Rodgers, Harden, Rathouz and Lahey studied how genetic and environmental factors interact with smoking during pregnancy to create externalizing problems in children (Bartels, et al. 2008) (D'Onofrio, et al. 2008). The use of structural equation models in biomedicine, especially genetics and proteomics is limited to a specific set of questions usually related to psychological outcomes. Structural equation modeling has also not made a real forte into mainstream analysis of high throughput microarray technologies either.

This chapter on statistical methods for analyzing immunosignatures provides a number of examples in which the structural equation modeling framework can be useful for helping to model complex relationships among peptide microarray data. It also presents a model for using confirmatory factor analyses or structural equation measurement models as a method for screening new samples for disease. It is also demonstrated that while simpler classical tests such as T-Tests are useful, the same level of information is not able to be obtained without using more complex modeling approaches such as those within the structural equation modeling framework. Structural equation modeling is necessary to untangle the complexity in modern biomedical data.

Experimental Study Background:

The human immune system is a rich source of information about the health and disease status of an individual (Johnson and Stafford 2009) (Metchnikoff and Binnie 2009) (Litman, Cannon and Dishaw 2005) (Legutki, et al. 2010). Immunosignaturing is a new technology that may be useful to decode the vast amounts of health information contained in the immune system. An immunosignature is a pattern containing multiplexed signals from chronic or recently matured antibodies. These signals come from a sufficiently diverse set of peptide targets on a microarray. Thousands of peptides of random sequence (mimotopes) provide the density and diversity sufficient to discriminate different diseases. An initial question, and the aim of this chapter, is how best to analyze and decode the information from immunosignaturing studies. Previous reports (Legutki, et al. 2010) (Johnson and Stafford 2009) used frequentist statistics (ANOVA or t-test) and cluster analysis (hierarchical clustering and Principal Components) to identify features that classify disease states. We examine other methods that may yield better performance in immunosignature analyses. Corrected T-Tests as well as logistic and multinomial logistic regression models have demonstrated an ability to differentiate between patients with different disease states even after stringent corrections for running multiple statistical tests (alpha inflation). Confirmatory factor analysis is an additional method which provides an abundance of information relating to the clustering of samples as well as providing an alternative method for categorizing and determining the disease state of a single sample. Descriptive statistics help to paint a better picture of the overall immune system activity. Finally, structural equation modeling and mixture models can help explain the underlying structure of an immunosignature.

For these analyses we examined a dataset containing breast cancer samples along with patients who had a second primary tumor (not a recurrence). The group with a second primary tumor was included in the analyses because if these patients could be diagnosed as having a high probability of developing a second tumor, they could be more closely monitored.

In an immunosignaturing study, sera samples are collected from participants and the physical information from the immune system is extracted using high density peptide microarrays. Each microarray contains a large number of peptides; in this case 10,375 peptides. The selection of these peptides was designed to give broad spectrum coverage of relevant antigens in the human immune system. The relevant nature of each peptide capitalized on early phage display research (Johnson and Stafford 2009). The decision was made to use a peptide microarray instead of phage library panning because of the increased speed and efficiency offered by a peptide microarray (Johnson and Stafford 2009). Ideally, if we can better understand the information captured by the peptide microarrays we may be able to develop quick, accurate, unobtrusive and inexpensive screening tests for many types of disease.

Classic peptide microarrays are created by spotting overlapping peptides corresponding to linear sequences of proteins known to be involved in an infectious disease. These arrays cannot identify non-linear epitopes. The epitopes are identified when B-cells produce antibodies (usually IgG) specific to 8-12 residue peptides that are components of the antigen protein. In contrast, immunosignaturing arrays utilize random-sequence peptides. Random sequence peptides have some specific and reproducible affinity to antibodies, and determining the level and pattern of binding is core to determining the difference between patients with different diseases.

Although much research has been done on statistical analyses using microarrays, immunosignaturing microarrays pose a number of novel challenges not encountered in traditional microarrays. In nucleic acid microarray technologies, binding is essentially only between two types of molecules of complementary sequence. For example, in a genotype array, genomic DNA binds to complementary nucleic acid probes that have either matches (e.g., perfect match, PM) or mismatches (MM) and the signals from the different probes are combined to make homozygous and heterozygous base calls for individual single nucleotide polymorphisms (SNPs). In a gene expression microarray, only a specific fragment of RNA will bind to the oligonucleotide on the array. With modern microarrays, as long as there is a sufficient abundance of RNA on the array, it will generally bind only to the specific complementary probe, with very limited non-specific binding.

With immunosignaturing microarrays, the intensity values are a continuous value from 0-65,000 and binding is not restricted to a single “complementary” molecule. Multiple antibodies in IgG could bind to the same 20mer peptide on the array. Also, although the immunosignaturing arrays are designed to measure IgG, there may still be competitive binding from other material in the sera and from other types of immunoglobulin. Competitive binding could result in an IgG antibody not binding at all or binding with a lower affinity. This could be potentially problematic if the auxiliary particle reducing binding affinity does not differ systematically across groups. Furthermore, a single antibody may also bind to multiple peptides on the array; a problem almost non-existent in genotype or gene expression arrays.

With the potential for so many different things to bind to a peptide on the array, it is not immediately clear how accurately traditional and more novel statistical methods would perform. One primary goal of the research reported here was to determine if the proposed statistical methods were capable of effectively analyzing the data and producing a correct pattern of results. For example, with a number of different things binding to a peptide and antibodies binding to multiple peptides it was initially uncertain if this would produce erratic signatures which would lead to incorrect results when certain methods were used.

Despite a number of new complexities created by immunosignaturing microarrays, these challenges give us the opportunity to test the performance of classically used methods such as factor analysis models in a different environment while also allowing us to ask new and fundamentally different research questions. In order to answer these new research questions, there is a need to use different statistical models not commonly used to analyze microarray data. This is because more traditional models used to analyze microarrays lack the versatility to adequately capture and explain the complexities of immunosignatures.

Here, we explore the use of structural equation models in order to try to determine whether the immunosignature formed by the fluorescent values of the 10,375 peptides is mostly random or if there is a consistent underlying pattern or factor structure to an immunosignature that correlated with disease. This research question is made possible because of the novelty in immunosignaturing arrays that that allow a single antibody to bind to multiple peptides on the array. This research shows that there are complex and consistently reproducible structures underlying peptides which differentiate groups. Such patterns can be used as biosignatures for disease as well as provide deep insight into antibodies and immune response to disease. Although there are new analytic challenges in immunosignaturing, it is these exact challenges that provide the promise of new discoveries while laying the groundwork for applications in future research and technologies.

In this chapter we present a range of statistical methods, their use and demonstrate what type of information they can provide researchers in immunosignaturing studies. We show the ability to classify samples into their respective disease categories and find peptides which significantly predict disease status. This provides a promising method for screening and potentially presymptomatic screening of disease. We also identify a number of latent factors using structural equation modeling. We hypothesize that the latent factors being modeled may represent specific antibodies that differ among disease classes.

Methods:

Patient samples are analyzed by applying the sera or plasma to the array at a 1:500 dilution, detected with an anti-human fluorescent antibody, and the signals are read using an Agilent C laser scanner. Images are processed using GenePix Pro 8 providing a text file of values for each peptide. Binding affinity is a continuous value from 0-65,500 (16 bit image). Genepix software was used to convert the 16-bit TIFF images to values, median non-background subtracted values were used and $\log_{10}$ transformation was done on the median normalized intensity values. Three distinct datasets were used in these analyses. One was a set of samples from a random group of individuals without breast cancer, a second set of samples is from a group with breast cancer and finally the third set of samples is from a group of patients who were diagnosed with a second primary tumor. The normal samples were a convenient sample of individuals without any known breast cancer history. The breast cancer samples were a sample of current breast cancer patients with different levels of disease progression and diverse demographic backgrounds. There were 52 samples from normal individuals without cancer, 98 samples from cancer patients with a single primary tumor and there were 21 samples with second primary tumors. Human subjects protection was observed, collaborators ensured all samples were collected under the same protocol. All of the samples came from females between the age of 45 and 54. The specific participant age for each sample was kept from us because of HIPPA and patient privacy concerns. All pre-processing was median-normalization per microarray slide, to adjust for global intensity bias. Data was also $\log_{10}$ transformed. The spot intensity was the median signal (obtained by GenePix Pro) with no local background subtraction. Background subtraction was not used because the arrays showed consistent background across the 1172 empty spots which were spread across the physical surface of the array. Technical replicates also showed greater reproducibility without background subtraction than with, indicating that the method for subtracting background was not useful. Additionally, the local and global background estimates were, on average, 150-300 RFU, which for any microarray is extremely low considering the 3+ logs of dynamic range.

It is common in similar lines of research, such as genotype experimentation to use a pattern matched experimental design. Matching participants in an experiment has the effect of increasing homogeneity among groups. As a result, the reduced within class variation which often accompanies matching designs has the effect of reducing the standard error and denominator of common statistical tests. This in turn leads to higher statistical power. Additionally, more homogeneous groups often enable easier classification in exploratory models. In the data analyzed here, the normal non-cancer samples were not matched to either the cancer groups, however research has shown that the signature of immune response is far less susceptible to the type of personal factors that genetic studies are – even HLA has only a minor effect on the consistency of a disease state immunosignature pattern (Johnson and Stafford 2009) (Legutki, et al. 2010).

Given that immunosignaturing is a new technology, early investigations, contrary to initial belief actually capitalize on the lack of rigid experimental designs. This is because additional sources of variance in the data allow us to better understand the robustness of the technology and related statistical analyses. If a method can perform well in a somewhat noisy environment with loose experimental designs, it is highly likely to perform even better when well curated studies (such as matched designs) are performed. In many respects, testing immunosignaturing data with loosely structured and curated data provides a much more stringent test of the technology and methods. Being able to obtain statistically significant results with the correct patterns of results from such unstructured data illustrates the versatility of immunosignaturing technology and the statistical methods tested here.

