5-2018

Precision Medicine: Viable Pathways to Address Existing Research Gaps

Kathryn Kennedy Pegues
Clemson University, kpegues@gmail.com

PRECISION MEDICINE: VIABLE PATHWAYS TO ADDRESS EXISTING RESEARCH GAPS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Industrial Engineering.

---

by
Kathryn Kennedy Pegues
May 2018

Accepted by:
Dr. Byung Rae Cho, Committee Chair
Dr. Mary E. Kurz
Dr. David Neyens
Dr. Robert Riggs

ABSTRACT

Precision Medicine (PM) seeks to customize medical treatments for patients based on measurable and identifiable characteristics. Unlike personalized medicine, this effort is not intended to result in tailored care for each patient. Instead, this effort seeks to improve overall care within the medical domain by shifting the focus from one-size-fits-all care to optimized care for specified subgroups. In order for the benefits of PM to be expeditiously realized, the diverse skills sets of the scientific community must be brought to bear on the problem. This research effort explores the intersection of quality engineering (QE) and healthcare to outline how existing methodologies within the QE field could support existing PM research goals. Specifically this work examines how to determine the value of patient characteristics for use in disease prediction models with select machine learning algorithms, proposes a method to incorporate patient risk into treatment decisions through the development of performance functions, and investigates the potential impact of incorrect assumptions on estimation methods used in optimization models.

DEDICATION

This document is dedicated to my children, Siena and Bryan, whose unwavering love and gentle encouragement to “hurry up” pushed me to the finish line.

ACKNOWLEDGMENTS

I owe a debt of gratitude to my exceptional network of family, friends, and colleagues who have supported me as I relentlessly pursued my dream. In particular, I would like to thank the following:

· J. Adam Pegues, my husband, for his faith in me, our family, and our future together

· Dr. Marian Kennedy, my sister and hero, for her push to turn in my application to Clemson University

· Dr. Cho, my advisor, for his support through the process

· Dr. Dylan Morris, my brother-in-law, for keeping all of us laughing with his quick wit and being the uncle my kids adore

· Mari & Michael Notley, my aunt and uncle, for their love and guidance throughout my life

·And last, but not least, the women who have been by my side as I have navigated the challenges of professional life and parenthood: Christina Painter, Elizabeth Weaver, Alexia Machina, Alicia Pruitt, Nikki Blystone, Gail Dwyer, and Christy Thomas.

TABLE OF CONTENTS

Page

Table of Contents (Continued) Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER ONE

INTRODUCTION TO RESEARCH

1.1 Introduction

1.1.1 Research Motivation and Scope

In 2008, the National Academy of Sciences outlined the top 14 priorities for research in the "Grand Challenges for Engineering." Of the research priorities highlighted in the document, three challenges are directly related to healthcare: advanced health informatics, engineering better medicines, and reverse-engineer the brain (National Academy of Engineering, 2018). The inclusion of healthcare challenges reflects the importance of medical advancements in comparison to other national and global issues. In spite of the research progress in this area, there remains a significant need for continued refinement in the understanding of how the human body operates and development of improved treatment techniques for identified ailments. Finding solutions to the identified healthcare research challenges necessitates a multidisciplinary effort that spans partners in government, industry and academia.

The primary aim of this dissertation is to outline the potential role of quality engineers in addressing healthcare challenges. Chapter 1 will briefly outline the motivation for this research endeavor by touching on the financial implications to society of the current healthcare system. In the following section, a brief explanation of prevision medicine, a relatively-new initiative aimed at improving medical care through by providing tailored care for groups of people with matching characteristic profiles. If

successful, this research effort will spur current innovators in the field of quality engineering to consider future healthcare centric research projects.

1.1.2 Drivers for Systematic Improvement within Healthcare

The cost of healthcare in the United States consumes a larger percentage of available wealth with each fiscal year. In 2015, United States' healthcare spending reached $3.2 trillion and accounted for 17.8% of the gross domestic product (Martin et al., 2016). While the rising healthcare costs are attributable to a wide range of causes, the three primary contributing factors were the rising percentage of the population using available healthcare resources, increased utilization of services by the individual, and growing cost for specific medical services (Martin et al., 2016). The increased cost for specific medical procedures most likely reflects changes to healthcare policy, adjustment to patient treatment protocols, or attempts by the medical establishment to more accurately distribute overhead costs to the individual consumer. One example of a procedure whose total cost increased over the past decade is pediatric spinal fusion. From 2000 to 2013 the price of the procedure grew from $29,930 to $56,920 (National Center for Health Statistics, 2016).

The relentless growth of healthcare costs has fueled concerns regarding long-term affordability of national healthcare programs at the national level and affordability of care by individuals. By addressing inefficiencies within the domain of healthcare delivery, research teams may potentially stabilize or decrease the total cost of healthcare to the nation and for individuals.

1.1.3 Precision Medicine: Emerging Approach for Transforming Healthcare

As identified by Berwick et al.(2008), addressing the cost of healthcare will necessitate improving quality of patient treatment and developing additional preventative measures for disease. The accomplishment of these two objectives results in a reduction of per-capita care costs. In order to reach these goals, research teams must first focus improving understanding of the complex system that is the human body and how the system reacts to both disease and treatment.

Medical professionals have a limited number of diagnostic tools and treatment options at their disposal with which to assess and treat patients. In the healthcare system that exists today, misdiagnosis, missed diagnoses, and poor response to treatment still occur. The reasons for each of these issues vary. However, at the heart of the matter is the need to be able to measure patient's health characteristics, the ability to relate the characteristics to a set health complication, and to be able to provide the patient a treatment protocol that will have a positive effective.

In 2015, the federal government launched the Precision Medicine Initiative (PMI), a research effort aimed at changing the treatment of patients from a one-size-fits-all approach to a treatment approach that takes into account individual differences between patients. Precision medicine (PM) is defined as "an approach to disease treatment and prevention that seeks to maximize effectiveness by taking into account individual variability in genes, environment, and lifestyle" (Hudson et al., 2015). The research initiative has been touted for its potential to revolutionize the treatment of disease. In his 2015 State of the Union address, President Obama ignited interest in PM by stating

“Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type – that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?”(Obama, 2015). His words provided a compelling vision of how the advancement of PM could revolutionize the care of patients and improve medical outcomes.

As the underlying knowledge needed to support the application of PM grows and as the application of PM becomes more common place within the healthcare system, more medical professionals will able to use patient features to select the best treatment protocol based on likelihoods of positive treatment response for populations with similar features. To that end, research is needed to illuminate which patient factors are the best indicators of health for each disease and additional research needs to assess treatment protocol response for groups identified by common patient characteristics. It is in this area of healthcare research that quality engineers will find a problem set that matches their skill set. The combination of the statistics, decision analysis, optimization, and process development are all critical components of increasing the probability of applying the best treatment for each individual at the right price within the shortest window of time possible.

PM is poised for greater gains in the coming years due to the increase of existing biologic repositories, improved analytic methods to identify subpopulations, and the refinement of computational tools used to find optimal solutions (Collins & Varmus,

2015). Unique patient groups routinely examined in medical research are defined by common patient features which may include genetic differences, environmental factors, or lifestyle choices. The study of response differences between patient groups allows researchers to identify patients groups who are more susceptible to a disease or respond differently to specific treatment plans. One significant advancement in the medical field that greatly affected the future of precision medicine was the ability to map an individual's deoxyribonucleic acid (DNA). Since DNA is unique to the individual and dictates how that particular body functions, this knowledge may explain why treatments are effective in some patients, but not in others. By determining which genes affect drug metabolism, an individual's genes can be used to screen out treatments which will not be effective. This course of action will increase the likelihood of a patient receiving immediate relief instead from the prescription medication and decrease the likelihood of a repeat office visit. There remains a need to find additional differences between individuals which affect treatment response so that patient profiles may be used to inform treatment selection to predict an individual patient's response to a specific drug treatment.

1.1.4 Overlap of Precision Medicine and Quality Engineering

Precision medicine will transform medical care in two ways. First, it will improve prevention and diagnosis by improving the ability to identifying differences between individuals that are healthy, at risk for a future complication, or have a health complication. Secondly, it will improve the likelihood of assigning the patient an optimal treatment strategy with the least number of remedial visits for the same ailment. In order for PM to be effectively applied in practice, medical professionals must have the ability

to compare the likelihood of effectiveness of different treatment regimens for different patient profiles. Quality engineers have the requisite skills to make significant contributions to in both of these areas.

Quality engineers have knowledge to explore possible causes for variance within treatment response and assess patient risk when undergoing treatment. In the past, quality engineers have focused primarily on applications within the manufacturing sector but their knowledge is frequently applied in other applications areas. The quality engineering field is well known for the design of processes, a tool which enables production at consistent high quality outputs with few defects. Quality engineers analyze processes to determine how to produce predictable and accurate results. In the advancement of precision medicine, quality engineers can help the medical community to determine the optimal level of a measurable characteristic for a specific subpopulation, investigate optimal treatment strategies that take into account multiple ailments and combinations of drugs, and predict the impact of treatment strategies were based on incorrect assumptions.

A literature review of major quality engineering journals revealed that most of the previously published research which combined these domains are editorials or offer an analysis at the macro scale. While the author cannot definitively provide a rationale for the lack of overlap, one reason might be the qualitative nature of medicine in the past century. At the start of the twentieth century, doctors were reliant on a limited number of sources to obtain information about s patient's health when making a diagnosis. Medical providers would gain data from a visual inspection of the patient, measurements from a limited number of medical tests, and qualitative information from the patient's

perspective regarding ailment and treatment response. The scope and sensitivity of measurement tools available to medical professionals continue to improve as technology continues to advance. As a result, practicing medical professionals now have more "data" at their fingertips than medical professionals had fifty years ago. In order to optimize patient care, doctors are in need of improved methods of analyzing data and advisement on how best to integrate the results into the decision making cycle. Quality engineers have both the technical acumen and comfort of operating in uncertainty to provide support to the medical community as it moves forward. Another reason for the lack of involvement of quality engineers within the medical domain could also be due the nature of the problem. In manufacturing endeavors, engineers seek to improve the output of a process. The variability between products is limited and the changes to the assembly line are assumed to impact items on the line in similar fashion. Unlike manufacturing, the product examined in the medical system, an individual's health, is measurable in a vast number of ways. In addition, the target value for a measured characteristic cannot be precisely determined. Instead, healthcare providers have utilized a range of values to assess an individual's health. The turn to precision medicine could reduce the measured differences between individuals as smaller subgroups are identified. The advancements to reducing variability in treatment response based on newly-measured characteristics could be as impactful as the improvement to transplant success once blood type differences are recognized.

1.2 Research Outline and Objectives

This dissertation explores how best to use quality engineering tools within the healthcare domain. Three separate, viable pathways for pursuing improvements in healthcare are examined with an eye toward evaluating patient risk. To start, chapter 2 summarizes literature published within the past five years which used machine learning for research on T2DM. The two most important outcomes are a consolidated list of the research goals and a summary of research limitations from the point of view the researchers. Next, in chapter 3, a new method for risk quantification using an adaptation of quality loss functions is described. Whereas quality loss functions were developed for point targets, a new performance loss function could be used for risk assessment when measure characteristics are evaluated within a target range. Performance functions would allow medical professionals to assess the potential impact of a treatment on a patient across multiple characteristics of interest. The last chapter explores the impact of incorrect assumptions during parameter estimation on optimization outcomes. In total, the dissertation provides insights into the potential impact and the challenges of medical research. Each of the chapters is summarized in more detail in the following paragraphs.

Chapter 2 is a systematic review of the use of machine learning efforts that support the advancement of precision medicine for a selected disease. This chapter outlines a basic methodology for assessing the state of medical research for an analytic tool as applied on a specific disease. Type 2 diabetes mellitus (T2DM) was selected as a focus area because of its high prevalence within the population. As of 2017, an estimated 9.4% of the United States' adult population is afflicted with diabetes mellitus (DM), a

group of chronic diseases which affect insulin production within the human body (Center for Disease Control and Prevention, 2017). In the most recent published assessment of the financial impact of DM by the American Diabetes Association, the organization estimated that DM cost United States at least $245 billion annually (American Diabetes Association, 2013). The tally includes both the cost of medical treatment for DM and cost of lost workforce productivity. To mitigate the future impact of DM, researchers must develop more effective means of preventing the development of the disease in more patients and improve treatment methodologies to improve patient outcomes. To achieve those goals, researchers are exploring the potential of PM, an emerging approach to patient care that seeks to customize medical treatments based on measurable and identifiable characteristics. If the promise of PM is realized, the shift of medical care from the one-size-fits-all convention to optimized care for specified subgroups will improve medical outcomes. This literature review examines the use of machine learning to achieve PM aims for T2DM. The paper outlines major T2DM research areas, the most common algorithms utilized for research, and the measures of effectiveness used to assess their performance. This work also provides insights into the limitations that decrease the potential of the current research efforts.

Chapter 3 lays out the motivation for continued research at the convergence of healthcare and quality engineering. It examines the state of healthcare, the needs which have motivated new research, and past efforts of quality engineers to influence medical research. At the end of the chapter, areas for potential research are identified.

In Chapter 4 the potential of robust design is examined when it is paired with conditions based selection of regression estimators. At the start of the chapter, alternative methods of estimating parameters when the underlying distribution is unknown are compared and contrasted. If the researcher's assumptions regarding the underlying distribution are correct, results found using optimization models developed using the estimated parameter will not be impacted. However, if the researcher's assumptions regarding the parameter are proven incorrect, the results of optimization efforts using the parameters will be impacted. This chapter explores the potential impact of inaccurate assumptions made during parameter development phase of research. For illustrative purposes, the hypothetical research team assumes that the underlying distribution is normal when, in actuality, the underlying distribution is skew normal. This chapter also provides insight into the impact of incorrect assumptions during made during early phases of research on final recommendations. The analysis is particularly important for parameter estimation supporting medical applications since researchers may not know the underlying distribution. While medical researchers work to better describe physical phenomena, a parallel effort within the engineering community should focus on the development of improved methodologies for parameter estimation when the distribution may be non-normal.

Each chapter practically describes how quality engineers could apply their skills to support the development of precision medicine. Chapter 2 explores how machine learning is being applied to bridge identified research gaps for one prevalent disease. The three most important products from this chapter are a list of research needs, an

assessment of the types of quality characteristics being used to gauge health, and assessment of the limitations of current research due to limited data availability. Chapter 3 illustrates how informed adaptation of current QE methodologies could improve the assessment of patient risk when undergoing treatment with known, measurable side effects. Chapter 4 critically considers the impact of incorrect assumptions early in the analysis. The dissertation provides a foundation from which other quality engineers will be able to craft innovative research efforts for the continued development of tools needed for PM.

CHAPTER TWO

TYPE 2 DIABETES MELLITUS: OBTAINMENT OF THE PROMISE OF PRECISION MEDICINE WITH MACHINE LEARNING METHODS

2.1 Precision Medicine Applied to Diabetes Mellitus

The term diabetes mellitus (DM) refers to a group of chronic diseases distinguished by hyperglycemia, abnormally high blood glucose. The rise of glucose within an individual's blood stream can be either attributed to the insufficient production of insulin within the body, a physical resistance to insulin, or a combination thereof (American Diabetes Association, 2017). The increased blood sugar negatively affects the function of important organs to include the heart, eyes, vessels and kidneys among others (Pippitt & Li, 2016).