Understanding the robustness of the technology provide guidance for future experiments using this technology while giving insight into the potential clinical use of immunosignaturing. Biologically, it is possible that healthy normal individuals with no active infection are responding immunologically to their environment, and persons with an infection have a focused immune response. It is likely that high variation in immune response to an environment would be present across individuals.

Therefore, in order to be clinically useful, it is imperative that the technology and methods are robust enough to function accurately outside of precisely controlled laboratory settings; as would be encountered during clinical deployment of the technology.

Descriptive Statistics:

The first set of methods presented illustrate the capabilities and limitations of classical models such as descriptive statistics, the T-Test and factor analytic models. Table 6 provides basic descriptive fluorescence intensity statistics of each of the three disease groups. Descriptive statistics of an immunosignature provide a significant amount of insight into the underlying immune response during disease states. Of particular biological interest in this sample is the difference in the range of values from the three groups. The normal and single tumor cancer samples have ostensibly the same floor value while the second primary tumor cancer samples have a much lower floor value. This may suggest a suppression of the immune system in second primary tumor cancer samples. The single tumor cancer and second primary tumor cancer samples have progressively higher maximum values which may suggest an increased immune response associated with cancer and a reoccurrence of cancer.

Table 6 Fluorescence intensity and descriptive statistics for the three disease groups

Group	Mean	Minimum	Maximum	Std. Deviation	Variance	Range
Normal	329	207	9672	93	10336	9465
Single Tumor	336	204	16702	115	16258	16498
Second Tumor	676	36	49880	549	339301	49844

Although there are large differences in the ranges, in order to have any predictive validity, the differences in ranges need to be consistent across samples within each group. For example, a high fluorescence value over 45000 in the second tumor samples needs to occur on a given peptide with regularity to produce a statistically significant result.

Classical Statistical Significance Tests:

There are a number of statistical tests which could potentially be used to test whether the differences between groups across peptides are significant beyond what would be expected by chance alone. Some of these methods include the T-Test, corrected T-Tests, Logistic Regression and Multinomial Logistic regression. The standard T-Test divides the mean difference between two groups by a standard error to produce a T-Statistic used for null hypothesis significance testing. One problem with the standard T-Test is that the test makes the assumptions that the variances in both groups are equal. The problem of unequal variances in a T-Test is commonly known in the statistics literature as the Behrens-

Fisher problem and has been researched for the better part of the last century in various contexts. If the assumption of equal variances is violated, the T-Statistic can be either inflated or deflated depending on the samples sizes in each group. As a result, the analyses were conducted using a Satterthwaite corrected T-Test. The Satterthwaite test is one of numerous corrections for unequal variances that have been proposed over the years. The Satterthwaite test works by adjusting the degrees of freedom in the test. The resulting correction produces an asymptotically correct T-statistic when groups have unequal variances. The Satterthwaite correction works by modifying the degrees of freedom via equation 1:

$$df = \frac{(w1 + w2)*2}{\frac{w1*2}{n2-1} + \frac{w2*2}{n2-1}} \qquad \text{Equation 7}$$

A Satterthwaite corrected T-Test and a number of similar test corrections which could have also been used such as a Brown-Forsythe correction in an ANOVA model tended to produce statistically significant results after a Bonferroni correction for multiple testing (alpha inflation). A Bonferroni correction was used to protect against alpha inflation because with a standard alpha level of .05, purely by chance alone, 1 out of 20 tests will be significant. The Bonferroni correction divides the alpha value by the number of tests run; in this case 10,375, or one for each peptide on the microarray. This resulted in a corrected p-value threshold of 4.819*10^{-6}. Nonetheless, despite this much lower p-value, highly significant results are still obtained for Satterthwaite corrected T-Tests comparing normal versus single tumor cancer samples, normal versus second diagnosis samples and single tumor cancer versus second primary tumor cancer samples. Table 7 shows the top 10 significant peptides for a Satterthwaite corrected T-Test comparing normal samples to cancer samples. Logistic and Multinomial logistic regression may also be of interest and an alternative method for comparing groups to the tests used here. One place in which logit models may be useful is if a researcher in future studies has a known set of covariates they wish to control for. For example, in the study of diabetes, it may be of interest to control for body mass index or HB1AC test results.

Table 7 Top 10 significant peptides for a Satterthwaite corrected T-Test comparing normal samples to cancer samples

Variable ID	**Peptide Sequence**	**T-Value**	**DF**	**P-Value**
V2833	HFRKWHKRRWKHHKKWKGSC	-6.51	132.4	1.4372E-09
V3113	HRFKWHWKHRFHHFHRFGSC	-6.29	144.41	3.5843E-09
V6772	QKFKHQQGSFKLPWLSMGSC	-6.29	144.84	3.5843E-09
V9732	WRRSTPVGPWTWFGKFLGSC	-6.12	146.1	8.1933E-09
V7196	RFGRPQHQHDFRRHAIYGSC	-6.06	146.8	1.1046E-08
V6978	QSHMTLAPGIRRYKKFNGSC	-6.06	146.32	1.1046E-08
V7387	RMGFGLYERLWGKTNHYGSC	-6.01	134.26	1.6532E-08
V9561	WKWKRHWKWPHRRKHFFGSC	-5.95	144.49	1.9475E-08
V6987	QSIGLGYSAFMPKWPFRGSC	-5.93	140.13	2.2543E-08
V3249	HWKRHHRPKHKHHRHKHGSC	-5.9	145.4	2.4586E-

					08

Exploratory Factor Analysis:

Factor analytic models have previously been used in analyzing immunosignatures and are quite common in analyzing high dimensionality microarray data (Legutki, et al. 2010) (Kustra, Shioda and Zhu 2006) (Blume 2010). Each of the models explored during this line of research were investigated in order to determine its feasibility for answering a specific research question. Exploratory factor analysis (EFA) was examined as a method to be able to differentiate samples based on disease states with no prior clinical knowledge of the samples. Estimation of EFA models was performed using ordinary least squares (OLS). EFA with Promax rotation proved significantly better than chance at classifying samples. EFA is a set of procedures that accounts for the relationship among a set of variables in terms of a smaller set of underlying latent constructs or factors. (For example, a factor is a disease state.) We specifically use principal axis factoring with iterated communalities. Although PCA and EFA are quite similar, an important difference between the two methods is that PCA makes the assumption that all of the variance in an item is a reflection of common variance shared among all items whereas EFA posits that each item shares some common variance with all other items but also has its own unique variance. Mathematically the difference between PCA and EFA is the addition of single matrix; D^2.

$$Rzz = A * Rf * A' + D^2 \qquad \text{Equation 8}$$

In equation 2 R_{zz} is the correlation matrix among the observed variables. A is a matrix of factor loadings, R_f is the correlation matrix among the factor loadings, the A’ denotes the transpose of the A matrix of factor loadings and thus AR_fA' is the matrix representation of the common factor structure. D^2 is a diagonal matrix that captures the unique variance weights and distinguishes EFA from PCA.

Varimax and Promax rotation methods were explored in depth. This is in part because Varimax is often a starting point for a Promax rotation. A sample is said to “load on” a given factor when the model suggests a strong fit on the given factor. Rotation in EFA is a method for making factor loadings more interpretable. Rotation methods change the relationship between items and the factors (which are geometrically represented as axes). Rotation does not change the relationship among the individual items. Since rotation methods only make changes to the axes and not to the communalities (variance accounted for), rotation does not mathematically change the initially obtained results. Rotation makes the factor loadings more interpretable.

Varimax uses a complexity function to maximize the variance of the squared loadings on each factor. This results in loadings with a more even spread across the factors; as opposed to having an overabundance of loadings on a first factor. Varimax is an orthogonal rotation that maintains the orthogonal (90 degrees) intersection of the axes. This has the result of keeping the correlation between the factors at zero because the cosine of 90 degrees is 0.

Promax is an oblique rotation that allows the angle between the axes to vary. In statistics, variance has to be accounted for in some part of the model. Allowing the axes to vary and thus a correlation between the factors is another path to account for variance. Allowing variance to be expressed in terms of correlations between factors has the result of not forcing variance between factors to be represented as between item variance. This can result in cleaner factor loadings. Additionally, the assumption that there is no correlation between factors, or in this analysis, disease states, is unlikely because there will always be some additional common variance and similarities in immune samples due to basic immune responses and structures present across all samples.

Unlike Varimax, Promax does not use a complexity function. Rather, Promax rotation is a procrustean rotation to a target matrix. In Promax, a pattern matrix of loadings (often derived from Varimax rotation) is taken to some power (i.e., squared, cubed etc.) to form a target matrix. The original loading components are then rotated to get as close as possible to the newly formed target matrix.

A number of EFA models with Promax rotation were run to investigate the utility of this method for differentiating between groups with no prior knowledge of group membership. Table 8 provides summary results. The number of factors was known to be 2 for each comparison.

Scree plots were used to validate the hypothesis. None of the plots suggested the presence of a strong third factor. A scree plot plots the eigenvalues for each component. The largest components before a leveling off is used to determine the appropriate number of factors. Factor loadings greater than .3 were said to load on a given factor. If loadings for both factors were less than .3 the sample was said to not counted as a correct classification on either match. Catell (1966) provides a more detailed description of how to use eigenvalues and scree plots for determining the number of factors (Cattell 1966).

Table 8 Exploratory Factor Analysis Results

EFA Model	Correct Classification
Single Primary Tumor and Second Primary Tumor Samples	93.45%
Non-Cancerous and Second Time Cancer Samples	84.4%
Non-Cancerous and First Time Cancer Samples	68%

An EFA between cancer samples and the samples from patients who had a second primary tumor produced a correct classification for 93.45% of the cases. Of the cases that were miscategorized, all of them except one were cancer cases that loaded more highly with the second primary tumor group. There are a few possible explanations for this. This could simply be model error resulting from the lack of homogeneity among the first time cancer group. However, it is possible that the miscategorized cases may represent individuals who will at some point in the future develop a second primary tumor or are unbeknownst to the researchers already in the process of developing one. All this says is that less than 10% of cancer samples are more closely related to the samples of individuals who had acquired a second primary tumor than the samples with a single primary tumor.