The importance of the disease may be attributed to the prevalence of the disease worldwide and the resultant costs. From 1980 to 2014, the number of individuals with DM has risen from approximately 108 to 422 million (World Health Organization, 2016). As of 2017, 7.2% of the United States' adult population was diagnosed with DM (Centers for Disease Control and Prevention, 2018). An additional estimated 7.2 million adults have the disease, but have not been diagnosed. In 2012, the American Diabetes Association funded research to quantify the total cost of DM for the United States. The final report estimated that within the United States over $176 billion was annually spent for direct medical costs of DM and another $69 billion was lost due to decreased productivity (American Diabetes Association, 2013). The DM related medical costs to the individual is estimated to be significant as well. The average annual medical expenditure for a patient diagnosed with diabetes in 2012 was on average $13,700 with

$7,900 of that cost directly relating to diabetes. Past research efforts have shown that early diagnosis and proper management of the disease can improve an individual's health and reduce risk for further complications.

Current medical research continues to focus on early diagnosis and treatment of DM. Within the field of medicine, one particular area, PM has shown promise for advancements in patient treatment. PM focuses on finding the best treatment for a patient based on the individual's characteristics which may include "genetic, biomarker, phenotypic, or psychosocial traits" (Jameson & Longo, 2015). The success of PM hinges on the ability to classify individuals into groups of susceptibility and treatability for a particular disease or a combination of diseases using measurable characteristics. Two challenges to PM include resolving competing healthcare system stakeholder interests and the challenge of dealing with a vast, continuously growing, and complex data set (Jameson & Longo, 2015). The first challenge will require changes to government policy to realign stakeholder interests into a more mutually beneficial system. To overcome the second challenge, efficiently and quickly, the involvement of other parts of the scientific community will be required. This challenge involves wrestling large medical data repositories in an attempt to find new medical knowledge through the identification of subpopulations and unpredicted responses to treatment plans. As such, the size of the data and the complexity of the problem should make working on medical problems a desirable application area. The complexity of medical care provides an interesting area for application of other skill sets. Hence, the classification of a disease within a patient will depend on more exact definitions requiring adjustment to decision algorithms.

Machine learning (ML) is a branch of artificial intelligence that gives machines the ability to learn and automate analytical models for classification and predictions with big data. PM research efforts focused on DM are making significant progress in identifying the portion of the population most at risk for developing DM and in improving treatment methodologies. This chapter seeks to document the state of current ML research efforts using published literature. It will outline the general research goals for type 2 diabetes mellitus (T2DM), the types of approaches, sources for data, and limitations of recent work. It is the hope of the author that this work will provide a foundation for future research efforts involving T2DM and ML, the study of algorithms to uncover insights within a dataset and to develop models for prediction.

In the following sections, an overview of DM and basics of ML will be discussed. Section 2.1.1 provides a brief overview of the disease and outlines the reasons for narrowing the scope of this research effort on T2DM, a single variant. Section 2.1.2 examines the broad categories of ML and briefly discusses the most common types of ML algorithms. The section is concluded with research considerations when attempting to use ML. Section 2.2 provides a methodology for the literature review. Specifically this section covers the search criteria, inclusion and exclusion criteria. Section 2.3 summarizes the results of this research effort. Section 2.4 provides a way ahead for T2DM research involving ML and suggestions for how to improve medical data repositories.

2.1.1 Diabetes Mellitus Overview

As stated earlier, the disease DM is caused by malfunctions affecting the amount of insulin within the human body. Insulin is a hormone produced by beta cells in the pancreas. The hormone regulates the amount of glucose within the blood stream. Too little insulin within the body results in high levels of glucose which is known as hyperglycemia. The impact of DM on the patient's health is affected by a variety of factors including the severity of type of DM, the speed with which DM is diagnosed and treated after initial onset, the effectiveness of the treatment protocol, and the patient's adherence to the prescribed treatment protocol. If a glucose level with the patient's blood stream remains above the recommended threshold for an extended period of time, the patient is at greater risk for serious health complications to important organs. The disease is linked to damage to eyes, kidneys, nerves, and the heart.

The most prevalent types of DM are defined by their etiopathogenesis and are referred to as type 1, type 2, and gestational diabetes. Classification is important since the disease progression for each variant of DM is different. Type 1 diabetes mellitus (T1DM) occurs when the immune system produces antibodies which attack beta cells in the pancreas. The presence of antibodies within the blood stream is an indicator of T1DM (American Diabetes Association, 2017). The disease progression for T1DM depends on the how early antibodies are detected and the number of antibodies detected. Once a patient has type 1 diabetes, the patient requires treatment with exogenous insulin to facilitate metabolic survival (Atkinson, 2014). T1DM affects approximately five percent of the population diagnosed with DM (Centers for Disease Control and Prevention,

2017). In comparison, T2DM occurs when either the body does not produce enough insulin or the body is resistant to the effects of insulin. This variant of DM accounts for approximately 90-95% of the DM cases (Centers for Disease Control and Prevention, 2017). The disease most often appears in individuals once they reach adulthood with only 132,000 people under the age of 18 in the United States diagnosed with any form of DM (American Diabetes Association, 2017). Like T1DM, this form of the disease creates an increase in glucose within the patient's blood stream. Unlike T1DM, insulin is typically only required for disease management, but not necessarily survival. The third most common form of DM is referred to as gestational diabetes (GD). GD occurs during pregnancies and is considered a temporary condition. Most often, the clinical signs of gestational diabetes will disappear after the birth of a child. Some patients, however, do progress from diagnosis of GD to T2DM after the birth of a child.

While the impact of all three forms of DM is significant, this chapter will focus specifically on T2DM because it affects a greater portion of the population. Of the 23.1 million people in the United States diagnosed with DM, 90-95% of those patients have T2DM (Centers for Disease Control and Prevention, 2017). This form of the disease is considered progressive with symptoms becoming more intense over time. The initial onset of the disease is not always recognized by the patient because the short-term symptoms may not be distinguishable. T2DM symptoms may include increased thirst, weight loss, or increased need to urinate more frequently. The impact of the hyperglycemia, the presence excess glucose in the bloodstream, on the patient, may progressively get worse over time as either the individual's resistance to insulin grows or

as the gap between the insulin needed by the body to regulate glucose and the amount of insulin produced increases. Long-term complications of having hyperglycemia may result in additional complications including blindness, loss of limbs due to poor blood flow, or kidney failure. However, unlike T1DM, T2DM may be partially preventable through behavior changes and the disease may go into remission given a reduced severity of the form of DM and patient response to treatment protocol. Accordingly, early detection of the disease and proper management is critical for the health of the patient. The continued investment in PM research for T2DM is made with the goal of improving detection of the disease and determining of the best treatment protocols for specific patient profiles.

Given the prevalence of the disease, the medical community has collected a vast amount of data concerning T2DM. As stated earlier, one of the primarily challenges of PM is how to effectively use the data to develop insights that illuminate patient characteristics which best align with increased incidence of the disease. Within recent literature, one prevalent method for investigating T2DM has been the use of ML. The following section provides a brief overview of the topic.

2.1.2 Machine Learning: The Basics

The field of ML exists at the intersection of computer science and data science. ML was built on the premise that computer systems have the ability to improve the specified task completion without the necessity of successive improvements to the initial implementation being programmed by the user. Computer systems use algorithms to develop knowledge about a dataset. Feedback on performance enables the computer to make adjustments to calculated predictions or decision recommendations. The increased

computational power realized over the past thirty years has added increased capability to the field. As a result, feedback on performance enables the computer to make adjustments to calculated predictions or decision recommendations.

Today ML is used to develop a greater understanding in research areas with a vast amount of data. The tool also helps researchers identify patterns within the dataset that are not obvious. In fact, ML has been applied to a broad spectrum of areas. It has been used for identification, speech recognition, and statistical arbitrage (Pazzani et al, 1998)(Graves et al., 2013)(Galindo & Aamayo, 2000) . ML also provides a means of quickly detecting oil spills from radar images of the ocean's surface (Kubat et al, 1998). Moreover, ML has helped to close the gap between automatic speech recognition systems in comparison to human performance (Deng & Li, 2013). In addition, ML has also been heavily relied upon to develop quantitative training strategies for financial assets to include hedge funds, the field referred to as statistical arbitrage (Krauss et al, 2017). The investment strategy looks for patterns within financial data streams to identify patterns for exploitation. The field of ML is growing quickly due to high interest from both the government and industrial sectors. In 2016, McKinsey Global Institute estimated that machine learning received between $5 and $7 billion dollars in investment funding (Bughin et al, 2017).

The term "machine learning" encompasses a broad field of work which uses multiple techniques for a wide range of applications. Within the field, learning is classified as one of three major tasks: supervised, unsupervised, or reinforcement. A supervised learning task infers a relationship between inputs and outputs. Algorithms

classified as “supervised” require feedback during the training stage. The training dataset includes both the input data and associated outputs, also termed “supervisory data.” The algorithm uses the training data to develop an inferred function that relates inputs to outputs. Once a base model is formed, the model is tested using a validation dataset, a portion of the training dataset held in reserve. Performance is then judged on the ability of the algorithm to correctly link inputs to outputs. The parameters of the model are then adjusted to improve the accuracy of the model, and the model is used for prediction or classification purposes. Common metrics used to compare supervised ML performance between algorithms are accuracy, sensitivity, and specificity. Robust functions will have the capability of correctly analyzing samples that were not specifically included within the training data set. Unlike supervised ML, unsupervised learning algorithms only use input data to develop knowledge about the data set. The algorithms deduce relationships between the predictor variables. Since this type of learning does not have known outcomes, there is no means of evaluating the accuracy of the final model. The third type, reinforcement learning examines the trade space between exploration and exploitation.

When discussing ML, the three most important aspects are the purpose of the algorithm, the type of learning, and the data set used to train the algorithm. Typically supervised ML algorithms are used to perform two types of tasks: classification or prediction. For classification, a model is developed that assigns inputs into the system into a predefined class in the system. For diabetes research, classification algorithms can be used to determine if a patient is at risk for developing diabetes. Whereas classification algorithms are used to predict the correct group, regression modeling is used to

strengthen the ability of predictive analytics. For diabetic research, machine learning can be helpful in determining the proper dosage of insulin for patients based on individual characteristics. In this way, ML has the potential to provide power to the application of PM within healthcare. This tool has proven effective for the use of both continuous, discrete, and mixed data sets.

2.1.2.1 Supervised Learning Algorithms

There are a wide variety of algorithms in use today within the field of ML. It is commonly acknowledged that there is no single algorithm that works better across the wide variety of supervised learning problems. This type of ML is used for both classification and regression purposes. For classification methods, the output value is a category. One example of classification for T2DM research is that based on input data a patient could either be classified as either having diabetic retinopathy or not having diabetic retinopathy. For the purposes of regression, however, the output value from the model would be a real number. If considering the application within PM, this type of model could be used to determine the optimal dose of insulin for a patient. When applying supervised ML techniques, researchers must be aware of two typical issues that commonly occur. First, the researcher finds a balance between over-fitting and under-fitting the data. This is often referred to as the bias-variance trade-off. Under-fitting occurs when bias exists causing the algorithm to not identify a relationship between the independent variables and their associated dependent variables. Contrarily, over-fitting occurs when the developed model fits the training data too closely. When the model is subsequently used with other data, these exists high sensitivity to small changes in the

input variables. Analysis of the bias-variance may be captured in a discussion of the algorithm's expected generalization error. Secondly, the final outcome is dependent on the quality of the training dataset used to develop the ML algorithm. The algorithm is used to develop a function to be used for predictions or classification by relating known input and output values. Training datasets that are too small may not provide enough instances for which the resultant model may not be robust to a variety of input combinations. If the training dataset has a large number of input variables, the chosen ML methodology must be able to effectively judge which features are critical to optimal model development. In addition, the dataset may have missing entries, infeasible values or outliers. Researchers must determine how to best process the data to ensure that it is adequate for the intended purpose.

For supervised learning, the comprehensive analysis of a dataset is extremely important in the development of the inferred solution because the training dataset connects the input data to output values. The supervised learning process is iterative. After the algorithm develops a solution based on the training input, the algorithm learns by comparing its generated output, the prediction for a given set of input values, against the true output value. The algorithm is "correct" if the function's output matches the training set output. Training stops when the algorithm reaches an acceptable level of performance based on output values. The algorithm's performance is then validated with the portion of the available data held in reserve. Development of a training methodology for model development and development of a validation strategy to assess model performance are critical steps in supervised learning. Probably the most important factor

to take into consideration when conducting supervised learning is the size of the training dataset. If the set is large, the researcher may choose to just divide the available data into two subsections: a training dataset and a validation dataset. The parameters for the model will be developed using the training set and the model will then be verified with the validation dataset. While an established standard for the division of the dataset does not exist, common convention dictates that the data be divided proportionally 2/3 to 1/3 (training to validation). The other method, known as cross-validation, divides the dataset into mutually exclusive sections of equal size. Iteratively, one of the sections will be held for testing performance of the algorithm trained by the other subsets. The final model will be formed by combining the results of each iterations. The following paragraphs provide a brief overview of the most prevalent ML algorithms.

a. Naive Bayes Algorithm

The naïve Bayes (NB) algorithm uses training data to develop frequencies for each possible outcome which provides the class prior probability (Rish, 2001). The algorithm can then determine the posterior probability for each of the possible outcomes. The outcome with the highest posterior probability then becomes the prediction. NB needs less training data in comparison to other types of algorithms. Three considerations when utilizing the NB algorithm are the necessity for independence of predictors, the reliance on all outcomes being observed in the training data, and the known performance of NB in producing estimators. If a variable is not observed in the training dataset, the probability for that outcome will be zero and the ability to make a prediction is

eliminated. The algorithm has proven to be valuable for both real time predictions and for instances where there are multiple classes.

b. Decision Tree Algorithm

The decision tree (DT) algorithm is one of the most prevalent ML algorithms and as such it has been extensively studied in literature. DT is non-parametric, simple to use, and can effectively be implemented with large datasets. The algorithm has been put to use for purposes of classification and regression. A decision tree is sequentially formed by segmenting the dataset into smaller groups based on the values of successive features of the data. The end result is a hierarchy of features with each node representing where in the decision process a specific features affects the final process outcome. The branches departing a node represents the split of the dataset based on the outcome of the test.

The use of DT algorithms is not constrained by the data type as it can be used with categorical and continuous variables. The methodology of how to split the data is dependent on the associated probability of a set outcome. For regression decision trees, sum squared error for the training samples is used to select the order of the predictor variables within the tree. For classification trees, the Gini function is used to determine the best choice of splits. It is a measure of difference between values of a frequency distribution.