A second EFA was run between the normal or noncancerous samples and the samples with a second primary tumor. The overall classification accuracy was 84.8%. Within this model, 74.1% of the normal samples loaded correctly on the same factor whereas 100% of the second primary tumor samples loaded on the correct and same factor.

A third EFA was run exploring the relationship between normal or non-cancerous samples and single tumor cancer samples. Using the same model specifications as in the first model, this EFA produced a 68% classification accuracy. Although this is quite low by traditional model building standards, there are a number of factors relating to the data which may make this a useful starting point. First, the normal patients were taken from a wide range of convenient lab samples. Some of the normal samples may have come from individuals outside of the age and traditional demographic background to even be remotely at risk for breast cancer. Secondly, the stage and progression of cancer patients was unknown. As a result, an additional possibility for the classification accuracy may be that the cross loadings represent a mixture of early stage cancer patients and those at high risk for or who are developing cancer.

Unfortunately, detailed information about the disease state of the samples is unavailable and thus makes any conjecture purely hypothetical. However, in all models, the results are significantly better than chance and illustrate in many ways the performance of the technology and approach under adverse conditions. The three models taken in concert illustrate that the lack of a concrete and well curated control group is likely responsible for the decremented classification accuracy in some models. This can be most clearly seen when considering that the single tumor cancer and second primary tumor cancer samples consistently exhibit stable factor loadings with relatively low cross loadings because the single tumor cancer samples serve as a much cleaner control group for the second primary tumor cancer samples than the normal do for either of the cancer groups. This early research suggests that future studies using more precisely selected control groups and experimental design would have even better ability to classify cancer patents.

Beyond classification accuracy, the similarity between different factor based models and rotations is extremely informative from a biological perspective. All combinations of PCA and EFA with Varimax or Promax gave highly similar results with respect to overall classification of groups across a number of different analyses. Although specific factor loadings certainly had different values, the overall picture and classification accuracy was relatively constant. Brief investigations into other rotations such as Oblimin were also explored in the context of EFA models and produced similar results to Varimax and Promax.

First, with respect to PCA versus EFA, the lack of difference suggests that the vast majority of the variance accounting for classification is at the factor level (ie. ostensibly disease state) and not the individual level. This is because as the D^2 matrix which differentiates the two methods captures the unique variance in an EFA model and as the D^2 matrix approaches zero, an EFA model approaches a PCA model. Therefore, since the D^2 matrix is the only difference in the equation and an analytic solutions exists due to Ordinary Least Squares estimation, we can conclude that the lack of difference was because there was relatively little unique variance present.

Confirmatory Factor Analysis:

Since EFA models showed the ability to differentiate samples, a logical clinical application of immunosignaturing would be to screen a single sample from an individual to determine his or her disease status. Confirmatory factor analysis (CFA) was chosen as an ideal method for investigating this question due in part to the similarity with EFA and because of the versatility to examine one specific sample in detail. EFA is an exploratory method that should be used when the number of groups or structure of the data is not well understood. Conversely, CFA is a confirmatory method that can be used when the structure of the data is well understood. As the name implies, exploratory factor analysis, EFA models should not be used as confirmatory model or to confirm a hypothesis.

Both CFA and EFA attempt to explain the underlying structure in a dataset. However, CFA and EFA approach the problem from two distinct directions. EFA makes almost no prior assumptions about the structure of the data and attempts to sort through the data to help a researcher determine what the underlying structure of the data is. In this research, the general group membership was known and thus the appropriate number of factors was specified apriori. In a CFA model, the researcher explicitly identifies not only the number of factors but which cases load on each factor as well as factor variances, covariance's between the factors and disturbances for each item. CFA models are not data mining approaches and require well formulated notions about the underlying structure of the data.

Mathematically, the simplest formulation of a CFA model in matrix notation is:

$$X = \Lambda * \xi * \Delta_L \qquad \text{Equation 9}$$

In Equation 3, X is a vector of observed variables, Λ is a matrix of factor loadings, ξ is a matrix of scores for each variable on a factor or latent construct and Δ is a vector containing measurement error.

In the CFA models analyzed here, one sample from each factor (disease state) was chosen at random as a scaling constraint in order to ensure identification in these models. Maximum likelihood estimation with robust standard errors was used to estimate these CFA models. The known disease status was the basis for defining the factor loading for each sample. A sample was allowed to load only on a single factor and fixed to zero on the other. Variances and covariance's between all factors were estimated. Summary results are provided in Table 9.

Table 9 Confirmatory Factor Analysis Results

CFA Model	Correct Classification
Single Primary Tumor and Second Primary Tumor Samples	89.9%
Non-Cancerous and Second Time Cancer Samples	93.1%
Non-Cancerous and First Time Cancer Samples	83.4%

For a CFA comparing single tumor cancer samples and second primary tumor samples, 89.9% of samples loaded on the specified factor. For a normal versus second primary tumor CFA, 93.1% of the samples loaded on the specified factor and a normal versus single tumor CFA produced sample loadings on the specified factor 83.4% of the time. The difference in classification accuracy between the CFA and EFA models is due to a number of factors; some of which include model variance and covariance specifications as well as different estimator types.

One primary advantage CFA models have over EFA models are fit indices which give some quantitative measure of how accurately the specified model is. Although there are a plethora of fit indices that have been proposed within the structural modeling framework that CFA models reside, the chi-square difference test, root mean square error (RMSEA) and standardized root mean error (SRMR) are among the most common and widely cited.

The chi-square test ostensibly tests how well the specified model reproduces the covariance matrix from the original data. The problem with this test is that it is so sensitive that it is nearly impossible to obtain statistically non-significant results. It is important to note that the null hypothesis of this test is that there is no difference between the specified model's covariance matrix and the covariance model in the actual data, a non-significant p-value is the desired outcome. Because it is of interest to find no difference between the specified model and the data, a non-significant p-value is the goal. The chi-square test for all of the CFA models was significant with $p<.001$ suggesting that there is a statistically significant difference between the specified model covariance matrix and the covariance matrix of the original data. However, the chi-square test is extremely sensitive and often detects trivial differences [8-9]. Noting the sensitivity of the test is not meant to suggest that in fact the specified CFA models are perfect fits or deny lack of fit. Rather, the test is noted because it is among the most common fit indices and the issues with the test are noted as a means of providing appropriate context for the results.

The Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR) are two common fit indices used in the Structural Equation Modeling framework description; of which CFA is a part of. The basis of the RMSEA is a non-centrality parameter. The simplest reduced form of the RMSEA equation is:

$$\text{RMSEA} = \sqrt{\frac{\left(\frac{x^2}{df}\right)-1}{n-1}} \qquad \text{Equation 10}$$

In equation 4, X^2 is the model generated chi-square value, df is the degrees of freedom and n is the sample size. Smaller RMSEA values suggest better fit. The SRMR measures the standardized difference between the observed covariance matrix and the model implied covariance matrix.

For the CFA model for single tumor samples versus second primary tumor samples, the RMESA was .083 and the SRMR was .071. For the CFA model comparing normal versus second primary tumor samples the RMSEA was .097 and the SRMR was .076 while the normal versus single tumor samples produced a RMSEA of .07 and a SRMR of .074. These are marginally significant results because traditional benchmarks cite .05 as a cutoff for statistical significance (Bollen 1993). RMSEA and SRMR values in the .05-.08 range are usually regarded as marginally significant. Although the results do not meet the rigid .05 level, they are actually quite impressive when considering the experimental design and the fact that a portion of the lack of fit may actually be representing natural biological patterns such as the development of a first or second tumor.

Perhaps the real utility of a CFA model for immunosignaturing could come in the form of diagnostic testing. Given the accuracy of the CFA model with this data, once a well curated set of samples for a certain disease or collection of diseases has been established, a CFA model could be specified where a new unknown sample could be allowed to load on both (or multiple) factors. By comparing the relative loadings on the factors, it would be possible to determine to which group the sample most likely belongs. For example, there are numerous subtypes of breast cancer and different stages of disease progression. If a collection of samples was available as a concrete reference set, a CFA model could be easily and accurately employed as a new method for aiding in the diagnosis as well as perhaps early detection of breast cancer.

Structural Equation Models:

While the method presented above are highly useful, they are inherently limited in the amount of complexity they can unravel. Structural equation modeling helps us better understand the complex underlying nature of immunosignatures. From EFA, CFA and descriptive statistics we know that the immunosignatures as a whole are in fact different across groups while corrected T-Tests show that there are statistically significant systematic variations. The logical question arising from these findings is how precisely do the immunosignatures differ from one another? Is there a clear, consistent and reproducible pattern underlying the differences in immunosignatures across disease states? Because a single antibody can bind to multiple peptides and different antibodies can bind to the same peptide, a coherent pattern of peptide fluorescence across an immunosignature is much more informative than the fluorescence of individual peptides on their own. Furthermore, being able to identify common relationships and covariances between groups of peptides is of even greater utility. This can be accomplished by modeling latent factors.

On a genotype microarray, the probe is directly measuring an individual’s genotype at a specific location. In contrast, the peptide probes on an immunosignature array are indirectly measuring immune response and antibodies present in the sera. When measures are not directly observed they are often referred to in statistical and structural equation modeling literature as latent factors. If there are clear, consistent and reproducible patterns caused by specific antibodies in a sera sample binding to peptides on an immunosignaturing array, it should be possible to model individual antibodies as latent factors. For example, when reading the tick marks on a mercury thermometer, one is not reading a direct measure of temperature but rather displacement of mercury. The latent factor measured by displacement of mercury is temperature because from a purely physics standpoint, temperature is the kinetic energy of an object; usually measured at the molecular level. Another example of a latent factor is depression. Psychologists cannot directly measure depression but they can ask a series of questions that cumulatively allow them to model the latent construct of depression. Each question in a depression inventory gets at one small piece of the latent factor depression in much the same way that peptides on an immunosignaturing array provide an indirect measure of immune response; as measured primarily by IgG antibodies.

Structural equation modeling is specifically designed for modeling latent variables. Structural equation modeling models have two parts: a path model comprised of regressing a set of variables on another and a measurement model in which CFA is used to form latent variables. When a set of measured variables is set to load on a given factor, the result is a latent factor. In structural equation models, the resulting latent variables can be treated as either endogenous or exogenous variables; depending on the research question of interest. A full structural equation model is a collection of equations defining each variable and their relation to one another. Since complex models can quickly generate a large number of equations, structural equation models are often represented graphically for quicker interpretation. Since confirmatory factor analysis is a major component in a full latent variable structural equation model, attempting to classify samples with factor analytic methods lent evidence to the feasibility structural equation models. These early models also provided a plethora of background information which aided in the testing of full structural equation models.

Initial Structural Equation Modeling Testing:

Despite evidence from previous factor analytic models that structural equation models should be feasible, since these are highly complex models, an incremental approach was taken to building and testing large structural equation models. To start with, a measurement model and full structural equation model was run using the top three peptides from the normal versus single tumor cancer samples (Table 7) to predict disease. The measurement model (ostensibly a confirmatory factor analysis) in a structural equation model tests the loadings of individual peptides onto latent variables. In this model one peptide was set as a scaling constraint and the other two were freely estimated. Three peptides were chosen because that is the minimum needed for model identification and provides for the simplest model. Because of the iterative nature of the maximum likelihood algorithms used structural equation models, starting with a simple model reduces computational time and aids in convergence. Furthermore, starting with the simplest model and building up is good practice in modeling.

Since a measurement model with 3 factors is just identified or has no extra degrees of freedom, fit indices cannot be calculated. However, all the variables load strongly on the latent factor with loadings greater than

.7. This finding suggests that the top 3 peptides are indicative of a single underlying latent factor.

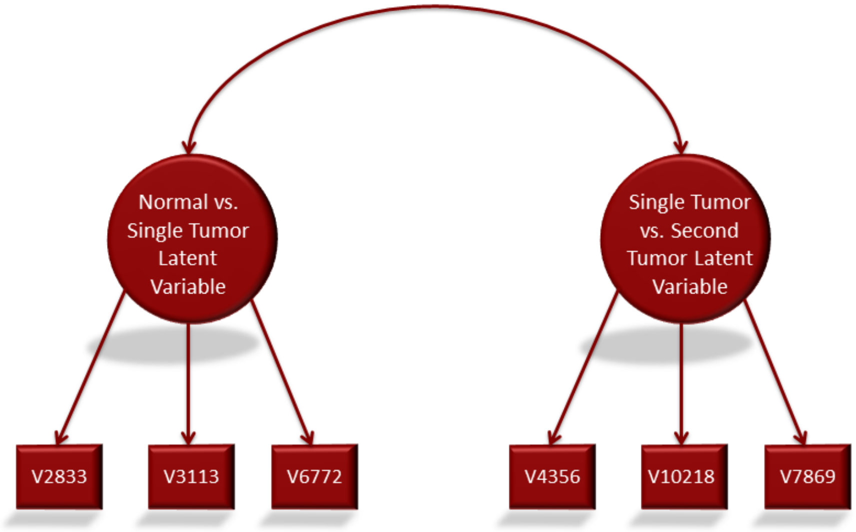
In order to help rule out the possibility that the consistent loadings in the first model were not type 1 error or false positive, the same model specification was run in an attempt to see if the top 3 peptides differentiating single tumor cancer samples from second primary tumor cancer samples. In this model the top 3 peptides also loaded on a single latent variable. Like the first model, the second model illustrated the same pattern of results with the top 3 peptides all significantly loading on a single latent factor.

The same pattern of results can be replicated with two disease contrasts. Replicating the finding with normal versus a single primary tumor cancer and second primary tumor cancer versus single primary tumor dramatically reduces threats to validity against causal conclusions proposed by structural equation models of immunosignaturing data.

When investigating models that differentiate two distinct groups from a baseline group (in this case single tumor cancer samples) there are three potential outcomes. First, a complete lack of model fit and no consistent underlying factor structure. In this case, none of the peptides would load consistently and correctly on either of two specified factors suggesting that peptide florescence is random. The second possibility is that all of the peptides would load on one factor. This result could result from any number of potential biases in the technology itself, printing or processing of the microarrays. Another reason all of the peptides might load on one common factor is that they are all part of a single latent factor. However, because the significance of each peptide varies quite precipitously across group contrasts, it seems unlikely that a single underlying latent factor would produce different significance values across disease contrasts. The third possibility is that the peptides significantly load on two separate factors and that the peptides for each contrast exhibit no cross loadings.

A series of analyses was run using significant peptides from normal versus single tumor cancer corrected T-Tests as well as second primary tumor samples versus single tumor samples combined into a single model. The first model was a measurement model which added the first two CFA's into one model. The top 3 peptides for normal versus single tumor samples and single tumor samples versus second tumor samples each were set to load on a separate latent factor. A covariance between the two latent variables was also estimated. The path diagram in Figure 2 illustrates this model. In Figure 2's path diagram, the square boxes represent measured variables, which, in this case are peptide fluorescent values. The large circles are the unmeasured latent variables. The arrows between the latent factors and measured variables show which measured peptides load on which latent variable. The curved arrow represents an estimated covariance between the two latent variables.

Figure 2 Latent & Measured Variables in Immunosignaturing


Normal vs. Single Tumor Latent Variable
Single Tumor vs. Second Tumor Latent Variable
V2833
V3113
V6772
V4356
V10218
V7869

In path diagrams, the arrows represent the causal flow of information. The arrows are pointing from the latent variables to the measured variables because the argument in structural equation models is that there is some unmeasured and underlying latent construct that is responsible for the observed results of the measured variables. The immune response and antibodies present in the sera samples is the ultimate causal factor of peptide fluorescence.

The model tested in Figure 2 was estimated using maximum likelihood estimation with robust standard errors (MLR). The model exhibits excellent model fit with an RMSEA of .063 and an SRMR of .031. In addition, the Chi-Square test was not significant, Chi-Sq=14.054, df=8, p=.0804. A non-significant Chi-Square test is the desired result. Again, this is because the null hypothesis of this Chi-Square test is that there is no difference between the observed covariance matrix (input data) and the covariance matrix implied by the model in Figure 2. These results strongly suggest excellent model fit and that the latent factors are unique constructs. Biologically, this suggests that a different latent factor is underlying each latent variable.

To further confirm the interpretation that the latent factors are different, one peptide from each factor was switched. V3113 and V10218 were set to load on the opposite factor from the first model. In this new model, there was a complete lack of fit. In addition to poor loadings, the fit indices dramatically decreased. The RMSEA was .354, the SRMR was .192 and the Chi-Square was 198.704, df=8, p<.0001. Thus further suggests two different underlying constructs rather than statistical anomalies.

An additional set of analyses were run using the top 5 peptides instead of just the top 3. The first models run in this sequence were Varimax and Promax exploratory factor analyses. Both models gave 100% classification with extremely strong loadings on each factor. Table 10 is the rotated factor pattern or a two group EFA taking the top 5 peptides from each disease contrast. This clearly illustrates the top five peptides strongly load on factor one while the last five strongly load on the second factor. The loadings of peptides are consistent with the groups from which each peptide was selected. For example, v4356, v10218, v7869, v8672 and v8170 were the top 5 most significant peptides differentiating first time cancer samples from second time cancer samples. In combination with earlier results, this very clear and consistent loading pattern strongly suggests that the top peptides for each class form unique latent variables and they are almost irrefutably measuring different constructs. Biologically, this suggests that the latent factor which is more active in single tumor cancer samples compared to normal samples is not the same latent factor that appears to be present in second tumor samples.

Table 10 Rotated factor pattern for two group EFA of significant peptides

Peptide ID	Peptide Sequence	Factor1	Factor2
V4356	KYQFAGQRSGKQYRWRIGSC	0.88773	0.05624
V10218	YQPPPRKAVIQMDWLSYGSC	0.92126	0.06844
V7869	SKFRDVLTFNEPSRFVSGSC	0.51657	0.04716
V8672	TVHESMIYRMRFMTFKHGSC	0.93261	0.04783
V8170	SWRRMRMHKNFMISNLDGSC	0.87997	0.06368
V2833	HFRKWHKRRWKHHKKWKGSC	0.11128	0.7436
V3113	HRFKWHWKHRFHHFHRFGSC	0.06271	0.82673
V6772	QKFKHQQGSFKLPWLSMGSC	0.12145	0.73203
V9732	WRRSTPVGPWTWFGKFLGSC	0.05844	0.88795
V7196	RFGRPQHQHDFRRHAIYGSC	0.035	0.88098

The same result was also found by running a two group exploratory factor mixture model with Geomin rotation. Geomin rotation is another oblique rotation method similar to Promax. A more complete discussion of the mathematical differences of rotation methods can be found in (Browne 2001). In this data, the observed peptides as a whole form a single distribution. In mixture modeling, the underlying notion is that the distribution formed by all of the observed data is the product of two or more underlying distributions; each of which represents a distinct class. Ostensibly, an exploratory factor mixture model is trying to answer the same question as PCA and EFA, PAF/Factor Analysis but via a different mathematical framework. Despite the complexity of mixture modeling, the basis of an exploratory factor mixture model is for a categorical latent class variable C, for a specific class k. The model estimated is:

$$Y_p = V_{kp}+\lambda_{kp}*\eta*\varepsilon_p \qquad \text{Equation 11}$$

In equation 5, for a variable Y_p,$V_{kp}$is an intercept parameter, $\lambda_{kp}$is a vector of loadings, η is a vector of latent factors and ε_p is a residual term. In addition, there is a correlation matrix Ψ_k for the latent factors η of class k along with a distribution for the latent class variable C: $P_k = P(C=K)$. In this equation, for a dependent variable P, the probability of C is equal to k. Also, other constraints are added to this basic framework for purposes of identification but are related to model specific decisions such as orthogonal or oblique rotation.