Implementation considerations when using this type of algorithm are the size of the final tree (number of nodes) and the level of accuracy expected from the algorithm. If too few nodes are included in the model, the accuracy of the model decreases. If too many nodes are included, the model's accuracy is higher, but the researchers run the risk

overfitting the dataset. Two possible methods for coping with the risk of overfitting are either artificially limiting the number of levels within the tree or pruning the tree once the algorithm has been run. While there are many alternative methods for how to best determine the split attribute, three highly effective methods include the greedy, gain ratio, or the distance-based measure. Finally, another option is to use the gain ratio which considers how broadly and uniformly the features splits the data. One drawback of using this algorithm is that DTs can have problems with high variance or increased bias.

c. Support Vector Machines

Support Vector Machines (SVM) are supervised learning models that examine data for classification or regression. SVM adds a dimension to the dataset as a way to make classes linearly separable (for linear applications). Simply put, a class is a subset of data identified by a common feature, or input variable. Given a classification context, SVM inserts a hyperplane between two classes. The selection of the best hyperplane to divide the dataset into classes is the difficult part of this method. In theory, the hyperplane is able to separate the two classes without error and the greatest margin. However, an error often occurs when a member of one class appears on the same side of the hyperplane as the second class. The margin is the distance between the closest point of each class and the hyperplane. When the data is non-linear, the SVM uses a kernel to convert a low dimensional feature space to a higher dimensional feature space, transforming the data, and enabling separation of classes by a hyperplane. In addition, SVM is known for being robust for outliers.

d. *k-Nearest Neighbors Algorithm*

The *k*-nearest neighbors (*k*-NN) algorithm is a non-parametric method used for classification and regression. The value or class of a point of interest is approximated "locally" within a defined feature space. The feature space is comprised of a set number of training points, denoted by *k,* closest to the point being examined that will be used to classify the test point. Small values of *k* can create many small regions which could lead to non-smoot decision boundaries or overfitting of the data. However, large values of *k* will leave larger regions and possible under-fitting of the data. For use in classification, the *k*-NN input is the *k* training examples that are closest to the point of interest and the output is a class membership. The class of the point of interest is determined using a similarity measure. The similarity measure for continuous variables is the distance between the test point and the point of interest. The similarity measure for categorical variables is the Hamming distance. If there is a mix of variable types, one solution is to use standardized distance on the same training set. For use in regression, the output value is the average of the selected feature of the *k* closest points. One method to validate the choice of *k* is to use cross-validation. A variant of the *k*-NN algorithm is the use of weights to weigh the values of the surround *k* points based on the distance from the selected point.

e. Random Forest Algorithm

The random forest (RF) algorithm is a form of ensemble learning. It is used to rank the importance of variables for either regression or classification problems. Initially developed by Breiman and Cultler (2007), the algorithm incorporates the results of several runs of the DT algorithm, each constructed with a unique subset of the initial

dataset, into a single result in order to decrease variance, decrease bias, or improve predictive power. The RF algorithm produces a final tree which combines the most common nodes. For the construction of trees using the DT algorithm, the dataset is sampled with replacement. When choosing the attribute from which to create the node, only a small, random subset of the available attributes is considered. Each tree within the forest is restricted to a subset of characteristics, thus reducing the dimensionality of the problem. This method is considered an improvement over decision trees because it reduces the tendency for overfitting that is often seen with decision trees. RF is considered robust to inclusion of irrelevant features and it is known to be capable of classifying a large quantity of data accurately.

Researchers typically find a balance between the performance, processing time and memory. The number of trees is related to the number of variables. The greater the number of variables, the greater the number of trees that may be developed for the dataset. However, it has been noted that increasing the number of trees does not necessarily improve performance. Research by Amit and German (1997) illustrated that the accuracy of RF algorithms is dependent on the individuals trees and the dependence between the trees. One advantage of using RF is that the method is capable of maintaining accuracy even with missing data. Common variants to this method include kernel RF, centered RF, and uniform RF.

f.Artificial Neural Network Algorithms

Artificial neural network (ANN) algorithms can be constructed for both supervised and unsupervised learning. The concept behind the algorithm was to create a

learning process modeled after the human brain. For example, if an artificial neural network is used to classify patients as diabetic using electronic health records, it develops its own set of relevant characteristics from iterations with the training dataset. This algorithm is best defined as a combination of optimization theory and statistical optimization. It seeks to find the best model from the set of models that minimizes the cost to traverse the network.

2.1.2.2 Unsupervised Learning Algorithms

Unsupervised learning assumes that there is a hidden structure within the data. As opposed to supervised learning in which input data is paired with supervisory data, unsupervised learning depends only on the input data. The goal of unsupervised learning is to learn more about the dataset. It is primarily used for clustering and association efforts. The method of clustering looks to discover groupings within the data. The method of association attempts to determine a rule (or rules) that can be used to describe a large portion of the data.

a. *K-means*

The *k*-means clustering (*k*-means) algorithm, or Lloyd's algorithm, divides the data space into *k* cells. To initialize the algorithm, *k* initial "means" are chosen. Once initialized, *k*-means consists of two iterative steps. In the assignment step, each point is assigned the cell that has the least squared Euclidean distance between the point and the mean. Once all the points have been assigned, the new means, called centroids, are updated. The observations are then sorted again and placed in the cell with the closest mean. The algorithm stops when observations are no longer being assigned to new cells.

b. *Apriori Algorithm*

The apriori algorithm is an example of association rule learning. It is used to find frequent "item sets." Since it was initially developed for datasets that contain a large number of transactions, it has been used within the field of healthcare for the detection of adverse drug reactions by creating association rules for the combinations of drugs on a specific subpopulation of patients (Harpaz et al., 2010).

2.1.2.3 Reinforcement Learning Algorithms

Reinforcement learning makes use of a small labeled dataset which includes supervisory data, and a larger unlabeled dataset with only input variables. This type of learning is based on the concept that the use of unlabeled data after training with labeled data can still provide incremental improvement in the results. This learning method is particularly valuable when the cost of labeling datasets makes labeling a full dataset prohibitive. This would be true in the development of algorithms to help review images for significant features or automatic image processing. Given a large number of images available, it would be costly to have a subject matter expert on the images review all available images. Instead, it is more likely that the subject matter expert would review a smaller sample of images examples to provide appropriate labels.

This section included summaries of common algorithms used for supervised, unsupervised, and reinforcement ML. As the field is still developing, the total number will continue to grow through both the addition of completely new algorithms and

development of variants of existing methods. Table 2.2, below, summarizes both the methods, available algorithms, and the purposes for which they are typically utilized. If an algorithm has been used for more than one type of learning, it was entered in the section for which its use is most prevalent.

Table 2.1. Common Machine Learning Algorithms

Machine Learning	**Supervised Learning**		**Unsupervised Learning**		**Reinforcement Learning**	
Learning Purpose	Classification	Regression	Clustering	Association	Prediction	Control
Type	▪ Logistic Regression (LR) ▪ Support Vector Machines (SVM) ▪ Discriminant Analysis (DA) ▪ Naïve Bayes (NB) ▪ Nearest Neighbors (kNN) ▪ Decision Trees (DT)	▪ Linear Regression (LR) ▪ Non-linear Regression (NLR) ▪ Ensemble Learning ▪ Neural Networks (NN) ▪ Nonlinear regression ▪ Decision Trees (DT)	▪ Hidden Markov Models (HMN) ▪ Neural Networks (NN) ▪Gaussian Mixture ▪ Principal Component Analysis (PCA) ▪ Single Value Decom -position (SVD) ▪ Hierarchical ▪ Self-organizing maps	▪ DCA ▪ Single Value Decomposition (SVD) ▪ K-Means ▪ Apriori	▪ Temporal Difference ▪ Tabular Temporal Difference	▪ Criterion of Optimality ▪ Brute Force ▪ Value Function ▪Direct Policy Search

2.1.3 Open Software for Machine Learning

There is a wide variety of open source software platforms readily available to assist researchers harness the power of machine learning algorithms. Open source software is accessible to the general public for use as is or maybe adapted for greater performance or a different application. Typically developed in a group, the software is free which makes the software attractive to individuals or teams without extensive financial assistance.

With the growth of open source ML platforms readily available and the decision on which software to utilize becomes harder. Considerations should include the availability of a specific algorithm within a library or framework and the researchers comfort with coding in general or a specific program. Table 2.2 contains a brief list of popular ML libraries and their associated platforms or frameworks. A library contains a set of objects for a particular use. A framework, on the other hand, is a collection of libraries designed to support a methodology. Software is defined as containing support programs, the existence of a code library, and reliance on a scripting language.

Table 2.2. Available open Source Machine Learning Software

ML Library	Language	Description	Creator	Established
Accord.NET	C#	.NET ML framework with audio and image processing libraries (.NET, 2018)	Cesar Souza	2009
Amazon ML		Guided platform built on proven, scalable ML technology which serves the parent company (free only with AWS) (AWS 2018)	Amazon	2015
Apache Mahout	Java, Scala	Project which produces free implementations of ML algorithms for filtering, clustering, and classification (Mahout, 2017)	Apache Software System	2014
Apache Singa	C++, Python, Java	Flexible architecture for distributed training (Apache Incubator, n.d.)	DB Systems Group	2015
H20	Java, Python, R	Open-source ML Platform focused on enterprise service (H2O.ai, 2018)	H2O.ai	2011
Oryx 2	Java, Scala, Apache Hadoop	Real-time large scale ML; packages for filtering, classification, regression, and clustering (Onyx 2, n.d.)		2014
Scikit-learn, TensorFlow, Theno*	Python	3 “most popular” ML libraries for use within python (Raschka, 2015)	Various	2016
Caret, randomForest, rpart*	R	Open source platform for statistical programming and applied ML (R Core Team, 2012)	University of Auckland	1993
Shogun	C++, Java, Python, C#, Ruby, R, Lua, Octave, Matlab	Open source ML library with range of ML methods (Shogun, n.d.)	Soeren Sonnenburg and Gunnar Raetsch	1999

2.2 Overview of Survey Methodology

This paper is the synthesis of a formal, systematic literature review of published research concerning machine learning and diabetes mellitus within the past several years. The basis for the research was a protocol developed to explore the breadth and depth of current research on T2DM using the techniques of machine learning. The protocol is explained in detail in the following sub-sections.

2.2.1 Research Focus

In order to understand the focus and research completed on T2DM using machine learning techniques, this work investigates the following research questions.

Question 1: What are the major T2DM research areas that are being actively pursued by the scientific community?

Question 2: What ML techniques are commonly being implemented?

Question 3: How is the effectiveness of machine learning research being assessed?

In the course of the literature review, it was noted that one of the greatest limitations on the research involved the data used as a foundation for the work. Therefore one additional research question was added which focused on the suitability of current data and the limitations on researchers and results due to datasets.

Question 4: What limitations exist that hamper the productivity of the current research efforts?

2.2.2 Repository Search Strategy

The purpose of this research effort was to explore the current research for the T2DM using ML technology. The search was limited to articles as part of the PubMed database. This online archive contains 28 million citations for biomedical literature from a variety of sources to include MEDLNE, journals, and books (National Center for Biotechnology Information, n.d.). The strategy used to identify search terms for an automated search within PubMed consisted of identifying major search terms from the research questions, identifying alternative spellings and synonyms for major terms, and then determining the best search phase to use within the selected database.

During a preliminary literature investigation, the author noted that T2DM is annotated with a variety of alternative phrases in published literature to include type II diabetes, type II diabetes mellitus, type 2 diabetes, type 2 diabetes mellitus, T2, and DM type 2. The inconsistent use of a reference term had the potential to remove relevant articles for a key search. The final search string selected for used in PubMed for this literature review was:

> “machine learning” AND (“diabetes” AND (“type II” OR “type 2” OR “T2DM” OR “T2” OR “T2D” OR “DM2”))

This search string resulted with 76 citations for review. Figure 2.1 provides an overview of the research methodology utilized for this paper. An initial search utilizing the aforementioned string was conducted to create the initial literature repository. Inclusion criteria were applied against abstracts of papers included in the initial

repository to form the base repository. The full papers were then examined with exclusion criteria to form the refined repository. The refined repository contained all relevant articles to the literature review.

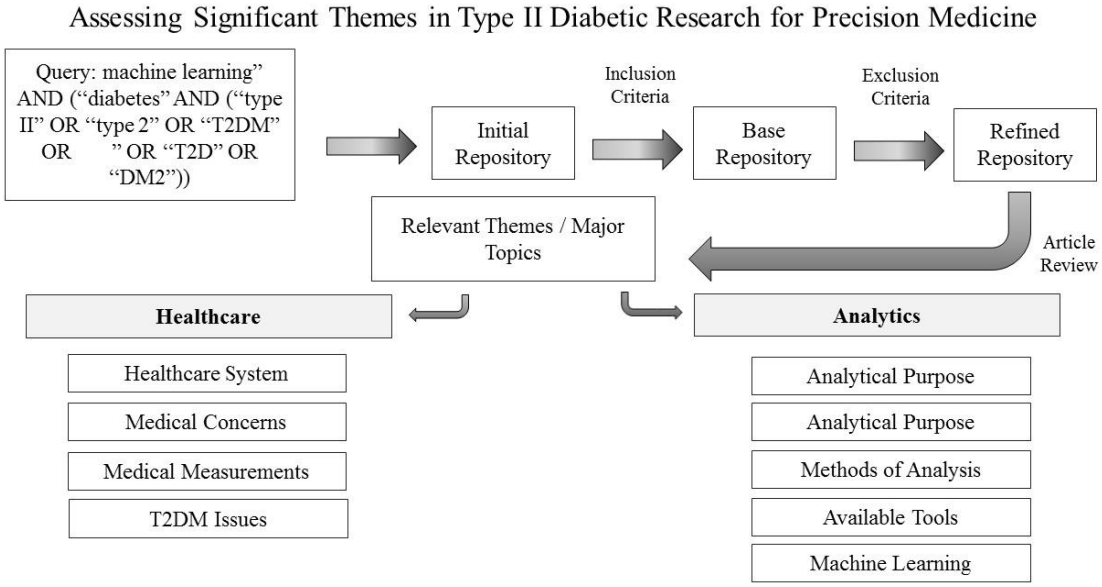


Figure 2.1. Research Methodology for Literature Review

2.2.2.1 Search Documentation

The title, author(s), journal title, year published, and title for the documents identified in PubMed using the chosen search string were stored in an Excel table designed as an initial repository for the remainder of the literature review. The list of articles was then evaluated using the inclusion criteria detailed in Section 2.2.2.2 against the information included in the published abstract. All papers that met this criteria were downloaded and the full papers were then reviewed using the exclusion criteria. Those papers were then examined for key pieces of information chosen for analysis using Table 2.2.

Table 2.3. Literature Review Checklist

Article Authors	All authors listed on the publication
Year	Year published
APA Reference	APA reference
Journal Title	Journal Title
Article Title	Article Title
Key Issue	Concern or gap that prompted research
Research Purpose	Goal of research paper
Model Dependent Variable	Dependent variable for analysis
Type of Learning	Supervised, unsupervised, semi-supervised
Machine Learning Purpose	Various
Machine Learning Algorithm(s)	Various
Dataset Methodology	Traditional or Train/Validation
Traditional Data Allocation	Percentage for training/percentage for validation or N/A
Cross Validation	Various
Model Performance Metrics	Various
ML Software	Software used for ML
Country of Dataset Origin	Various
Dataset Name	Various
Data Time Frame	Start Year - End Year
Dataset Time (Years)	Number of Years
Data Issues	Researcher recognized issues with data collection or format.
Instances	Number of subjects or cases included within the dataset.
Instances Used in Model Development	Number of subjects/cases used in analysis
Instances (Dependent Variable)	Number of subjects/cases used in analysis with T2DM
# of Features	Number of variables/features in the dataset
# of Features used in Model Development	Number of variables used in the modeling portion of the analysis
# of Features in Model	Number of variables included in the final model
Types of features	Types of variables/features included in the dataset
Study Limitations	Study limitations acknowledged by the author
Benefits of Analysis	Stated benefit of the analysis
Recommendations for Future Work	Recommendations for future work by the author
Sources of Funding	Various

2.2.2.2 Inclusion Criteria

When reviewing abstracts, the author focused on the following considerations:

Inclusion Criteria 1: Publications that describe research in which ML techniques are used to investigate a research question focused on the prevention, diagnosis, or management of T2DM.

Inclusion Criteria 2: Documents published after 2012.

2.2.2.3 Exclusion Criteria

When reviewing published papers, the author focused on the following considerations:

Exclusion Criteria 1: Identified article was published as a conference proceedings.

Exclusion Criteria 2: Identified article was a literature review.

Exclusion Criteria 3: Identified article was qualitative.

Exclusion Criteria 4: Publications/reports for which only an abstract is available.

Exclusion Criteria 5: Research that was not primarily focused on T2DM.

Exclusion Criteria 6: Dataset must include data from human subjects

Exclusion Criteria 7: Research methods did not include ML.

Exclusion Criteria 8: Method of medical diagnosis is not recognized by the American Medical Association

2.3 Literature Survey Outcomes

2.3.1 Summary of Articles Included

This literature review limited the scope of the search of articles included within PubMed through March 22nd, 2018. Of the 76 documents included in the initial repository, only a total of 39 articles were included in the refined repository on which the rest of this paper is based. 36 articles were excluded from the refined repository, and 14 articles were rejected for failure to meet the established inclusion criteria. Of these articles, 11 were published prior to 2013. Another 22 articles did not meet the exclusion criteria. The most prevalent reasons for exclusion of an article from the literature review was that the research did not focus of T2DM.