EFA mixture models were estimated using Maximum Likelihood with Robust Standard Errors (MLR) estimation and 20 random start values. Random starting values were used in part due to the complexity inherent in mixture models and to check for local solutions. By running the analysis with multiple random start values log likelihood (LL) values can be compared. To the extent that different LL values are obtained, the random start values can be directly input into the model and the results can be compared to the best fitting LL model. This is useful because if different start values produce dramatically different results, this might suggest that the algorithm converged at a local maxima instead of a global maxima or that the results are unstable.

Fit statistics such as the Bayesian Information Criterion (BIC) provide a more quantitative analysis of model fit for a series of nested models. EFA mixture models were estimated for one, two and three class models. This approach allows us to confirm that a two class model is in fact the best fit for the data.

The series of EFA mixture models suggested the same pattern of results as traditional EFA models; that there are two distinct and separate underlying classes formed by the top 5 peptides for each disease contrast. In addition, mixture models also produce a statistic for the average latent class probability:

$$P (Yp = j \mid C = K) = \varphi - 1 (T *kpj) - \varphi - 1 (T*kpj -1) \text{ Equation 12}$$

In equation 6 T^{*}_{kpj} is a threshold parameter on a standardized correlation metric and φ is a matrix of residuals for the latent factors [11]. For both two and three class models, the average latent class probability for the most likely latent class membership was greater than 99% for both class 1 and class 2. In other words, for the subgroup of samples classified as being part of class 1 by the model, more than 99% of the time, class 1 was also their most likely class membership. This further reaffirms the excellent model classification. The three class model produced nearly identical average latent class probability values because the model did not classify any of the peptides as belonging to the third class.

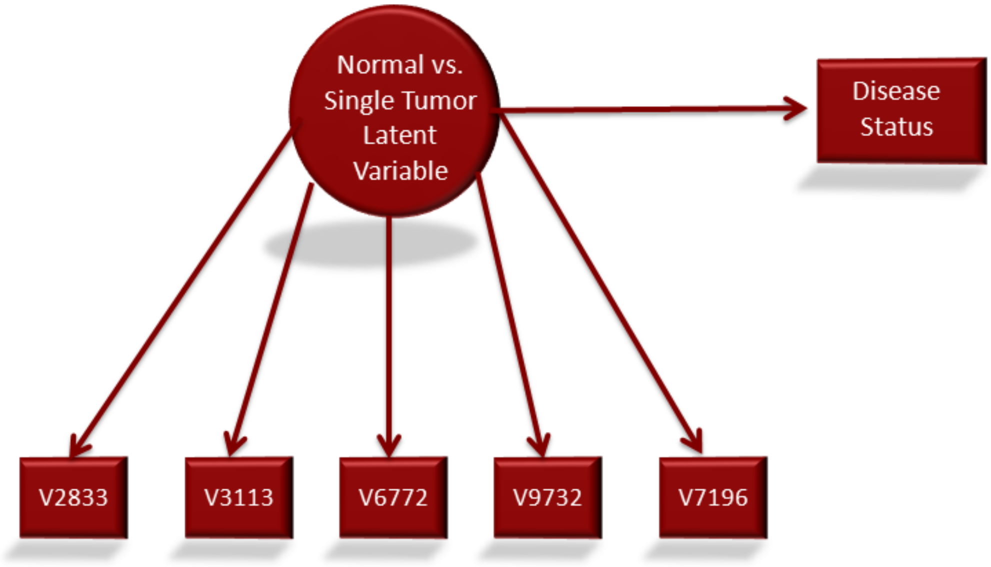
The BIC was used to assess the best fitting model. The BIC is estimated as follows:

$$BIC = - 2 * LL + p * \log(n) \qquad \text{Equation 13}$$

In equation 7 LL is the log likelihood value of the model, p is the number of parameters and n is the number of observations. The lower the value of the BIC the better the model fit. Often times, BIC values or plots are used ostensibly in the same fashion that scree plots and eigenvalues are used in PCA or traditional factor models where a researcher looks for the point at which the decrease in values levels off. However, in this analysis, the two class model had the lowest BIC and somewhat unexpectedly, the three class model actually saw a slight increase in the BIC This result further reaffirms the excellent fit of a two class model.

As is common in model building, a series of full structural equation models were run in increasing levels of complexity. To start with, the two latent variables were regressed on their respective disease states in individual models. A path diagram for the normal versus single tumor samples is presented in figure 3.

Figure 3 Path diagram for normal versus cancer peptides SEM model



These models were estimated using MLR. The latent variable regression was performed using logistic regression and was significant, p<.001. Additionally, the odds ratio was 1.841. This suggests that having the attributed measured by the latent variable makes an individual 1.841 times more likely to develop breast cancer. The same model specification for single tumor versus second tumor samples produced similar results with p<.001 and an odds ratio of 3.49. In other words, there appears to be a latent factor that is present in those who have a single tumor that is not present in those samples with a second primary tumor.

Furthermore, another structural equation model was run combining the above two analyses so that the two distinct latent variables were used to predict disease status. The estimation of disease status was done via multinomial logistic regression. This was done because when the models were combined there were three levels of disease. In a multinomial logistic regression model, one level (in this case single tumor samples) was set as the reference group. Then n-1 logistic separate regression equations are run; where n is the number of levels of the dependent variable. Therefore, since each latent variable was regressed on disease status, there were two logistic regression equations run. Both latent variables predicted their respective disease status with p<.01. Again, this suggests that normal, single tumor cancer and second tumor cancer samples are separated by different sets of latent variables.

The first set of structural equation models provided an initial proof of concept for full structural equation models. This laid the groundwork for the more interesting question of what the underlying structure looks like for unique parts of the immunosignatures. Since further investigations are meant to look at the overall differences in immunosignatures as a whole, it is hypothesized that the latent factors differentiating groups are specific antibodies present in the sera samples; as explained above. Two experimental tests were conducted: a series of structural equation models and an examination of the peptide means across groups.

Structural Equation Models of Significant Peptides and Antibodies:

Next, all of the peptides that were statistically significant after a Bonferroni correction in the normal versus single tumor and second tumor versus single tumor contrasts were selected for further analysis. Following the same pattern as before, exploratory factor analysis models were run to determine how many underlying factors appeared to be present. This was done because selecting the top peptides might yield more than one factor; suggesting more than one antibody. For the normal versus single tumor contrast there were 176 peptides that were significant and there were 30 significant peptides for the second tumor versus single tumor contrast. The eigenvalues and scree plots suggest a three factor solution for the normal versus single tumor contrast and a one factor solution for the second tumor versus single tumor contrast. In other words, for the normal versus single tumor, the hypothesis is that there are three antibodies that differentiate the groups while there is only a single antibody differentiating the second tumor versus single tumor groups.

In the second tumor versus single tumor contrast, factor loadings from exploratory factor mixture models and Promax EFA models confirm an unstable second factor. This is because the loadings on the second factor are generally low and minimally larger than the first factor loading on the same peptide. Additionally, a two factor solution produced Heywood cases in which there were communality estimates greater than one; suggesting a problem with the two factor model. When single factor models were run, all of the peptides loaded highly on the one factor. As a result of the EFA models suggesting a single factor solution, a full structural equation model was run in which all of the top 30 peptides were set to load on a single latent variable which was then regressed on disease status. In this model, the stable latent factor significantly correlated with disease status, p<0.001. The odds ratio of 3.148 suggests that the single hypothesized antibody confers significant risk for acquiring a second tumor. Also, the means for all of the peptides in the second tumor samples were lower than the means for the single tumor samples. This suggests immune suppression. In other words, there appears to be an antibody present in samples with a single tumor that is not present in samples with a second tumor.

The normal versus single tumor samples is a bit more complex. A full structural equation model containing all three hypothesized factors was unable to be estimated because there were more peptides than samples. Therefore, there were not enough degrees of freedom to run a full model containing all 3 groups. As a result, subsets and individual factors were tested individually. When tested individually, all of the three factors/hypothesized antibodies significantly correlate with disease, p<0.01. Two of the latent factors positively correlated while the third negatively correlated with disease status.

Within the 176 significant peptides for normal versus single tumor samples, 162 peptides increase or have a higher mean in the cancer samples than in the normal samples increase. Conversely, 14 decrease or have a higher mean in the normal group than in the cancer group. In other words, there appears to be two new antibodies present in cancer samples not present in normal samples and one antibody present in normal samples that is not present in cancer samples. Immunosignatures are unique in analysis of the humoral response in that they can detect decreases in reactivity relative to normal levels.

One finding of particular note is a high covariance between the two positive factors (or proposed antibodies present in cancer that are not in normal samples). The high covariance and multicollinearity suggests that the two are very similar. When regressing both of the positive latent variables on disease, in every instance, only one of the latent factors was significant with $p<.05$. This is likely due to the way in which multiple regression partitions variance. In a multivariate regression model, the effect of one variable (x) is the unique contribution of that variable with all others held constant. Because there is so much common or shared variance, a vast majority of the variance is used up or accounted for by the first factor, not leaving enough unique unexplained variance left for the second factor to be significant as well.

A two level measurement model was used to test whether the two factors were measuring a similar underlying construct. In this model, the two latent factors were set to load on a third latent variable. The reasoning behind this test was as follows: if the two latent factors loaded on a single second level latent factor, then the two original factors would be measuring the same underlying construct. One way this could occur is if the antibody had a highly complex structure. This model was not, however, statistically significant, RMSEA = .21, SRMR = .09. This suggests that the two factors are unique albeit highly similar.