Table 2.4. Impact Table of Inclusion and Exclusion Criteria

Criteria	Explanation	Articles Impacted
Inclusion 1	Publications that describe research in which ML techniques are used to investigate a research question focused on the prevention, diagnosis, or management of T2DM.	3
Inclusion 2	Documents published after 2012.	11
Exclusion 1	Identified article was published as a conference proceeding.	3
Exclusion 2	Identified article was a literature review	0
Exclusion 3	Identified article was purely qualitative.	1
Exclusion 4	Publications/reports for which only an abstract was available online.	0
Exclusion 5	Research not primarily focused on T2DM.	13
Exclusion 6	Dataset must include data from human subjects.	3
Exclusion 7	Research methods did not include ML.	1
Exclusion 8	Method of medical diagnosis not recognized by the American Medical Association.	1

Figure 2.2 shows the growth within the research conducted over the given year period. In 2013, only four articles were published, but by 2017, 16 articles were published for the T2DM research using machine learning techniques.

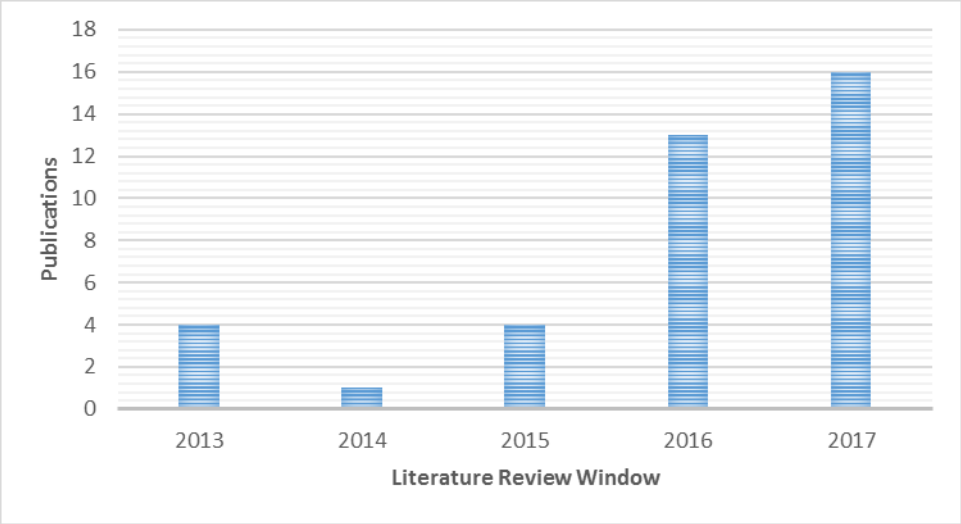


Figure 2.2. Annual Comparison of Published Articles on Type 2 Diabetes Mellitus and Machine Learning

Of 39 published articles that met this review's inclusion and exclusion criteria were published in 31 different journals. Table 2.4 lists all of the journals in which the papers were published. The journals with more than one paper published on the topic included Big Data, IEEE Journal of Biomedical and Health Informatics, Journal of Diabetes Science and Technology, and Medical Care. In looking at the authors, the 29 papers had over 244 contributors. On average six authors were listed as contributors in each paper. What was surprising in looking at the authors was that only eight authors had published research papers on the intersection of T2DM and ML. Of those the highest number of published contributions within that window was three. Only two authors had

successive publications which were published in the same journal that was IEEE Journal of Biomedical and Health Informatics.

Table 2.5. Summary of Journals with Publications on Type 2 Diabetes Mellitus and Machine Learning Between 2012 and 2018

Journal	Number of Articles
Artificial Intelligence in Medicine	1
Big Data	2
BMC Nephrology	1
BMJ Open	1
Briefings in Bioinformatics	1
Computational Biology and Chemistry	1
Diabetes	1
Diabetes Care	1
Diabetes, Obesity and Metabolism	1
Diabetologia	1
Diabetology & Metabolic Syndrome	1
Health Informatics Journal	1
Health Informatics Research	1
IEEE Journal of Biomedical and Health Informatics	3
IEEE Transactions on Biomedical Engineering	1
Information Sciences	1
International Journal of Biostatistics	1
International Journal of Medical Informatics	1
Journal of American Medical Informatics Association	1
Journal of Biomedical Informatics	1
Journal of Clinical Epidemiology	1
Journal of Diabetes Science and Technology	5
Journal of Translational Medicine	1
Medical Care	2
Medical Decision Making: An International Journal of the Society for Medical Decision Making	1
Medical Physics	1
NPJ Genomic Medicine	1
Plos One	1
Sao Paulo Medical Journal	1
The Lancet: Diabetes & Endocrinology	1
Translational Psychiatry	1
Total	**39**

2.3.2 Major Topics Covered

A key component of this research effort was determining areas of interest for medical research and narrowing down the most frequently used analytic approaches to tackle those issues. Before delving into the articles themselves, a simple keyword analysis was performed. While eight articles chose not to list key words, the other 31 articles cited 112 key terms. After removing repeated and similar terms, the words were placed into one of two primary categories consisting of healthcare or analytics. Once placed into the categories, the words were again sorted by general topic areas. For healthcare, the important topics, outlined in Table 2.5, covered medical measurements, medical concerns, healthcare systems, and T2DM related issues. Determining the correct features in which to measure physical health and disease progression is a challenging and complex issue. First and foremost, choosing the best measurements comes from understanding the disease at the heart of the study. The medical measurement terms included within this bin were either general terms (phenotype and genotype) which indicated whether the research was focused on physical or genetic characteristics or was a unique or non-traditional indicator of T2DM.

Table 2.6. Healthcare words cited in published literature focused on Type 2 Diabetes Mellitus and Machine Learning

Medical Measurements	Medical Concerns	Healthcare	T2DM Related Issues
anthropometric measurements	disease progression	healthcare	diabetes mellitus type 2
anthropometry	early disease prediction	high throughput	diabetic kidney disease
arterial markers	high throughput	primary care	diabetic retinopathy
biomarkers	medication adherence	privacy	disease complex
body mass index	missing heritability		disease progression
continuous glucose monitoring	noisy labels		glycemic variability
fats	noninvasive treatment		healthcare
genotype	patient centered medicine		hypoglycemia prediction
glomerular filtration rate	patient similarity		impaired glucose tolerance
glycemic control	population screening		kidney failure
glycemic variability	privacy		metabolic syndrome
hypertriglyceridemic waist phenotype	risk assessment		microvascular complications
optical coherence tomography	risk classification		pre-diabetic state
phenotyping	risk predictions		renal insufficiency
photoplethysmorgaphy	screening		
prognostic tool			
protein			
protein-protein interaction			
serum creatinine			
SNPs			
triglycerides			

Table 2.7. Analytic key words cited in published literature focused on Type 2 Diabetes Mellitus and Machine Learning

Analytical Purpose	Available Tools	Machine Learning	Methods of Analysis	Analytical Concerns
risk assessment	big data analytics	artificial neural networks	big data analytics	positive predictive value
evidence based medicine	cohort study	boosting	comparative effectiveness research	selection bias
medical informatics	comparative effectiveness research	classification	contextual anomaly detection	sensitivity
predictive analytics	data mining	classification and regression tree	data mining	size constraints
predictive models	database research	ensemble learning	interaction network	time dependent confounding
predictive models	electronic health records	FDSP	inverse probability weighting	
predictor	feature engineering	feature learning	joint image-region-map model	
propensity score	machine learning	Gini importance	Kallikrein-Kinin system	
risk classification	medical informatics	predictive models	marginal structural model	
risk predictions	modeling	random forest	Markov-Gibbs random field	
	regulatory feature data	random forest feature contribution method	metric learning	
		Semi-supervised clustering	mixture of generalized linear effects	
		super learning	multivariate model	
		supervised decision techniques	non-negativity-constrained autoendcoder	
		supervised machine learning	prediction	
		support vector machine	predictive models	
		survival tree	signal processing	
			statistical learning	

2.3.1.3 Insights on Current Type 2 Diabetes Mellitus Research

The articles in this literature review focused on finding solutions on gaps between today's medical capabilities and identified medical needs in the diagnosis and treatment of T2DM. As part of the introduction for published articles, the authors outlined the purpose of the research. Each cited a critical gap, explained implications of leaving this aspect of medical care at the status quo, and then proceeded to explain their research methodology. For the most part, the issues addressed by the researchers are not unique to T2DM, but if the gaps could be bridged for patients with T2DM the payoff would be more impactful due to the prevalence of the disease. While the solutions proposed by the researchers are important, the documentation of the gaps themselves provides valuable insight into the research areas that are seen significant enough to attract funding support, areas and for which current technological capabilities may potentially be able to solve. As a majority of the research was funded, the gaps identified were seen as significant to wider audiences than just the authors. The gaps generally fell into three categories.

The first gap category included papers that looked at potential applications or uses for new medical knowledge. For example, Acciaroli et al. (2018) focused their research on how to best use glycemic variability indices to classify patients. An individual's glucose levels fluctuate throughout the day based off food consumption and exercise. The glucose level increases after meals and decreases after cardiovascular events. Glycemic variability is the measure of change in glucose swings. Researchers discovered that some patients have greater glycemic variability than other patients. Another region of growth is the advancement of molecular technology. As aptly summarized by Leung et al. (2013) a

major challenge is synthesizing new knowledge so that it can impact clinical practice. The increased ability to measure patient characteristics at the molecular level provides more detailed information, but increased knowledge has not fully been translated into better medical practice. Continued research needs to be done to link more refined measurements to more precise care.

The next bin includes research that attempts to address how best to decrease the time that a provider must spend on a single patient to identify the ailment and recommend treatment. Doctors must analyze a patient's medical history and current laboratory results to narrow down possible ailments and associated treatment plans. Researchers are examining the medical diagnosis process to determine if they are possible efficiencies within the system. Efficiencies exist where technology can replace the human in the loop in assessing routine data collected to look for abnormalities. While the doctor's opinion will remain central to the decision of the final treatment plan, automatic review of portions of a patient's record potentially improves the speed of the decision making, or call a doctor's attention to a critical component of the record. One area of growing interest is the application of machine learning technology as part of the medical image review process. Available medical imaging includes x-ray, computed tomography, magnetic resonance imaging, and ultrasound imaging. To assess a patient's current physical health, a medical professional will need to review each image to look for signs of disease. If the initial image review is automated, doctors would only need to review files for which the images indicated risk of a particular disease. The change to the process would reduce the time an individual needs to spend reviewing multiple images.

EITanboly et al. (2017) focused research efforts to develop a computer-aided diagnostic system for optical coherence tomography images. By detecting retinal changes in T2DM diagnosed patients earlier in the development of the disease, the improved prediction model would provide patients additional time and awareness to make appropriate decisions that could either delay or prevent the onset of diabetic retinopathy, a complication associated with hyperglycemia. One common concern with articles that addressed automating medical tasks was ensuring that the process would perform better than a human's performance on the same task. Automated processes that do not perform as effectively medical professionals could negatively affect patients and should not be implemented.

Another gap between current medical capabilities and medical needs is the ability to assess a risk for individuals with regard to developing a disease because once having the disease an additional risk of developing specific complications associated with the disease. Current prediction methods do not capture all patients who will develop T2DM. However, given the advancements of analytics and medical measurements, there exists the possibility of refining prediction models to better identify individuals at risk. Also, technological advancements have broadened the available characteristics for medical professionals to better assess patient health by increasing the number of characteristics measured. Measurements that may be used as indicators are broken down into two distinct groups: genotypes and phenotypes. Genotypes are an individual's heritable genetic identity. Phenotypes are observable characteristics that include physical appearance, development, and behavior. Researchers are working to link specific

genotypes and phenotypes to individual diseases. Allalou et al. (2016) noted that 20-50% of patients with gestational diabetes progress into T2DM within 5 years. The characteristics that make gestational diabetic patients more likely to become T2DM are not well documented. By finding the characteristics common to patients who progress from gestational diabetes to T2DM, clinics would be able to better treat mothers during and immediately following pregnancy. Li et al. (2016) noted that using anthropometric measurements, human body measurements, to predict T2DM remains controversial. Their work was unique in that it offered to reexamine the use of non-invasive measurements as indicators. The benefit of noninvasive measurements is that they are less expensive to obtain and could potentially be used to screen portions of the populations that are not routinely able to access healthcare.

Anderson et al. (2016) explored how to best use existing EHR data to gain new knowledge about the progression of T2DM. Li et al. (2016) attempted to use limited EHR data to evaluate patient risk for T2DM in order to better protect patient privacy. Farran et al. (2013) noted that the medical community needed an effective way to stratify patients by classifying potential risks of developing complications over time. Advances in the development of better methods for assessing the patient risk allow for the medical system to focus prevention and treatment efforts on patients with the greatest level of risk. It also creates the potential for patients to be aware earlier of their risk so that if properly motivated they can make alterations to their behaviors to lower their risks.

2.3.2 Insights on Medical Data for Research

The medical datasets used as a basis for the research came from a variety of sources. Each dataset had its own limitations due to the structure and completeness of the available information. The following section provides a review of the data sources used, the types of variables used within the research, and the associated limitations. Analysis of existing data resources provides a foundation which can be leveraged to recommend changes to how data is current collected and stored within the medical system.

2.3.2.1 Data Sources

The articles in this literature review utilized data from four primary sources: electronic health records, national health studies, completed research efforts, or data collected for the purposes of the specific research paper. The first three sources of information provide data at potentially reduced cost to researchers, but research utilizing this type of data may be limited in that researchers may have to adjust the scope of their research effort to conform to the data at hand.

Ethical and security concerns have informed the construction of a detailed approval process for the use of medical data. A critical detail in the construction and use of medical data is ensuring that the data is used in a manner consistent with the way the data was initially collected or approved secondary uses. Most medical data is collected with the intent to assess a patient's health and inform treatment decisions. The data collected for this type of use, also termed as the original use, may contain a patient's family history, laboratory test results, records of procedures, financial information, prescribed medications, and clinical notes on issues such as treatment compliance. Any

other use of medical data obtained in this manner is termed as the secondary use. This term reflects the use of any data for non-clinical applications. There is considerable debate about the extent to which an individual's medical information can and should be used for research purposes. In order to ensure that medical data will be used within the secondary use guidelines of the controlling institution and in a manner that protects patient information, the majority of data used for medical research is considered restricted access. Researchers will need to obtain approval for their research plans through both their institution's institutional review board (IRB) and receive an additional IRB approval from the organization that owns the data repository.

Assuming that the researcher chooses to use an existing dataset there are a variety of different data repositories available through both public and private agencies. Choice of repository depends on the research question. In comparison to using electronic health records, the use of data from a study or trial provides concentrated data about a specific type of patient. Studies allow for the aggregation of interesting cases. Casanova et al. (2006) chose to use the Jackson Heart Study from the University of Mississippi Medical Center to uncover potential predictors of T2DM in African Americans. Hertroijos et al. (2017) elected to use data repositories from two Dutch Diabetes Care networks to develop glycemic trajectories for patients recently diagnosed with T2DM. The inclusion of a second, similar repository allowed the team to validate the model developed using the initial repository. By using data from a specific study, the researchers can focus their efforts on highlighting unique aspects related to a subpopulation. In some instances,

studies allow researchers to study the effectiveness of treatment protocol since the patients included within the study receive standardized care.

In comparison, the electronic health record (EHR) system aggregates individual patient treatment data for a larger population. The medical repository may be defined by medical facilities within a geographic location or, more likely, by agreements between large health providers. The establishment of an EHR system has created a more standardized, readily accessible repository of medical information on a large sample of patients. Researchers have recognized that use of data compiled as part of an EHR could support clinical research by providing longitudinal treatment data for a robust patient population. These records have the potential to allow researchers to evaluate treatment outcomes and develop screening criteria for known health risks. The transition from paper to electronic records made the aggregation of medical data easier. Each EHR is unique to the controlling medical system and may include important information like a patient's medical history, diagnoses, current medications, immunizations, and laboratory reports. Lack of standardization of what medical data must be included within an EHR would limit aggregation at the national and international level and causes challenges for how to externally validate developed models with region EHR data. In addition, researchers have recognized that use of data compiled as part of an EHR could support clinical research by providing longitudinal treatment data for a robust patient population. These records have the potential to allow researchers to evaluate treatment outcomes and develop screening criteria for known health risks.

Researchers continue to explore existing medical databases to find new novel indicators for unique sub-populations and develop greater knowledge of non-invasive indicators. The growth of electronic data bases, better methods for the measurement of known characteristics, increased knowledge of genetics, and the increased availability of data analysis tools capable of handing the scale of medical data have contributed improvement in this area.

2.3.2.2 Medical Dataset Variable Categories

The availability of the data used in the majority of the research efforts was dependent on the data collected and stored in an existing data repository. The majority of the variables within the datasets fell into nine general categories. The following summarized types provides a general overview of the types of data that may be included for that category for a given study.

a. *Demographic.* This type of variable includes information such as a patient's age, ethnicity, gender, time to diagnosis, and source of medical insurance. Most of the variables included within this type are nominal data types such a gender or ethnicity.
b. *Clinical laboratory tests.* These variables includes outcomes for tests performed on samples of blood, urine, or other tissues take from the patient. The laboratory tests are performed with the intent to diagnose a disease or to monitor the patient for changes in their health conditions.
c. *Prescribed medications.* Patients under the care of a medical provider for a health condition may be prescribed drugs to control physical reactions. For example a patient may be diagnosed medication to control blood pressure. Information

contained within this section should then include the type, frequency, and quantity of each prescribed drug.

d. *Vital signs.* As part of an initial screening when meeting with a medical professional at a care clinic, a patient's vital signs are recorded. Variables annotated within medical repositories may include a patient's height, weight, blood pressure, or resting heart rate. These variables provide the medical staff important information about a patient's health particularly when observations are conducted over a period of time.

e. *Genomic Data.* Identifiable variations in an individual's genes can provide medical professionals with important information that can impact medical care decisions. While still not common practice to gather genotype data for all patients, this collection and analysis of this type of data is growing.

f. *Anthropometric measurements.* This type of measurement is used to assess the size and shape of the human body. Common measurements may include waist circumference, waist-to-hip ratio (WHR), waist-to-height ratio (WHtR), and body mass index (BMI).

g. *Diagnosis codes.* The U.S. healthcare system uses the International Classification of Disease (ICD) codes to annotate a patient's disease and heath conditions when the patient is seen by a medical provider (National Center for Health Statistics, 2018). The ICD system as a whole is a comprehensive tally of health conditions that standardizes entries in patient records. First adopted in 1893, the ICD allowed researchers to assess health trends across time and space. The codes are published in two different manuals for separate purposes. ICD-10-CM is used for outpatient

coding and contains over 68,000 codes. While ICD-10-PCS contains the procedural classification system for use in an inpatient setting. Finally ICD-10-PCS contains over 87,000 codes. T2DM had 86 distinct codes. Table 2.8 contains a sampling of the codes to illustrate the details regarding a patient that can be gained from the ICD code.

Table 2.8. Classification Codes for Diabetes Mellitus (National Center for Health Statistics & Centers for Medicare and Medicaid Services, 2018)

ICD Code	Description
E11621	Type 2 diabetes mellitus with foot ulcer
E11630	Type 2 diabetes mellitus with periodontal disease
E11641	Type 2 diabetes mellitus with hypoglycemia with coma
E11649	Type 2 diabetes mellitus with hypoglycemia without coma
E1165	Type 2 diabetes mellitus with hyperglycemia
E1169	Type 2 diabetes mellitus with other specified complication
E118	Type 2 diabetes mellitus with unspecified complications
E119	Type 2 diabetes mellitus without complications
E1110	Type 2 diabetes mellitus with ketoacidosis without coma
E1111	Type 2 diabetes mellitus with ketoacidosis with coma
E1122	Type 2 diabetes mellitus with diabetic chronic kidney disease
E1129	Type 2 diabetes mellitus with other diabetic kidney complication
E11311	Type 2 diabetes mellitus with unspecified diabetic retinopathy with macular edema

h. *Specialized tests.* Depending on the medical issue at hand, medical providers have additional medical tests which may provide information about a particular aspect of a patient's health. Variables pertinent to this category include information obtained from tests like an echocardiogram, a sonogram of the heart, or electromyography, a test to analyze nerve and tissue electrical activity.

i. *Medical Images.* Pictures of certain structures within the body may provide additional information about a disease. Common medical images that are used for diagnosis are x-rays, computed tomography scams, magnetic resonance images ultrasounds, and nuclear medicine imaging. Those images are reviewed by medical specialists trained to look for abnormalities.

2.3.2.3 Insights on Training and Validation

Prior to using the dataset, the researchers must develop a data plan to describe how the available data will be used to train, validate, and test the model developed using machine learning. There are two commonly used methods for how to use available data to train and validate model performance. Traditionally, the available dataset is separated into three sections and training dataset is used to develop the model. The data is then provided to the machine learning algorithm as a set of examples from which parameters are identified. The validation dataset is used to adjust the parameters used for a classifier. This dataset can also be used for feature selection. The test set is held in reserve to judge the performance of the model. However, this method only works when there is a large volume of data. The second method, the development of model parameters through the use of cross validation, may allow the researcher meet both the test and validation requirements without losing modeling or testing capability. In general the data set will be divided into small, equal sections. The data sections will then be allocated into training and validation test sets. These tests will be used to develop the model. Once completed, the sections will be re-allocated to form a new training and validation test sets. The process will continue until either all possible ways to divide the original sample into

training and validation sets is complete (exhaustive cross validation) or until a pre-determined number of runs is complete (non-exhaustive cross validation).

Within this literature review the preponderance of the researchers chose to use k-fold validation by dividing the training dataset into *k* subsections. The model is then built using *k*-1 sections. Once the model is complete the model is then tested using the remaining section and the researcher annotates the model performance. The process is then repeated until each section has been held in reserve as the validation set. The overall model performance is the average of the model performance when tested using the section held in reserve.

2.3.2.4 Insights on Limitations Due to Dataset/Structure

Within each article, the researchers annotated the limitations of their work due stemming from the available data. The primarily limitations were due to the size of the available data set, frequency of missing values, class imbalance, large dimensionality, and limited variable types. All of the papers selected to be part of this literature review cited missing entries as a limitation in performing the analysis. Researchers had three options for how to deal with missing data. First, the researcher could choose to remove all cases for which there was missing data. Farran et al. (2013) only used patients with complete data for the variables within the model. From over 270,000 patients, termed "hospital visitors," the number considered in the model was 10,632. This option, although valid, removed a significant portion of the available data from the analysis. Han et al. (2017) also chose to delete vacant data. As a result, the sample size shrunk from 9,562 to 7,913. Secondly, the researcher could choose to approximate the missing data based on

other data entries. Lastly, the researcher could choose to impute the data. Before opting for this method, the researcher must determine why the data is missing. Is the missing data related to other available information about the subject or is the missing data dependent on the value? One possible reason for missing data is that the patient chose to censor particular information such as a family history of a disease. The second most prevalent issue that researchers needed to address was how to deal with class imbalance. This particular issue occurs when there is a large difference between the size of groups with or without a feature. For example, there may be 200 entries of patients that screened for T2DM, but only 5 percent of the patients showed large glycemic variability. Datasets that link output variables to input variables are particularly valuable in the development of risk models.

2.3.3 Recent Machine Learning Algorithms Used for Type 2 Diabetes Mellitus

ML is a method which automates model building based on the idea that systems can learn from data to identify patters and make informed recommendations. In this context ML is used to improve performance in T2DM prevention, diagnosis, and management. Choice of the best type of ML to use was dependent on both the available data and the research questions. In most cases, the research questions attempted to answer how best to predict health risk for patients. The dataset chosen to support the analysis included the known classification of the patient which made the data amenable to use supervised learning. The research questions that used genetic information as possible predictor variables applied unsupervised methods to determine which features were most important.

The preponderance of the research used supervised ML for the purposes of classification and prediction. Acciaroli et al. (2018) used supervised ML to build prediction models with the capability to distinguish between three classes of individuates: healthy people, patients with impaired glucose tolerance, and patients with T2DM. The researchers chose to use logistic regression build with 5 fold cross validation since the size of the dataset was too small for the traditional division of the dataset into a training, validation, and test set. Contrarily, Allalou et al. (2016) used ML to develop a metabolomics signature for the prediction of patient progression from gestational diabetes to T2DM. Using feature selection, their team was selected the top 22 variables out of 182 known variables to develop an accurate prediction model.

Anderson et al. (2015) used unsupervised ML to explore relationships within the dataset. The end result of their work was the development of a prediction model for the progression of pre-diabetes into T2DM using variables found within EHRs. Finally, Argwal et al. (2016) applied a reinforcement ML technique in an effort to examine an alternative method to manual labeling to create training sets. The research team correctly surmised that labeling a dataset for use with machine learning was prohibitive due to both cost and availability of medical specialist to review the requisite number of files.

2.3.2 Software Utilized by Research Teams

Researchers working at the intersection of T2DM most often use open software. The two packages used the most often are R and the Waikato Environment for Knowledge Analysis (WEKA). Developed by Bell Laboratories, R is a free software package that is capable of working on a variety of platforms to include Windows and

MacOS. The frequent use of R can be attributed to its ability to handle large datasets, its graphical capabilities, and the thought that went into the development of its programming language. Additionally, ready-made functions for ML algorithms to include neural networks, deep learning, recursive partitioning, random forests, regularized and shrinkage methods, and support vector machines are available for download. Table 2.9 includes a small sample of over 150 available R ML packages. Even though the table is not comprehensive, it hints at the breadth of what exists.

Table 2.9. Examples of existing ML packages for R (Hothorn, 2018)

R Package	Purpose	Authors
randomForest	Classification and regression based on a forest of trees using random inputs	Leo Breiman, Adele Cutler, Andy Liaw, Matthew Wiener
rpart	predictive models by indirect classification and bagging for classification	Andrea Peters, Torseten Hothorn, Brian D. Ripley, Terry Therneau, Beth Atkinson
tree	classification and regression trees	Brian Ripley
nnet	software for feed-forward neural networks with a single hidden layer and for multinomial log-linear models	Brian Ripley, William Venables
ROCR	flexible tool for creating cutoff-parameterized 2D performance curves	Tobias Sing, Oliver Sander, Niko Beerenwinkel, Thomas Lengauer
caret	training and plotting classification and regression models	Max Kuhn
svmpath	computes the regularization path for the two-class SVM classifier	Trevor Hastie
kernLAB	Kernel-based ML for classification, regression, clustering, novelty detection, quantile regression, and dimensionality reduction	Alexandros Karatzoglou, Alex Smola, Kurt Hornik
glmpath	a path-following algorithm for L1 regularized generalized linear models and Cox proportional hazards model	Mee Young Park, Trevor Hasite
CoxBoost	routines for fitting Cox models by likelihood based boosting for a single endpoint or in the presence of competing risks	Harold Binder
BayesTree	implementation of the Bayesian Additive Regression Tree	Hugh Chipman, Rober McCulloch

Like R, WEKA is a free licensed software that is used for data analysis. Written in Java, the program has a variety of tools for data pre-processing, classification, regression, clustering, association rules, and visualization (WEKA, n.d.).

2.4 Further Research Insights and Common Concerns

At the end of their papers, research teams highlighted their concerns regarding the findings and the potential applications of the work. One prevalent concern expressed was the need to limit negative consequences to the patient when employing machine learning findings. The researchers were particularly concerned when the final result of the analysis was intended to replace a human within the analysis portion of diagnosis and treatment decisions. There was a common understanding that the final model needed to perform at least as well as medical professionals before the model should be used in clinical practice. Another concern expressed by researchers was whether the work would have meaningful impact. Some researchers considered whether the available prediction model would provide meaningful warnings to patients with enough time for the patient to change their behavior to avoid undesirable consequences.

The results of the research included within this literature review aptly illustrated the positive impact that the integration of supervised ML into medical research can have on the identification of important variables for the purposes of classification and prediction models for T2DM applications. Continued advancements using ML will depend on fostering a small community of researchers that consistently explore the use of ML for T2DM application, the development of improved medical datasets to support the

research, and the cost to assess the predictive value of an indicator against the cost to take the measurement and store the data.

Of the 266 authors who contributed to the literature included in this study, only a small subset had published more than one paper involving ML for T2DM applications. This may indicate that there needs to be greater support, both financially and intellectually, to encourage more analysts to pursue research in this area. The creation of a community of practice for the application of ML for DM. The group will be able to facilitate changes to medical database construction to support future research efforts. Furthermore, the community of practice will be able to initiate discussions with the medical community to solicit input from subject matter experts on proposed research questions and methodologies. The engagement will also provide researchers to share their findings and potential influence changes to the medical system.

To produce the most benefit, datasets needs to be complete and comparable. A large number of missing data entries within large data repositories can create variation in the final predictive models. If cases with missing data are deleted, the available data for analysis is significantly smaller than the initial data set. Placing values on improving completeness of electronic health records will improve the resultant models. In addition, the limited availability of research databases has restricted researchers from verifying their findings or those researchers within a reasonable timeframe and also limits their ability to compare the results of single analytic method with more than one dataset. This could be especially important in the ability to compare the results of research utilizing electronic health records.

Analysis conducted in isolation may result in recommendations that cannot be applied in practice due to practical considerations. Future work that seeks to find best variables to use for prediction of T2DM should consider selecting multiple sets of variables for various scenarios. When comparing prediction models that rely on the availability of select pieces of data to make a prediction, it is imperative that researchers need to consider the portion of patients which are likely to have the variables in their datasets and the cost of those tests.

CHAPTER THREE

DEVELOPMENT OF PERFORMANCE LOSS FUNCTIONS FOR HEALTHCARE APPLICATIONS

3.1 Introduction

Relatively recent advances in both medical knowledge and increased technological capability to measure changes within the human body have made the role of a healthcare provider increasingly difficult. Doctors are expected to digest excessive amounts of data and, from that data, develop actionable recommendations in a timely fashion. This global "expectation" of healthcare providers creates a demand for increased involvement of other specialties within the scientific community in the development of better methodologies for transforming data into information and then using the resultant information to develop optimal treatment plans. The Precision Medicine Initiative (PMI) outlines a goal of tailoring medical care for the individual patient. For this effort to be successful, it is incumbent upon research teams to think critically about the problem at hand, determine what aspect of patient treatment their field could provide assistance in improving, and start an open dialogue with the medical community. This paper seeks to establish an informative exchange as to how quality engineering methodologies can be applied to treatment protocol selection by examining how to adapt the quality loss function for use within the healthcare domain. In support of a larger effort to develop improved metrics for health assessments and patient's physical performance this article develops the concept of reference interval-based performance functions.

3.1.1 Research Motivation and Scope

The rise of the financial cost of healthcare in the United States has intensified the desire to find efficiencies within the medical system to lower overall costs and, at the same time, improve the quality of medical care for patients. The phrase *healthcare costs* encompasses all funding related to the management of the complex healthcare system as well as the costs stemming from the lack of productivity of the ailing portion of the population. Non-optimal treatment plans, ineffective treatment, or lack of care can lead to longer patient recovery time, multiple cycles of treatment, loss of life, or the inability of the patient to recover to their pre-ailment physical condition.