There are a number of potential interpretations of this result. One of the more plausible biological hypotheses is the presence of subpopulations. Among two different cancer subtypes of single tumor breast cancer, there are likely two distinct antibodies; one for each subtype. If subpopulations are present in the data, it seems plausible that these two antibodies are highly similar because both are, in the end, I responding to breast cancer. The variations that lead to different subtypes may in fact be what makes the two positive latent factors separate and distinct from one another. The multicollinearity may be because they vary together, not that they have a similar sequence and see the same antigen. If two different antigens consistently arose in a tumor they would raise antibodies that varied together in samples but would see completely different antigens.

A second possibility is that the high multicollinearity is a result of modeling different times in the disease progression. As disease progresses it is likely different antigens are presented by the tumor to the immune system. If so, the relative amount of particular marker antibodies will also change.

Implications of Structural Equation Modeling:

We have explored a number of statistical models for analyzing immunosignatures. Each method explored helps answer a different research question relating to the analysis of immunosignatures. Descriptive statistics about an immunosignature can provide high level information about the general immune response in a signature. Exploratory factor analytic models (PCA and EFA) can be useful for classifying immunosignatures into different disease groups without any clinical information. CFA models can classify samples onto specified factors and could be developed into a useful model of disease. These structural equation models identify interesting and robust latent factor structures underlying immunosignatures which warrant further investigation.

Latent factors can be reliably extracted from immunosignatures. These latent factors are clear, consistent and replicable patterns which differentiate disease states in terms of their statistical significance fashion. These latent factors can serve as strong biomarkers for disease. Given the design of immunosignaturing and the fact that antibodies are binding to peptides on immunosignature arrays, it is highly plausible that the latent factors are modeling individual antibodies.

Although future research is needed to conclusively confirm the relationship between modeled latent factors and antibodies, the potential of having a high-throughput bioinformatics-driven method for antibody discovery creates countless potential avenues for future applications. The primary benefit of this methodological approach is to reduce the time it takes to identify antibodies associated with various clinical situations. Reducing this time reduces cost and increases the speed of advancement in biomedicine. Additionally, the increased speed of analysis and resulting reduced cost may permit applications that were not conceivable just a short time ago. For example, a method for quickly and inexpensively detecting an antibody could play a crucial first step in developing personalized vaccines.

Below we present a multi-step procedure for detecting latent factors and potentially antibodies in an immunosignaturing study. The first step is to run an exploratory factor analysis on the data with rotation. Various rotations should be explored but Promax or Geomin are recommended. An EFA model is a useful starting place for multiple reasons. First, it ensures that the groups are different constructs and significantly different from one another. This determination can be made by looking at scree plots and eigenvalues to assess the probable number of groups in the model; these should be equal to the number of known disease states. The samples should load correctly on a given factor with a high classification rate.

At this point, cross loadings in an EFA model can be investigated. If clinical data exists, it would be of use to try to assess whether there are potential reasons why a specific sample may be cross loading. For example, is there a history of cancer in a normal sample that cross loads on a cancer sample which might suggest the person is in a transition phase? This may be a way of detecting aberrant cases or outliers. That said, haphazardly removing cases from a dataset is NOT advocated. Cross loadings were not analyzed in this application due to a lack of additional information and clinical data upon which to draw any relevant conclusions.

From here, a test of statistical significance comparing the groups for the peptides of interest can be done using an appropriate test statistic. T-Tests or logistic regression and their multivariate extensions, ANOVA and multinomial logistic regression, are a few potential methodological tools. The specific test should be picked with respect to the features of the data being analyzed. For example, in this chapter, we used a Satterthwaite corrected T-Test because of unequal samples sizes and variances. The Satterthwaite test was chosen on the basis of the Monte Carlo simulation run in chapter 3. A correction should be made to protect against alpha inflation. Although a number of tests exist for this purpose, the Bonferroni correction is among the most common; even if it may be somewhat conservative.

A traditional EFA model or an exploratory factor mixture model can be used to infer the structure of the significant peptides within each group. This information can be used to create a full structural equation model. However, as part of good model building practices, starting with a CFA measurement model is recommended; especially because the iterative nature and complexity of these models may lead to convergence problems. Additionally, information from these simpler models can be used to specify starting values in full structural equation models if convergence problems occur. CFA measurement models specify which peptides load together on a given latent factor. Checking the fit of the measurement models can confirm the accuracy of the model. However, given that CFA is so similar to EFA methods, it is unlikely that differing results would be obtained.

Once a working measurement model has been obtained, a full structural equation model can be created by regressing the latent factors on disease state. It is important to test a full structural equation model for a number of reasons. Although EFA and CFA models may suggest that a group of significant peptides are related in some way, without a full structural equation model, there is no way of knowing whether the relationship is a significant predictor of a specific disease state. In the absence of predictive validity for a specified disease state, any relationship among the peptides is trivial and would not suggest that it is because of a common antibody. The same conclusion can made if the latent factor is predictive of disease states beyond the hypothesized state.

If a significant structural equation model can be obtained, wet lab validation can then attempt to determine whether the model is correct. One potential way of testing this in the wet lab would be to use the designated peptides to affinity purify the antibody from the sera. The prediction is that the different peptides would purify the same antibody. This could be tested by immunosignaturing the antibodies purified.

Screening and Presymptomatic Screening for Disease:

The relative ease from which samples can be classified and differentiated with all of the methods explored here makes this technology an excellent use for disease screening. Whether examining the loadings of new samples in a CFA model or as part of a larger structural equation model, this technology can allow researchers to screen patients in a variety of contexts. This initial research suggests that immunosignaturing could be developed into a quick and inexpensive method of screening for cancer. Taking a small sera sample from an individual is much less expensive and intrusive than traditional screening methods such as mammograms. One early potential use for immunosignaturing would be to help follow at risk populations; such as those individuals with a family history of cancer. Immunosignatures could be taken at regular intervals between regularly scheduled mammograms. If the generated immunosignature from an interim test started to suggest a closer similarity to cancer, this could prompt physicians to follow the patient more closely or advise additional screening. Immunosignatures could be used in the same way for individuals who already have cancer. In this case, if an immunosignature suggested the person was developing an antibody signature indicative of a second tumor (or more closely loading on a latent factor biomarker), the individual could be followed more closely to detect the presence of a second primary tumor.

Screening for a specific disease state is fairly straightforward. A well curated collection of disease samples would form baseline control factors. A sera sample would be taken from an individual and their sample would be allowed to freely load in a CFA model across relevant disease conditions. A significant loading on a disease factor would provide strong evidence for the person having a given disease.

There are a number of ways in which a presymptomatic screening test could be developed from immunosignatures. This could be done by collecting a longitudinal or time series sample of sera from an individual and following the factor loadings on a disease state over time. As the loadings on a disease factor tend to increase the individual could be watched more closely and additional screening for a disease could be recommended by a physician. A number of statistical methods and time series analyses such as latent transition analysis (LTA) could be employed to model this.

Discussion:

Immunosignaturing is a novel approach for understanding disease. A number of statistical methods including, exploratory factor analysis, confirmatory factor analysis, descriptive statistics, corrected t-tests, ANOVA, logistic and multinomial logistic regression, mixture models and structural equation modeling have shown promising abilities for analyzing different dimensions of immunosignatures. Immunosignaturing in the context of breast cancer has been shown to be a good platform for differentiating groups of samples based on disease status, determining the disease status of specific samples as well as potentially serving a role in the discovery of antibodies for specific diseases.

Despite many new challenges posed by immunosignaturing microarrays such as competitive binding and binding to multiple sites, the analyses conducted here clearly illustrate the usefulness of classical analytical methods to produce accurate results. The results are particularly noteworthy because of the lack of structure in the data and lack of a full pattern matched experimental design. The early results of structural equation modeling are very promising. Although wet lab validation is needed for the proposed methodology of antibody discovery, even if the latent factors turn out not to be a specific antibody, the model can still serve as an excellent biosignatures for disease screening.

Early detection of cancer is among the best predictors of survival. Continued development of immunosignaturing into a screening and presymptomatic screening diagnostic tool will aid in early discovery and help turn the corner in the fight against cancer. Future research in this field should aim at validating the hypothesis that the latent factors modeled here are in fact antibodies and to develop the technology into a diagnostic screening tool.

High Performance Computing:

In this study, each sample were loaded into a relational database. The database connectivity methods were used to access the data. With the multiple tests run, this eliminated the need to create a separate data file every time a different format was needed for a different analysis. In addition to reducing the time needed to export data files and reimport them into SAS, this dramatically reduced storage space from extra files because the total data size for this experiment was over 10 gigabytes.

Multi-core and distributed grid processing was used to process the large number of pairwise comparisons. The syntax in chapter 2 on distributed grid processing was derived from the syntax used in this chapter. Also, multi-core processing was used to perform the Monte Carlo chapter used to decide on the use of the Satterthwaite T-Test presented in chapter 3 and used in this chapter.

High performance computing methods reduced storage and dramatically reduced the time necessary to process the large number of calculations necessary to test the more than 10,000 peptides present on the immunosignaturing array.

Conclusion:

While simplistic methods can be a good starting point for modeling, methods such as the T-Test cannot fully unravel the complexity underlying the data when latent factors are present. Using structural equation modeling allows researchers the methods necessary to finally be able to ask questions not possible with simpler methods. Beyond being able to unravel complexity of latent factor not before possible, the use of the structural equation modeling framework may also prove useful for developing diagnostic devices to help screen for complex diseases.

CHAPTER 6

CONCLUSION

The living world we inhabit is massively complex. For years scientists lacked the data to ask many fundamental research questions in biomedicine. With the recent advances in computers, databases and new technologies such as microarrays we finally have access to the data necessary to investigate many areas of biomedicine including population health dynamics and the role genomics and proteomics play in disease. Unfortunately, now that we have the data, one major question is how to make sense of it amidst the complexity.

Complexity is an inherent trait of the natural world we inhabit and in part a byproduct of sequential evolutionary change as well as from the enormous size and structure of modern biomedical data. New methods are necessary to help cope with the ever increasing size of biomedical data and to be able to effectively make sense of what the data means. However, before beginning to address the complexity in the natural world, new tools need to be made more accessible to average analysts to help cope with the volume of modern biomedical data.