This chapter examines how the adaptation of an existing analytical methodology within quality engineering could facilitate the establishment of better decision tools for healthcare providers. Improved decision tools that combine laboratory results, available treatment options for the ailment, and associated treatment risks would assist medical professionals in the selection of the best treatment option for the patient.

The advances made during past century in medical knowledge and practice have made a medical professional's role increasingly difficult. Doctors are expected to digest vast amounts of data and, from that data, develop actionable recommendations in a timely fashion. Based on an assessment of the number of medical articles published in the 20 major clinical journals in 1992, researchers estimated that doctors would need to read 17 articles a day to keep up with advances in medicine (Davidof et. al, 1995). Since the publication of this research effort, the number of medical articles published per month

has continued to increase. The continued growth had made staying abreast of the most recent advancements even more difficult for practitioners. For this reason, it is important to develop improved decision tools for medical professionals to harness the available information. The involvement of other disciplines will serve to innovate current processes and, if successful, improve the quality of patient care through decreased diagnosis time and improved treatment efficacy. The importance of involving other parts of the scientific community is best highlighted by the establishment of PMI. The initiative's research platform outlines the vision of tailoring medical care to the individual (Ashley, 2015). The multidisciplinary research effort leverages medical databases to find better techniques for both diagnosis and treatment informed by individual patient characteristics. For PMI to be successful multidisciplinary research, teams must think critically about the problems at hand, determine what aspect of the problem their respective academic fields can assist in solving, and then start an open and productive dialogue with other teams working on the same problem and practitioners. This paper seeks to establish a productive discussion centered the application of quality engineering methodologies to treatment selection within healthcare. Existing analytical tools refined within the field of quality engineering for manufacturing applications have potential to improve healthcare system efficiency. In particular, the adaptation of the quality loss function (*QLF*) for use within the medical field would provide a means of mapping a measured physical characteristic to either physical performance loss or an increased risk of future health complications. The technique could allow providers to assess the value of a treatment protocol on a patient's overall health prior to selecting the best treatment.

Before applying the manufacturing quality methodologies to the healthcare field, researchers will need to adjust developed techniques to account for differences in the initial problem construction between a manufacturing application and a medical application. This paper provides a brief review of *QLFs* and their development for use within the manufacturing sector. After outlining the unique aspects of medical quality characteristics, the author proposes a new methodology for assessing performance of an individual based on measurable physical characteristics. Lastly, insights for the potential use of performance functions for both univariate and bivariate healthcare assessments are discussed.

3.1.2. Manufacturing Loss Functions

In the manufacturing sector, quality engineers are often tasked to develop and monitor a process whose output needs to adhere to a pre-identified specification value for a select measured quality characteristic with minimum variance. The desired value is a single number commonly referred to as the "target" or "target value." The examined item is only usable by the customer if its measured characteristic meets the pre-identified requirement. If an item's measurement exceeds the specification limits, the product must be scrapped or reworked. If the item does not meet the specification requirement, the manufacturer incurs a financial loss related to that item since the entity is not able to sell the item. Suppose that a company manufactures nails to be used for wood frame construction of residential homes. If the nail manufacturing process is flawless, the measurable characteristics of each nail produced on the assembly line are the predefined targets. In this case, the manufacturer incurs no financial penalty related to production

defects. In reality, production lines are not perfect and sources of variance exist within the process. Few nails produced in the factory are exactly the target length of 1.45 inches long. The consumer base, however, does not need nails that are exactly 1.45 inches long. As long as the nails are within 0.05 inches of the target's desired length, they can still be used safely for frame construction. Any nails that do not fall within the $\pm$0.05 tolerance window must be scrapped. Quality engineers work diligently to ensure that the manufacturing process produces nails with the smallest ratio of defective items to usable items. Their job, in essence, is to reduce the loss to the manufacturer. The quality engineer mathematically relates the cost of defective items to process performance using the target value and process variance. This methodology will be described in greater detail in Section 2. The relationship between the measured characteristic and the cost enables the manufacturer to identify strategic points within the production process for investment to reduce process variability and improve desired target achievement.

Like the manufacturing sector, the healthcare industry strives to improve the effectiveness of treatments through both target acquisition and variability reduction. Variability in patient response to treatments has the potential to incur additional expense on the part of the patient, the treating medical organization, society at large, or a combination thereof. The purpose of this paper is to examine how *QLF*s could be used within the healthcare community to relate an analyte's measurement and associated healthcare costs. An analyte is a substance that is analyzed by finding the measurements of its chemical subcomponents (Merriam-Webster, 2017). By relating the measured value

of physiological characteristics with costs patients and medical professionals will be better able to assess the value of a treatment protocol.

3.2 Quality Loss Functions

3.2.1 Development and Application in Manufacturing

Loss functions are a mapping of an event to an associated cost. Traditionally loss has been defined from the viewpoint of a manufacturer and, as such, occurred when the item produced was unable to provide value to the manufacturer without additional investment. Using the example discussed in the introduction, let us consider again the case of a manufacturer who produces nails for wood frame construction. The target value for the nail's length is 1.45 inches. Natural variability in the assembly process affects production and the end result is that no two nails produced are exactly the same length. As long as the nail's measured length is within a specified tolerance of the target, in this case $\pm$ 0.05 inches, the nail can still be used safely for frame assembly by the customer

and the nail maintains value to manufacturer. If a nail exceeds the predetermined tolerance limit the item would be scrapped and the value of the nail to the manufacturer would decrease. A nail that measures less than 1.39 inches or more than 1.51 inches could not be used by a consumer for the intended purpose. A traditional step loss function relates the nail length to the manufacturer's cost, *L(y)*. Other practical scenarios for loss functions may include the additional possibility of reworking an item in addition to that of scrapping the item when the tolerance is exceeded.

3.2.2 Traditional Step Loss Functions in Manufacturing

Traditional step loss functions (Taguchi *et al.*, 2004) vary by the type of quality characteristic chosen: nominal-the-best type (*n-type),* smaller-the-better type (*s-type*), and larger-the-better type (*l-type*). The three types of characteristics are explained in more detail below.

3.2.2.1 Step Loss Functions: Nominal-the Best Type Quality Characteristic

For nominal-the-best type (*n-type*) quality characteristics the manufacturer accepts all items whose measurement of a chosen characteristic lies within the pre-determined upper and lower specification limits. The specification limits are defined by the allowable tolerance, Δ, from a desired target value, τ. Loss is incurred by the manufacturer only when a tolerance limit is exceeded. The step loss function for an *n-type* quality characteristic can be written as:

$$L(y) = \begin{cases} 0 & \tau - \Delta \leq y \leq \tau + \Delta \\ A & otherwise \end{cases}$$

L(y) is the loss associated with *y*, the measured value of the desired quality characteristic. The cost for scrapping or reworking the product is *A*, a constant cost. The lower specification limit, *LSL*, is the lowest value that a characteristic can be without the item being scrapped or reworked. The upper specification limit, *USL*, is the largest value that a quality characteristic can be without the product having to be scrapped or reworked. Figure 3.1 illustrates a step loss function for an *n-type* quality characteristic.

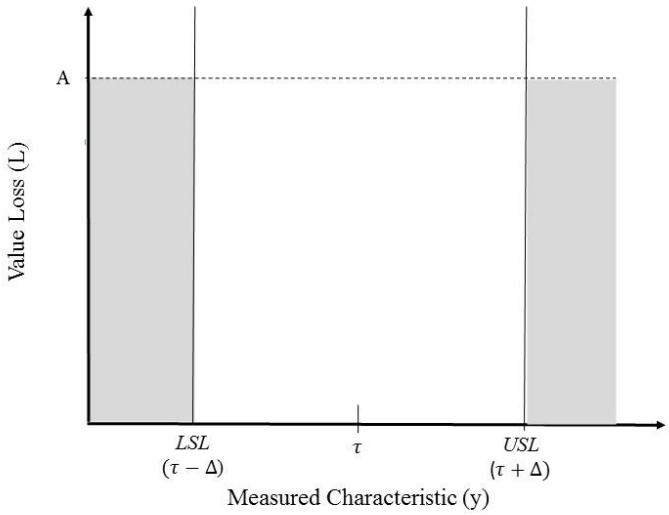


Figure 3.1 Traditional Step Loss Function: Nominal-the-Best Type Quality Characteristic

The light grey area illustrates the additional cost that a manufacturer incurs from scrapping or reworking a product. As can be seen, the loss value is uniform when the measurement exceeds the specification limits. As long as the measurement of the characteristic is within the tolerance window the product's value to the manufacturer is not affected by either the absolute distance between the measurement and the desired target or by the absolute distance between the measurement and the closest specification limit.

3.2.2.2 Step Loss Function: Smaller-the-Better Quality Characteristic

For smaller-the-better type (*s-type*) characteristics, the company strives to manufacture items whose measured characteristic is as small as possible. Traditionally the target value for an *s-type* characteristic is set to zero. A practical example of an *s-type* characteristic could be the sound created by a lawnmower. In this case, the manufacturer desires the noise volume for engine to be as small as possible. If the sound emitted by the

machine exceeds the hearing safety threshold or, more likely, exceeds the threshold for which consumers are willing to purchase, the product will need to be scrapped or reworked before being sold. The traditional step loss function for an *s-type* quality characteristic is expressed mathematically as:

$$L(y) = \begin{cases} A & y > \tau + \Delta \\ 0 & otherwise \end{cases}$$

The loss to the manufacturer, denoted by *A*, is incurred by the manufacturer only when the measurement of the characteristic exceeds the tolerance window. Figure 3.2 illustrates a traditional step loss function for an *s-type* quality characteristic.

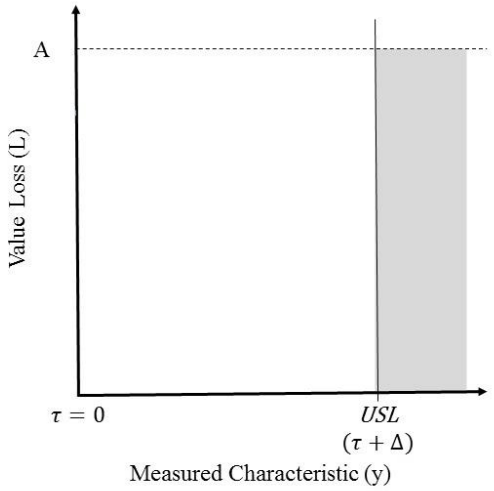


Figure 3.2 Traditional Step Loss Function: Smaller-the-Better Type Quality Characteristic.

The light grey area illustrates the additional cost that a manufacturer incurs from scrapping or reworking a product. As can be seen, the loss value is uniform when the measurement exceeds the upper specification limit. As long as the measurement is less

than the *USL* the product's value to the manufacturer is not affected by the absolute distance between the measurement and the desired target or by the absolute distance between the measurement and the upper specification limit.

3.2.2.3. Step Loss Function: Larger-the-Better Type Quality Characteristic

For the case of a larger-the-better type (*l-type*) quality characteristic, the manufacturer desires to produce products with the largest possible measurement for the chosen quality characteristic. Practical examples of an *l-type* characteristic could include a product's useful lifespan or the amount of resistance an exercise band could endure before snapping. The traditional loss step function for an *l-type* quality characteristic may be expressed mathematically as:

$$L(y) = \begin{cases} A & y < LSL \\ 0 & otherwise \end{cases}$$

As seen in Figure 3.3, the manufacturer incurs a loss when the measured characteristic is less than the established specification limit.

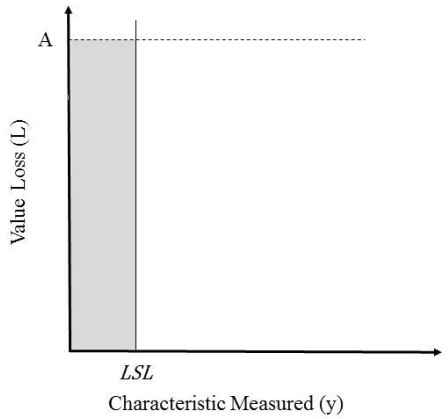


Figure 3.3. Traditional Step Loss Function: Larger-the-Better Type Quality Characteristic

There are two notable shortcomings to the traditional step loss function. First, the manufacturer cannot use a known value of the loss to determine the measured characteristic's value. The uniform formulation results in an uncountable number of different measurements which could result in the same loss. The manufacturer cannot use loss to determine how the process should to be tweaked to obtain better results in the future. Secondly, the value of the loss, *L(y)*, only takes into account manufacturer's financial loss. The loss to the customer is ignored by the traditional step function. When a product fails to match expectation of the customer, the value of the product decreases in the eyes of the customer. A company advertises a product with specific characteristics and the customer chooses to purchase the product based on the advertised characteristics. The difference between reality and expectation could create frustration in the consumer. While the manufacturer does not incur an immediate loss when the item does not meet the target but remains within the specification limits, the deviation from the target value has the potential for future losses due to recalls, returns or warranties.

3.2.3 Development and Description of Quality Loss Functions

Taguchi (Taguchi *et al.,* 2004) believed that any deviation of a measured quality characteristic from a desired target value results in loss. He articulated a more comprehensive value of the loss by adding the customer's perceived loss (the difference between the expected value and observed value) to the manufacturer's loss. Deviation from a characteristic's target value, even if the measurement remains within the accepted tolerance window, can result in a loss of goodwill from the customer due to the variance in quality. Taguichi's quality loss function (*QLF)* incorporates both the viewpoint of the

manufacturer and the customer. The loss function increases in value as the amount of deviation from the target increases. Inclusion of the customer's loss provides recognition that not all products produced within specification limits are equal. In his work, Taguchi looked at the best way to calculate the value of loss for three different types of quality characteristics: nominal-the-best (*n-type*), smaller-the-better (*s-type*), and larger-the-better (*l-type*). This methodology is explained in more detail below.

3.2.3.1 Quality Loss Function: Nominal-the-Best Type Quality Characteristic

For *n-type* of quality characteristics, the closer the measurement of the quality characteristic to the desired target the smaller the summative loss to the consumer and manufacturer. Taguchi opted to use the quadratic function to represent the approximate loss between the specification limits. A function to estimate the loss is necessary since true relationship between measured value and the resultant loss is unknown when the measurement lies within the tolerance window. The mathematical representation of the loss function for an *n-type* characteristic can be written as:

$$L(y) = \begin{cases} k(y-\tau)^2 & \tau - \Delta \leq y \leq \tau + \Delta \\ A & otherwise \end{cases}$$

The approximate loss is calculated by multiplying the loss coefficient, also known as the proportionality constant and denoted by *k*, by the square of the difference between the measurement of the characteristic and the desired target value. If the value of the characteristic exceeds either the lower or upper specification limit, the loss occurred is a constant value of *A*. As can be seen in Figure 4, the *QLF* accounts for the loss between the specification limits (customer's loss) in addition to the loss that occurs when the specification limits are exceeded (manufacturer's loss). Whether the deviation is to the

left or to the right of the target value does not impact the loss calculation because of symmetry. The value of *k* is chosen by the analyst to relate the measured characteristic's numerical value to the loss incurred at that measurement by the customer. As such, the value is unique for each problem. As stated earlier, the true relationship between the measured value and the customer's value of loss is unknown. The curve merely estimates how the loss changes within the tolerance window. For manufacturing applications with assumed symmetric loss above and below the desired target, the constant *k* relates the loss associated with the specification limit with the distance of the limit from the target value. The mathematical form of a loss coefficient for a symmetric *n-type* characteristic can be expressed as:

$$k = \frac{c}{d^2}$$

The constant *c* is the loss associated at a specification limit and *d* is the absolute distance between the specification limit and the desired target value of the characteristic's measurement. If the loss incurred is not the same at the upper and lower specification limits or the target value is not in the center of the tolerance window, there should be different values of *k* calculated for the customer's estimated loss if the measured value is below the target and if the measured value is above the target.