While it is imperative that researchers actually be able to physically process all of the data in a timely fashion, the result is not of high value unless the results obtained are accurate. Statistical methods are based on many assumptions and when violated they produce incorrect results; which are often not easily predictable in a precise fashion. Determining the most appropriate model is difficult in the univariate case. The complexity added to the task is increased dramatically by the size and variability in modern biomedical data to the point that the question of interest is not simply which model makes the most sense but rather to which model performs the best across the entire range of the data. Monte Carlo Simulations combined with sampling methodologies are advocated as a way to test model performance and to pick the model which will provide the most accurate results across the spectrum of the dataset.

Another added source of complexity in biomedical data is multi-level or clustered structures in the data. Clustering in the data if left uncorrected can cause inaccurate results in the form of inflated type 1 error rates. Additionally, there can also be disaggregated relationships in which ignoring clustering can produce parameter estimates not only with incorrect magnitudes but also incorrect signs. Unfortunately, despite the known hazards of not using multi-level modeling when clustering exists, the method is still underutilized in biomedicine. When multi-level modeling is used, it is not clear that it is used correctly and many critical pieces of information needed to accurately assess the models are omitted from journals.

Beyond the limited use and arguably correct use of the method, multi-level models and experiments designed with them in mind are necessary to more fully understand the complexity in the data. For example, simply knowing that both physicians and patients have some causal role in adherence to treatment is only of limited use. Understanding the interaction between the two and being able to know how much variance in adherence comes from physician and how much comes from patient factors is critical to being able to develop the most effective interventions possible. Failing to correct for clustering will give incorrect model results, but designing experiments to avoid clustering means the researcher is limiting the amount of information they are using and are ostensibly ignoring the inherent complexity of our natural world. The necessity of multi-level models is not simply to obtain correct parameter estimates but also to help us more fully understand and unravel the complexity amidst modern biomedical data.

Finally, structural equation modeling is presented because the presence of latent factors makes an in depth analysis of modern biomedical data highly complex. Simpler classical methods are incapable of detecting or adequately modeling latent factors. Ignoring the presence of latent variables in biomedical data fails to explain the complex relationship among variables in the dataset. Method such as multi-level modeling and structural equation modeling are needed to unravel the complexity. In addition, confirmatory factor analysis or the measurement

Future Directions

The methods presented in this dissertation are highly useful, but do not constitute an exhaustive treatise on the analysis of biomedical data. Additionally, this dissertation focuses more on individual pieces rather than how they can be integrated together into an enterprise workflow. As a result, a major future direction for this research is in integration of methods and methods into workflow.

With respect to integration of methods, multi-level modeling and structural equation modeling can be integrated into a single model when necessary. The implementation of high performance computing in each chapter discusses how complexity of voluminous and computationally intense data often co-occurs with other complexities such as clustering or latent class variables. In the same context, clustering sometimes co-occurs with latent class data. For example, assume that in the immunosignaturing study that the sera samples had come from different labs and that the expression was significantly different based on which lab the sample came from. Samples coming from participants from sets of families or diverse racial backgrounds may also lead to clustering in genomic the data due to the inherent genetic variability within the human population. These examples may all require the use of a multi-level model to correct for clustering. However, the need to use a multi-level model would not preclude researchers from also then investigating latent class variables. Studying the multi-level structure might produce potentially novel new information while still allowing for the investigation of latent factors.

The methods are presented as individual chapters but are not entirely separate and unrelated to one another within the context of an overall workflow for data analysis. More emphasis can be placed on integrating analysis into the workflow. Beyond the initial mention of integrating an analytic package such as SAS with a database, once the data is aggregated and loaded into a database or analytic program, automated processes could be written to perform a Monte Carlo simulation and select the most appropriate model. Then, the selected model based on a predefined target algorithm could be used to run a screening experiment on all of the samples in the dataset; in the same way the Satterthwaite T-Test was run on the peptides in the immunosignaturing study based on the Monte Carlo results. The results would then all be output into one aggregate report upon completion. The high performance computing method such as pipeline parallelism and multi-core processing could be integrated into this process.

Providing tools to integrate the methods presented in this dissertation into a streamlined workflow is useful for a number of reasons. Providing simple automated tools will to help these methods to become a more normal part of standard research practices. This is because performing one task is less complex for the end researcher who may not have an informatics or statistics background. Providing a single validated package for researchers to use is also less error prone than having researchers perform many sequential manual tasks. Additionally, integrating and automating the process reduces time for an experiment to run.

However, beyond one experiment, having automated analytic processes and supporting methods (high performance computing) that can be widely applied to data within a given domain is a key part in building community wide informatics pipeline for integration and analysis of biomedical data. Such a system could be placed in the cloud to allow for real time and on demand processing and reporting of predefined data streams within biomedicine; thus removing the informatics burden from researchers such as epidemiologists, biochemists, geneticists in order to enable them to focus more directly on their core area of expertise.

In the case of population health in which reports are received from multiple hospitals or labs where clustering is endemic to the data, multi-level models could be directly integrated into the process flow. As data is received, it is loaded, normalized to a common data structure, variables denoting clustering are tagged during loading and normalization period which would be passed directly to a module performing multi-level modeling and finally the results of the model can be sent directly to the necessary recipient upon completion. Such results based on more accurate multi-level models would give a better reflection of differences between hospitals and labs or how covariates of interest are influencing spread of disease.

Although the data is complex and the models presented in this dissertation are also complex, there are many common denominators. Specifically, there are many recurrent forms of complexity such as the voluminous size of data, heterogeneous variances, clustering or latent constructs. By better understanding how and when they occur through concrete examples presented in this dissertation, we can use the methods presented to help address common complexity problems in a more automated process flow.

Complexity from natural from natural and synthetic sources will continue to be a challenge for analyzing biomedical data. The advancement of the methods presented in this dissertation as well as integrating them more seamlessly into workflow and informatics pipelines will increase the pace of biomedical discovery. The driving goal behind this dissertation is that by advancing analytic methods and giving biomedical researchers tools to help facilitate analysis, the bench to bedside timeline can be decreased; ultimately ameliorating the quality of life four countless individuals.

REFERENCES

Affymetrix. *Affymetrix Microarray Solutions.* 2 22, 2012. http://www.affymetrix.com/estore/browse/level_one_category_template_one.jsp?category=35796&categoryIdClicked=35796&parent=35796 (accessed 2 22, 2012).

Anderson, Phillip. "Some thoughtful words (not mine) on research strategy for theorists." *Physics Today*, 1990.

Ary, D. V., D. Toobert, Wilson, W., and R. E. Glasgow. "Patient perspective on factors contributing to non-adherence to diabetes regimen." *Diabetes care*, 1986: 168-195.

Bartels, M., J.T Cacioppo, J.J. Hudziak, and D. I. Boomsma. "Genetic and environmental contributions to stability in loneliness throughout childhood." *American Journal of Medical Genetics Part B*, 2008: 385-391.

Behrens, W. "A contribution to error estimation with few observations." *Landwirtschaftliche Jahrbücher*, 1929: 807-837.

Bender, B. G., and C. Rand. "Medication non-adherence and asthma treatment cost. Outcome measures." *Current Opinion in Allergy & Clinical Immunology*, 2005: 191-195.

Benjamini, Y: Hochberg, Y. "• BenjamControlling False Discovery Rate: A Practical and Powerful Approach to Multiple Testing ." *Journal of the Royal Statistical Society*, 1995: 289-300.

Berman, J. *Biomedical Informatics.* Sudbry, MA: Jones and Barlett Publishers, 2007.

Bernstam, Elmer, Jack Smith, and Todd Johnson. "What is biomedical informatics?" *Journal of Biomedical Informatics*, 2010: 104-110.

Blume. "A Factor Model to Analyze Heterogeneity in Gene Expression." *BMC Bioinformatics*, 2010: 368-397.

Bollen, Kenneth. *Structural Equations with Latent Variables.* USA: Wiley-Interscience, 1989.

—. *Testing structural equation models.* United States of America: 1993, 1993.

Brown, J, P Stafford, S Johnson, and V Dinu. "Statistical Methods for Analyzing Immunosignatures." *BMC Bioinformatics*, 2011.

Browne, M. "An Overview of Analytic Rotation in Exploratory Factor Analysis." *Multivariate Behavioral Research*, 2001: 111-150.

Cabana, M., et al. "Why Don’t Physicians Follow Clinical Practice Guidelines?" *JAMA*, 1999.

Cattell, R. "The Scree Test for Number of Factors." *Multivariate Behavioral Research*, 1966: 245-276.

Chen, H, S Fuller, C Freidman, and W Hersh. *Medical informatics: knowledge management and data mining in biomedicin.* United States of America: Springer, 2010.

Choudhry, NK, RH Fletcher, and SB Soumeral. "Systematic Review: The Relationship between Clinical Experience and Quality of Health Care." *Annals of Internal Medicine*, 2005: 260-273.

Cios, K, and G Moore. "Uniqueness of Medical Data Mining." *Artificial Intelligence in Medicine*, 2002.

Cohen, J, P Cohen, S West, and L Aiken. *Applied Multiple Regression Third Edition.* USA: Routledge Academic, 2002.

Corwin, John, Avi Sukberscgatz, Perry Miller, and Luis Marenco. "Dynamic Tables: An Architecture for Managing Evolving, Heterogeneous Biomedical Data in Relational Database Management Systems." *JAMIA*, 2007: 86-93.

Crick, Francis. *What Mad Pursuit.* United States of America: Perseus Books, 1988.

Dahly, D, L Adair, and K Bollen. "A structural equation model of the origin of blood pressure." *International Journal of Epidemiology*, 2009: 538-548.

DiMatteo, M., P. Giordani, H. Lepper, and T. W. Croghan. "Patient adherence and medical treatment outcomes: a meta-analysis." *Medical Care*, 2002: 794-811.