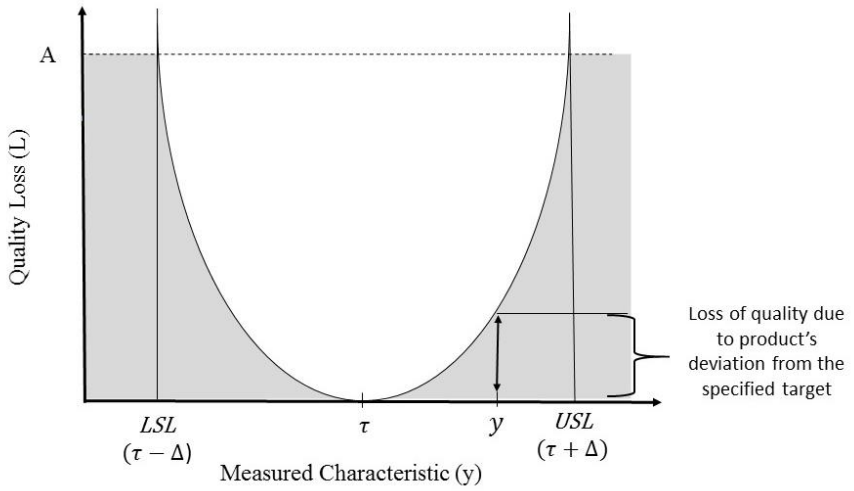


Figure 3.4. Quality Loss Function: Nominal-the-Best Type Characteristic

3.2.3.2 Quality Loss Function: Smaller-the-Better Type Characteristic

For the *s-type* quality characteristic, it is desirable that the measured value of the characteristic be as small as possible. Traditionally the target value for an *s-type* characteristic is set to zero. The quality loss function imposes a penalty for any deviation of the measured reading above zero (the readings cannot be negative). Once the deviation surpasses the upper specification limit the loss reaches a maximum value of *A*, the loss to the manufacturer. The mathematical formulation for the quality loss function for an *s-type* quality characteristic can be as:

$$L(y) = \begin{cases} A & y > t + \Delta \\ ky^2 & otherwise \end{cases}$$

The loss between the target (zero) and the specification limit is calculated by multiplying the loss coefficient, *k*, by the square of the measured value of the selected quality characteristic. As shown in Figure 3.5, the loss increases as the difference between the measured quality characteristic and the desired target grows.

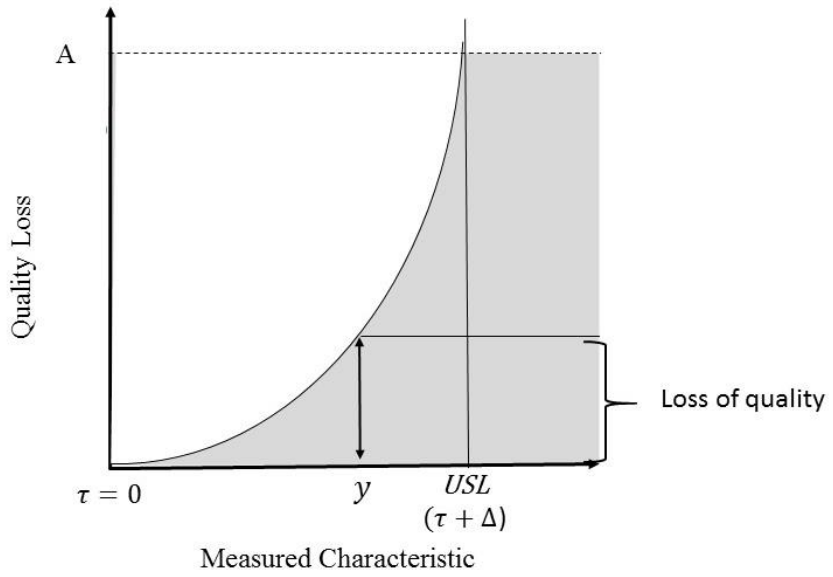


Figure 3.5. Quality Loss Function: Smaller-the-Better Type Quality Characteristic.

Practical examples of characteristics which could be considered *s-type* include noise levels, the weight of an item, and the breaking distance of a car. Manufactures desire to produce products in which identified *s-type* characteristics as small as possible. The measurement must, however, always be positive.

3.2.3.3 Quality Loss Function: Larger the Better Type Quality Characteristic

For the case of an *l-type* quality characteristic, the manufacture's goal is to maximize the measurement of the quality characteristic, *y*. Like the traditional step loss function, the manufacturer's loss is denoted by the constant *A* and is incurred by the manufacturer when the measurement of quality characteristic, *y*, is less than the specification limit. The quality loss function takes into account the loss to the consumer when the quality characteristic deviates from the desired target, but is still within the acceptable tolerance window. The *QLF* imposes a penalty for all values of *y* that are less

than the target, but more than the specification limit. The quality loss function for an *l-type* characteristic is defined as:

$$L(y) = \begin{cases} A & y < \tau - \Delta \\ \dfrac{k}{y^2} & \tau - \Delta < y < \tau \end{cases}$$

The loss estimate is calculated by multiplying the loss coefficient, *k*, by the reciprocal of the square of the value of the measured characteristic. Practical examples of *l-type* quality characteristics in manufacturing include a product's useful life, reliability of component parts, or the strength of a component material. As can be seen in Figure 3.6, the customer's loss grows as the distance between the measured characteristic and the target grows.

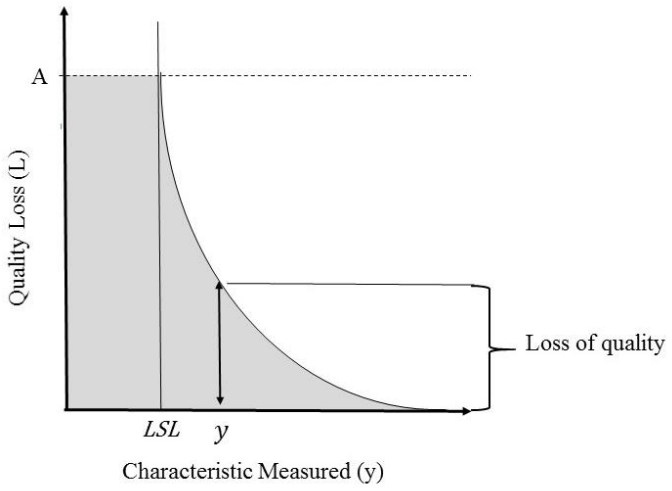


Figure 3.6. Quality Loss Function: Larger-the-Better Type Quality Characteristic

The quality loss function proposed by Taguchi provides two distinct advantages over the traditional step loss function. First, in using the traditional step function the analyst is forced to accept the assumption that all products whose measured characteristic is within the tolerance window have the same value to the customer. Only the loss of the manufacturer is taken into account and the voice of the customer is lost. The quality loss

function allows for the inclusion of the voice of the customer. In addition, the traditional step loss function does not provide gradated differences in loss related to the measured value. The loss of value to the customer due to the measurement's deviance from the target when the measurement with within the tolerance window is not included in the overall loss estimate. In comparison, Taguchi's methodology allows for the calculation of the total loss which incorporates both the loss to the manufacturer and the loss to the customer.

Figure 3.7 provides a visual means for which to compare the differences of the two methodologies. The picture shows measurements (*a*-*d*) for a single, measured characteristic of four products (*A*-*D*), the target value for the characteristic (τ), and the upper and lower specification limits ($\tau - \Delta$, $\Delta + \tau$). As can be seen in the illustration, three of the measurements span a large portion of the tolerance window and one measurements sits to the right of the upper tolerance. If using a traditional step loss function, the loss values for products *A*, *B*, and *C* would all be zero. Product *D* would be the only product with a valued loss of *A*, the manufacturer's loss. Only using the quality loss function can differences between the products be articulated to the manufacturer using the estimated loss value. Given the differences in measurements, one should expect that the customer would perceive a difference in product performance when using product *A* versus product *C*. The figure shows that there is relatively the same amount of distance between the measurements of *a* and *b*, and the measurements of *c* and *d*. Common sense would dictate that the loss incurred by product would be closer for two products whose measurements are relatively close. Using the traditional step function to

value loss, the loss value for the product *C* would be closer to loss value for the products *A* and *B* than product *D*.

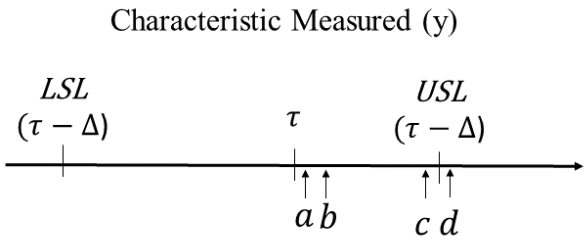


Figure 3.7. Illustrative Diagram: Measurements Relative to Specification Limits.

Since the true value of the loss to a customer is not measurable, Taguchi developed a reasonable methodology to estimate loss utilizing the quadratic function. The quality loss function, while not perfect, was an improvement upon the traditional step method for loss valuation.

3.3 Motivation to Alter Manufacturing Loss Functions for Future Healthcare Application

3.3.1 Uniqueness of Healthcare Characteristics

Like manufacturing, the healthcare industry relies on the use of physical measurements to assess a product. In order to determine the proper diagnosis and treatment plan for a patient, medical professionals use qualitative and quantitative data to form and validate hypotheses about the patient's health. Qualitative information is collected as medical professionals observe the patient during the initial screening and subsequent meetings. Based on the information provided during the screening and knowing possible ailments which could presumably cause the identified reaction or ailment, the medical professional may then order select laboratory tests to help confirm

the suspected cause. Once the sample is analyzed, the laboratory provides the resulting measurement and the associated tolerance window to the medical professional responsible for the patient. Almost 80% of medical decisions made by doctors are influenced by information from laboratory reports (Katayev, 2010). To assess the measurement's significance, the measurement the medical professional uses the provided the reference interval, decision limit, or the reference change value to assess whether the measured physiological characteristic is considered to be within a normal range. The differences between these measurements will be discussed more in depth in Section 3.2. The assessment about the physical characteristic provides information to the doctor to narrow down possible causes for the ailment.

A key to improving both the speed and quality of diagnosis and effectively choosing an appropriate treatment protocol is the development of a deeper understanding of important measurable characteristics in the human body. The knowledge regarding healthcare quality characteristics has been limited by four distinct challenges. First, medical knowledge is still a growing field. As an example, deoxyribonucleic acid, DNA, profiling was not developed until 1984. Today DNA is used to test for an individual's susceptibility to known hereditary diseases. Within the span of 30 years, the improved test expanded the breadth of the medical community's capability drastically. Medical professionals are still learning about the building blocks the human body and their importance to peak functioning. Secondly, researchers continue to develop and refine techniques to measure specific physical indicators. As the testing process improves the amount of information that a doctor can glean from a patient's sample will continue to

grow. Furthermore, the increased medical data repositories and computational capability provided by modern systems will allow medical research teams to identify physiological differences between subpopulations which can either indicate a smaller tolerance window for a measurable physical characteristic or might indicate the selection of a specific treatment protocol. Normal ranges for known characteristics potentially may be refined so that inter-population variability will no longer mask ailments. Some tolerance windows for specific physical characteristics have been established for each gender. Continued research focused on the identification of important sub-groups and associated reference ranges is needed to narrow current reference ranges. Wide tolerance windows do not help medical professionals efficiently treat patients since they provide little information. Lastly, medical professionals are forced to primarily rely on univariate analysis to support multivariate decisions. Many of the references available, to include reference intervals, are compiled while holding all other factors stable. For example, doctors are not able to provide a reference interval for characteristic *a* given that characteristic *b* is not within the normal range. Additional research needs to be conducted so that medical professionals are better able to understand how to treat patients with multiple ailments using on or more approved drugs. As medical knowledge grows, technology improves, and subpopulations are identified medical professionals will be able to better use the results from laboratory measurements to diagnose and treat patients. Specifically, the ability to understand the significance of a single value in light of a patient's personal history will allow doctors to possible diagnose patients sooner or to treat a patient more efficiently. Over time, continued research on healthcare quality

characteristics will enhance our understanding of the human body and how it reacts when stressed.

In comparison to quality characteristics within the manufacturing sector, there are two distinct facts make healthcare characteristics unique. First, the target for a healthcare characteristic is defined as a range rather than a point value. The optimal level of glucose in the blood stream for a specific individual is not empirically known. Diligent medical research teams have been able to specify a range within which the glucose reading for 95% healthy population would fall. Reference intervals (RIs) for select quality characteristics are established through well-documented studies and will be explained in more detail in Section 3.1.2. RIs are currently the most prevalent metric used to assess a patient's health. Unlike manufacturing where engineers compare a measurement to a point target, doctors compare the laboratory result to a target range. The use of an acceptable range aligns with the concept that each human is unique and the realization that human understanding of medical science is limited. Secondly, a naturally occurring inter-variability between the physical qualities of people and possible intra-variability between successive measurements of a single patient. Differences between successive measurement can be due to error in measurement technique or biological changes (in a single patient) or biologic differences (between patients) (Ricos et al., 2004). In healthcare, the biologic differences have the potential to impact the choice and effectiveness of treatment plans. The following paragraphs provide a brief summary of the different comparison metrics that are used by providers to understand laboratory results.

3.3.2 Healthcare Measurement References

There are three different numeric values which medical professionals may use to assess the significance of a patient's laboratory result. The comparison references are used to determine if the sample's measure is atypical. The three numeric values are discussed in more detail in sections 3.3.2.1 to 3.3.2.3.

3.3.2.1 Reference Intervals

Reference intervals (RI) are the range within which 95% of the values of the population from which the sample was taken are estimated to fall. In the case of medical RIs, the reference group from which samples were taken was limited to selection from a pre-screened healthy sub-population. This type of numeric reference is the most widely used yardstick used to help determine a patient's health. When test results for an individual's laboratory assessment are sent back each measure will be paired with the appropriate reference interval.

RIs can be published both by the manufacturer of the equipment used to perform the test and by independent laboratories that employ the equipment. Before processing analyte samples, it is recommended that laboratories establish a laboratory specific reference interval or verify existing reference intervals from another facility are applicable to the serviced population (Clinical Laboratory Standards Institute, 2008). It is important to note that the RI for a given characteristic may vary between locations. Variance between the published reference intervals for similar reference populations at different laboratories can be attributed to the use of different types of test equipment, the

types of chemicals used in the analysis, or technician technique (American Association for Clinical Chemistry, 2017).

Recent advancements in technology and published methodologies for determining and verifying intervals have improved the quality of the intervals over the past few decades. Often cited and referenced, EP28-A3c: *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory* offers the most comprehensive examination of laboratory protocols which ensure usefulness and reliability of reference intervals (Clinical Laboratory Standards Institute, 2008). The focus on standardization has helped to reduce the variability between RIs from different locations. Differences in reference intervals between laboratories and medical facilities have the potential to induce additional confusion into the decision-making process (Plebani, 2016). The following paragraph will briefly describe the process for establishing an RI. First, the laboratory must fully document its proposed methodology to include criteria for the reference sample population. Next, at least 120 reference individuals from a reference population must be used to form the reference sample group (Clinical Laboratory Standards Institute, 2008). The reference sample group is screened to ensure that they meet the minimum health standards outlined by the documented protocol. In a numerical study using National Health and Nutrition Examination Survey (NHANES) data, Horn et al. (2013) was able to show that inclusion of unhealthy subjects could increase the width of the RI by as much as 30 percent. A wider RI can result in an increased number of individuals who are not appropriately diagnosed. The reference sample group then provides the necessary samples. The RI is found by including the central 95% of the

values found within a sample. The lowest and highest 2.5% of samples are excluded. While separate methodologies for creating reference intervals for parametric and nonparametric data are included in EP28-A3c, the guide recommends the use of the nonparametric methodology for simplicity (Clinical Laboratory Standards Institute, 2008).