D'Onofrio, B.M., Hulle, C.A., I.D. Waldman, J. L. Rodgers, K.P Harden, P.J. Rathouz, and B.B Lahey. "Smoking during pregnancy and offspring externalizing problems: An exploration of genetic and environmental confounds." *Development and Psychopathology*, 2008: 139-164.

Elmasri, Ramez, and Shamkant Navathe. *Fundamentals of Database Systems 5th Edition.* USA: Addison Wesley, 2006.

Enders, Craig. *Applied Missing Data Analysis.* USA: Guilford Press, 2010.

Enders, Craig, and Davood Tofigi. "Centering predictor variables in cross-sectional multilevel models: A new look at an old issue." *Psychological Methods*, 2007: 121-138.

Fan, Xiato. *SAS for Monte Carlo Studies.* Carey, North Carolina: SAS Institute, 2002.

Fisher, R.A. "The Correlation of Relatives on the Supposition of Mendelian Inheritance." *Transactions Royal Society of Endinburg*, 1918: Endinburg.

Fisher, Ronald. "The Fiducial Argument In Statistical Inference ." *Journal of Eugenics*, 1935: 391-398.

Goldstein, H. "Multilevel Variance Component Models." *Biometrika*, 1986.

Gossett, Willaim Sealey. "The Probable Error of The Mean." *Biometrika*, 1908: 1-25.

Gossett, William Sealey. "An experimental determination of the probable error of Dr Spearman's correlation coefficients." *Biometrika*, 1921: 263-282.

Gossett, Willian Sealey. "Probable Error Of A Correlation Coefficient." *Biometrika*, 1908: 302-310.

Hager, G, and G Wellein. *Introduction to High Performance Computing for Scientists and Engineers.* United States of America: Chapman & Hall, 2010.

Haldene, J. "A Note on Non-Normal Correlation." *Biometrika*, 1949: 467-468.
Hall, Asaph. "On an experimental determination of PI." *Messenger of Mathematics*, 1872.

Halpern, M T, et al. "Recommendations for evaluating adherence and persistence with hypertension therapy using retrospective data." *Hypertension* 47 (2006): 1039-48.

Jeffreys, Harold. "Note on Behrens Fisher Formula." *Annals of Eugenics*, 1940: 48-51.

Johnson, S, and P Stafford. "Immunosignaturing to Profile Humoral Responses." 2009.

Johnson, W., J. Brown, G. Harootunian, D. Petitti, Y. Qui, and T Sama. *Phoenix Healthcare Value Measurement Initiative Report.* Phoenix: Center for Health Information and Research, 2011.

Keppel, G, and T Wickens. *Design and Analysis 5th Edition.* United States of America: Prentice Hall, 2007.

Kowalski, C. "On The Effects of Non-Normality on the Distribution of the Sample Product Moment Correlation Coefficient." *Journal of the Royal Statistical Society*, 1972: 1-12.

Kustra, R, R Shioda, and M Zhu. "A Factor Analysis Model for Functional Genomics." *BMC Bioinformatics*, 2006: 207-216.

Lee, J Y, et al. "Assessing medication adherence by pill count and electronic monitoring in the African American Study of Kidney Disease and Hypertension (AASK) pilot study." *Am J Hypertens* 9 (1996): 719-25.

Legutki, J, M Magee, P Stafford, and S Johnston. "A General Method or Characterization of Humoral Immunity Induced by a Vaccine or Infection." *Vaccine*, 2010: 4529-4537.

Litman, G, J Cannon, and L Dishaw. "Reconstructing immune phylogeny: new perspectives." *Nature Review*, 2005: 866-879.

Lohr, S. *Sampling: Design and Analysis 2nd Edition.* United States of America: Duxbury , 2009.

Longford, N. "A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects." *Biometrika*, 1987: 817-827.

Mallion, J M, J P Baguet, J P Siche, F Tremel, and R de Gaudemaris. "Adherence, electronic monitoring and antihypertensive drugs." *Hypertens* 16, no. suppl (1998): S75-S79.

Martin, L. R., S. L. Williams, K. B. Haskard, and M. R. & DiMatteo. "The challenge of patient adherence." *Clinical Risk Management*, 2005: 189-199.

Meichenbaum, D, and D Turk. *Patient compliance; Medical personnel and patient; Sick; Patient Education.* United States of America: Plenum Press, 1987.

Metchnikoff, E, and F.G. Binnie. *Immunity in Infective Diseases.* Cambridge: Cambridge University Press, 2009.

Metropolis, Nicholas. "The Beggining of the Monte Carlo Method." *Los Alamos Science*, 1987: 125-130.

Metropolis, Nicholas, and Stanley Ulam. "The Monte Carlo Method." *Journal of the American Statistical Association*, 1949: 335-341.

Mills, E, et al. "Adherence to Antiretroviral Therapy in Sub-Saharan Africa and North America." *JAMA*, 2006: 2479-2485.

Musen, M, and J Van Bemmel. "Challenges for medical informatics as an academic." *Methods in Information Medicine*, 2004: 1-3.

Muthen, Bengt, and Linda Muthen. *Mplus Version 6.* USA: Mplus, 2011.
Nationanl Cancer Institute. *cBIG.* 2 20, 2012.

https://cabig.nci.nih.gov/community/concepts/caDSR/ (accessed 2 20, 2012).

New England Health Institute. "Improving Physician Adherence to Clinical Practice Guidelines." 2008.

Neyman, Jersey, and Egon Pearson. "On The Use And Interpretation Of Certain Test Criteria For Purposes of Statistical Inference." *Biometrika*, 1928: 263-294.

NIH Common Fund. "About the NIH Common Fund." June 16, 2011. http://commonfund.nih.gov/about.aspx (accessed September 3, 2011).

NIH. "Meeting the Challenge of Big Data in Biomedical and Translational Science." August 2, 2011.
http://commonfund.nih.gov/InnovationBrainstorm/?tag=/biomedical-data (accessed August 3, 2011).

Parker, J, et al. "Supervised Risk Predictor of Breast Cancer Based on." *Journal of Clinical Oncology*, 2009.

Pearson, E.S. "Some Notes on Sampling Tests with Two Variables." *Biometrika*, 1929: 337-360.

Pearson, E.S. "The Test of Significance for the Correlation Coefficient." *Journal of the American Statistical Association*, 1931: 128-134.

Quackenbush, John, Helen Causton, and Alvis Brazma. *Analysis, Microarray Gene Expression Data.* United States: Wiley-Blackwell, 2003.

Rao, C, H Toutenburg, Shalab, C Heumann, and M Schorn. *Linear Models and Generalizations: Least Squares and Alternatives.* United States of America: Springer Series in Statistics, 2010.

Raudenbush, S, and A Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods.* USA: Sage Publications, 2001.

Rider, P. "The Distribution of the Correlation Coefficient in Small Samples." *Biometrika*, 1932: 382-403.

Ron, D. H., et al. "The impact of patient adherence on health outcomes for patients with chronic disease in the medical outcomes study." *Journal of Behavioral Medicine*, 1994: 347-360.

Rubinstein, R, and D Kroese. *Methods, Simulation and Monte Carlo.* United States of America: Wiley Series in Probability and Statistics, 2007.
SAS Institute. *SAS Corporate Statistics.* 2 26, 2012. http://www.sas.com/company/about/statistics.html (accessed 2 25, 2012).

SAS Institute. *SAS User Guide 9.2 Second Edition.* Carey, North Carolina: SAS, 2011.

Seliegman, B, and J Brown. *Health and Human Services NCI Basal Breast Cancer Diagnostic Grant HHSN261201000131C.* Grant Phase 1 Report, Tucson: High Throughput Genomics, 2011.

Seock-Ho, Kim, and Allen Cohen. "On The Behrens Fisher Problem: A Review." *Psychometric Society*, 1995: 1-40.

Shamlin, D. " Threads Unraveled: A Parallel Processing Primer." *SUGI*, 2004: 217-229.

Shamlin, D. "Threads Unraveled: A Parallel Processing Primer." *SUGI*, 2009.
Shortliffe, Edward, and James Cimino. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine.* USA: Springer, 2006.

Stokes, J, D Bradstreet, and M Hill. "SAS/Connect® Simply Stated." *SUGI*, 2002: 109-127.

Storey, J. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society*, 2002.

Stulberg, J, C Delaney, D Neuhauser, D Aron, and S Koroukian. "Adherence to Surgical Care Improvement Project Measures and the Association with Postoperative Infections." *JAMA*, 2010: 2479-2485.

Stulberg, J, C Delaney, D Neuhauser, D Aron, P Fu, and S Koroukian. "Adherence to Surgical Care Improvement Project Measures and the Association with Posoperative Infection." *JAMA*, 2010: 2479-2485.

Stulberg, J. Delaney, C., D. Neuhauser, D. Aron, P. Fu, and S. Koroukian. "Adherence to Surgical Care Improvement Project Measures and the Association With Postoperative Infections." *JAMA*, 2010: 2479-2485.

Tabachnick, B, and L Fidell. *Using Multivariate Statistics Fifth Edition.* USA: Allyn & Bacon, 2006.

Tu, Yu-Kang. "Is Structural Equation Modeling A Step Forward for Epidemiologists." *International Journal of Epidemiology*, 2009: 549-551.

Tukey, John. "A survey of sampling from contaminated distributions." *Contributions to Probability and Statistics*, 1960: 448-485.

Tukey, John, and D McLaughlin. "Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorizing." *Sankhya*, 1963: 331-352.

Vermeire, H., H. Hearnshaw, P. Van Royen, Denekens, and J. "Patient adherence to treatment: three decades of research. a comprehensive review." *Journal of Clinical Pharmacology Therapy*, 2001: 331-342.

Wilcox, Rand. "ANOVA: A Paradigm For Low Power and Misleading Effect Size?" *Review of Educational Research*, 1995: 51-77.

Wright, Sweall. "Correlation and Causation." *Journal of Agricultural Research*, 1921: 557-585.

Zolnierek, H., and M. & DiMatteo. "Physician communication and patient adherence to treatment: a meta-analysis." *Medical Care*, 2009: 826-834.