Accurate reference ranges are needed by the medical community for patient care and to interpret data from vaccine trials (Kibaya et al., 2008). The accuracy of a reference interval have the potential to impact treatment decisions for patients. As highlighted by Brewster *et al.* (2007), patients can be misdiagnosed when inappropriate reference limits are used. The team analyzed serum creatine kinase (CK) from ethnically diverse sample to validate the applicability of the published reference interval. Their findings indicated that specific ethnic subgroups had naturally higher CK level activity than the general population. If the RI established for the general population was used for diagnosis, the patient's laboratory readings would block the patient from participation in statin therapy. This study illustrated that RIs had the potential to impact the availability of treatment options for patients. In order for a RI to be useful, the sample reference population must be reflective of the population for which the RI will be used. If the patient population is not adequately represented in the sample used to establish the reference interval, the resultant limits can lead to a suboptimal decision on whether to start or continue treatment.

Past studies, like the one by Brewster et al. (2017) bring into question the applicability of common reference intervals to all patients. While it has been

acknowledged that the quality of RIs is better than at any point in history there is still ample room for improvement. Establishing quality RIs is hampered by the availability of adequate sample reference groups and by a limited understanding of which factors impact the levels of a substance within the human body. It has been noted that establishment of RIs for the pediatric population is difficult due to ethical considerations of sampling healthy patients. In addition healthy reference sample group for the geriatric population is difficult due to the high percentage of that population who do not meet the minimum health requirements to provide usable sample (Ceriotti, 2012). A recent National Health and Nutrition Examination Survey (NHANES) study showed only one out of ten subjects in the 70-80 age range could meet the requirements to be part of the reference sample group (Horn & Pesca., 2003). In addition to the complications in gathering data from subpopulations due to ethical concerns and health issues, the lack of understanding of what factors can affect measurements hinders the identification of appropriate subpopulations which could allow for narrowed RIs. As pointed out by Ceriotti only 47 out of the 296 analyte reference intervals provided in the Tietz textbook included a separate RI for each gender (Ceriotti, 2017). This statement suggests that while gender does influence the levels of 47 analytes, not all analyte levels are influenced by gender. In summary, reference intervals provide a range in which a healthy individual could expect the reading to fall. Wide RIs hinders prompt diagnosis by masking abnormal values. The identification of subpopulations with potentially narrower RIs in comparison to the general RI may have a profound impact. Of all the researched limits for medicine, RIs are the most widely documented and researched.

3.3.2.2 Decision Limits

Initially termed “discrimination value” by Sunderman, decision limits are used to mark the difference between the “healthy” and “diseased” population (Sunderman, 1975). While reference intervals focus on describing a physiological state of a healthy person, decision limits were created to help medical professionals determine the risk of disease (Ceriotti, 2008). The two conditions which affect the identification of decision limits are the clinical question for which the lab was ordered and the patient category (Ceriotti, 2008). There are currently three different methods used to establish decision limits: Bayesian, epidemiological, and physiopathological. The Bayesian approach uses knowledge of the diagnostic test, distribution for the analyte in a healthy population, distribution for the analyte in an unhealthy population, and the cost of misdiagnosis to determine an appropriate decision limit for the clinical question. The epidemiological approach is based evidence from population studies. The limits are determined by consensus. The last approach, physiopathological, is based on clinical experience. To date, only eight analytes have universally accepted decision limits. Those analytes are total cholesterol, LDL cholesterol, HDL cholesterol, non-HDL cholesterol, triglycerides, apolipoprotein A-I, apolipoprotein B, and glycated hemoglobin (Ceriotti, 2017).

Standards continued to be refined and are updated as new information becomes available. For example, in 1997, an International Expert Committee recommended changes to criteria used to diagnose diabetes (Kahn, 2003). Specifically, the panel recommended that the fasting plasma glucose level that distinguished between those with diabetes and those without be lowered. The recommended change reflected the

knowledge gained from examination of data that clearly showed diabetic retinopathy, a diabetes complication that affects the eyes and can result in blindness, occurred frequently at a lower reading of fasting plasma glucose. The lower threshold for fasting plasma glucose will directly influence the number of people who are diagnosed with diabetes. The change could potentially result in less people loosing eyesight since they are more aware of the importance of proper control of blood sugar.

3.3.2.3 Reference Change Values

Reference change values are the difference in an individual's analyte measurement over a period of time. A reference value may be used to monitor individuals who have been diagnosed with either an acute or a chronic condition. The utilization of reference change value increases a provider's sensitivity to an individual's pathologic changes in comparison to the use of reference interval. If an individual is taking medication to alter the level of an analyte within his or her body, comparison of a laboratory measurement to a target range value may or may not show that the medication was having the intended affect. Comparison of an individual's successive lab results has the potential to illuminate physical change that might not be apparent because the patient's natural variability falls within the bounds of the population variability. The use of the reference change value is limited to instances when successive readings of the same analyte are taken from one individual.

3.3.3 Need Unmet by the Quality Loss Functions in the field of Healthcare

Quality loss functions are used within the manufacturing application area to determine where limits should be set in order to minimize the loss to society. In

transferring the concept of loss functions to a healthcare setting it is important discuss two important aspects of loss functions: the purpose of the measurement and the meaning of the measured value. Prior to use, there must be agreement on the purpose of the loss valuation within healthcare. The purpose of the loss value for manufacturing serves as motivation to improve the process for long term gains to society as a whole. The quality loss function maps an event or measurement to a cost. When establishing a summative loss function for use within healthcare, it is important to keep in mind that the value will help medical professionals compare the impact of treatment protocol options. The measurement of a select physical characteristic maps to a valuation of health. Unlike manufacturing there only two perspectives need to be included, the use of an adaption of the loss function within healthcare will need to take into account three perspectives: the provider, the patient, and society at large. Loss functions allow the user to clearly see what will happen if a quality characteristic does not meet the target. For example, use of a loss function in healthcare could provide insight into the meaning of a laboratory test does not fall within the allowable bounds. A loss function would allow for health providers quantize a patient's physical performance based on the measured characteristic. Readings outside a specified tolerance window would indicate health implications, the possible need for further tests, or the need for immediate treatment. The second influential decision use of loss functions within healthcare is the meaning of the measured characteristic. When loss functions used within the manufacturing sector, the manufacturer specifies both the target value and the limits for the product. The target value is a point target. The product performs best when the measured characteristic is at

the target value. The limits delineate the lowest and highest measurement within which the product can still be used for its intended purpose. In healthcare, the loss function would need to be adapted for use with a target interval. Medical knowledge has not advanced to the point where the best value of an analyst for an individual is known. At this juncture, there is only general consensus that a healthy value of a measured analyte exists within a specified range. The limits for the loss function could be based either on decision limits or the reference interval. The following two sections propose univariate and bivariate loss functions for use within healthcare. The performance function is an adaptations of the loss functions with a target interval for use within the medical field.

3.4 Proposed Univariate Performance Functions

3.4.1 Applications of Univariate Performance Functions

Performance functions provide a means of mapping a patient's physical performance outcome given a measured physical characteristic. A clinician could reference a performance diagram to make an assessment of how to further investigate, diagnose, or treat the patient. Readings outside the specified tolerance limits would indicate possible health risks.

3.4.2 Healthcare: Nominal-the-Best Type Characteristic

As explained earlier, an *n-type* characteristic for a quality loss function consists of a target value with established upper and lower specification limit. An *n-type* characteristic for a performance function would consist of a target range and with upper and lower specification limits. If the patient's measurement falls within the target range, the patient's measurement would be considered optimal. A patient's value for the

measured characteristic may however fall above or below the specified target range and indicate possible associated health risks. If performance functions are adapted by the medical community, the following three physical characteristics would be designated as *n-type* measurements.

a. *Heart Rate:* Heart rates typically are between the ranges of 60 to 100 beats per minute(bpm) (American Heart Association, 2018). The rate measurement for an individual may vary based on a large number of factors to include gender, fitness, current emotional state, and the individual's position while the reading was being taken (sitting, standing, or lying down). Some issues that can be indicated by heart rate are rhythm disorders include tachycardia (heart rate exceeding 100 bpm), bradycardia (too low heart rate), pre-mature contraction, and Adam-Stokes disease (very fast and steady).

b. *Ferritin:* A protein found in in reticuloendothelial cells, ferritin stores iron and releases it to the rest of the body in a controlled fashion. The amount of iron can be measured by radioimmunoassay. Typical values for adults by gender are 20 to 200 nanograms per milliliter (ng/mL) for men and 20 to 120 ng/mL for women (US National Library of Medicine, 2018). A test result that is less than the established normal range could indicate chronic iron deficiency. A slight increase above the normal level could indicate renal disease. Levels above the normal range could be an indication that the patient has acute or chronic hepatic disease, iron overload, leukemia, or an acute or chronic infection.

c. Vitamin D: This substance plays a role in the control of calcium and phosphate levels within the human body. The amount of Vitamin D is measured through a blood test. The normal range of range is between 20 and 40 ng/mL (Mayo Clinic, 2018). A lower than normal level can be due to the lack of exposure to sunlight, insufficient diet, or liver or kidney diseases. Low levels of Vitamin D have also been linked to greater risk for cardio vascular disease. The use of certain medications such a phenytoin, an anti-epileptic drug, can also result in a lower reading of Vitamin D. If Vitamin D is too high, a patient could experience adverse symptoms which could include nausea or kidney stones.

d. Thyroid Simulating Hormone (TSH): Produced by the pituitary gland, this hormone signals the thyroid to generate and release triiodothyronine (T3) and thyroxine (T4) in to the blood steam. T3 and T4 help to control the body's metabolism. To test for the substance, the patient provides a blood sample. The normal range for TSH in an adult is from .4 to 4.0 milli-international units per liter (Mayo Clinic, 2018). If TSH is below this range, the measurement could mean that the thyroid is not producing enough thyroid hormone and possible hypothyroidism. If the TSH is too high, the measurement could indicate that the thyroid is too active and possible hyperthyroidism.

3.4.2.1 Healthcare: Step Function for Nominal-the-Best Type Characteristic

For an application in healthcare, the traditional step loss function is renamed as a "performance" function. The function provides a numeric estimate of how well the body is performing based on the measured physical characteristic. Performance values close to

zero are considered optimal. A higher performance value indicates decreased performance of the body and associated increased risk of future health complications. Figure 8 graphically illustrates the proposed performance function for an *n-type* characteristic. The horizontal axis is broken down into three distinct zones. Each zone is indicative of the performance that may be achieved with the given characteristic measurement. Zone 1 coincides with the specified target interval and should align with the established reference interval for measured substance. As can be seen, the performance loss within zone 1 is zero. As the measured value deviates outside zone 1, the recommended target interval, the individual experienced decreased performance or increased future health complications.

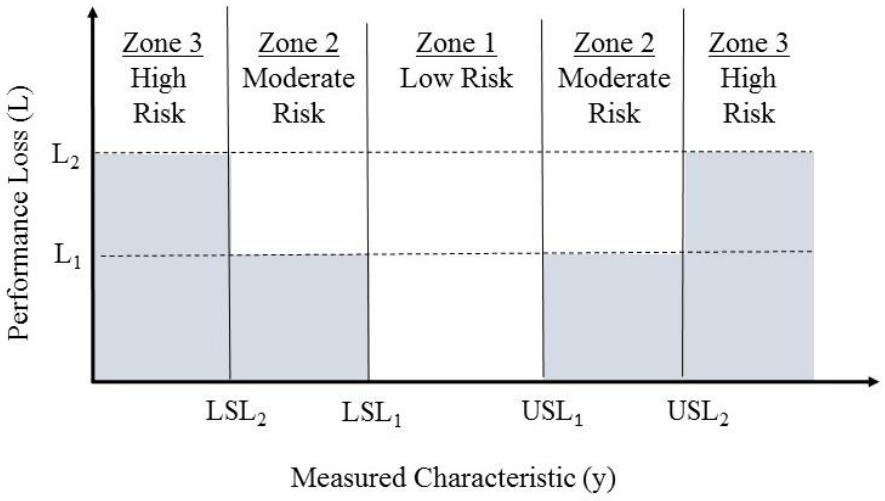


Figure 3.8. Performance Step Function for Nominal-the-Best Type Characteristic.

The performance loss function would be unique to each physical attribute measured. Some health characteristics are known to have a greater impact on an individual's physical performance or are associated with higher risk for long term health complications. The mapping of values outside of the established tolerance window would

result in a larger value of performance loss. Variations of the performance function could include more zones to communicate known risks with set measurements outside the tolerance window. Other physical characteristics might exhibit asymmetric performance degradation. Careful analysis will need to be conducted for each health characteristic to determine the most appropriate number of zones, to verify symmetry of performance, and determine the "magnitude" of the estimated performance loss and increased risk. The mathematical form of the performance step function for an *n-type* characteristic would be:

$$Performance(y) = \begin{cases} 0 & LSL_1 \le y \le USL_1 \\ L_1 & LSL_2 \le y < LSL_1, USL_1 < y \le USL_2 \\ L_2 & y < LSL_2, USL_2 < y \end{cases}$$

Like the traditional step loss function, the performance function does not offer a good estimate of performance degradation in the middle of a zone. The advantage of this method is that it is easy to calculate performance loss and increased risk with a quantitative measurement.

3.4.2.2 Healthcare: Performance Function for Nominal-the-Best Type Characteristic

In order to make the performance function more sensitive to changes in the measured physical characteristic, the step function is replaced with a smooth continuous function. Figure 3.9 provides an illustrative example of a loss function for an *n-type* characteristic with symmetric loss outside of zone 1. Like the step function shown in Figure 3.8, zone 1 is the accepted tolerance window for a normal measurement. As the measurement increases or decreases from the boundary of zone 1, the performance loss grows.

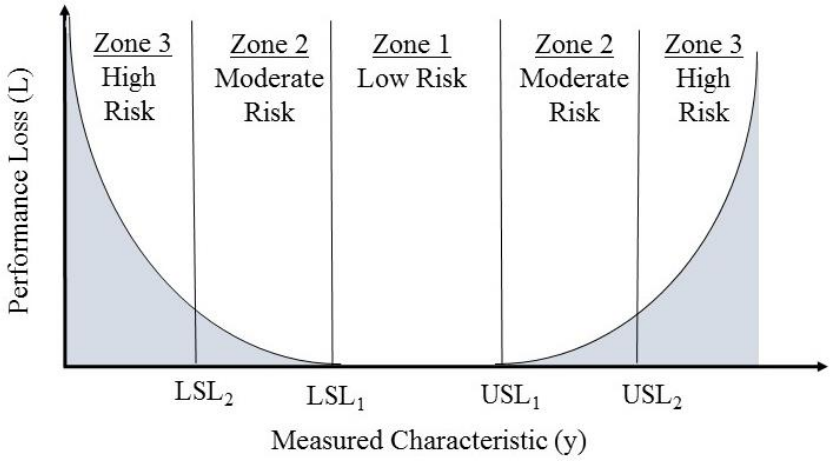


Figure 3.9. Performance Function for a Nominal-the-Best Type Characteristic.

The mathematical form of the performance function for an *n-type* characteristic would be as follows:

$$Performance(y) = \begin{cases} L & y < LSL_2 \\ k(y - LSL_1)^2 & LSL_2 \le y < LSL_1 \\ 0 & LSL_1 \le y \le USL_1 \\ k(y - USL_1)^2 & USL_1 < y \le USL_2 \\ L & USL_2 < y \end{cases}$$

By using the quadratic function to model the decreased performance, the loss would be increased by an amount proportionate to the absolute value of the deviation of the measurement from the specified target. Since performance loss would most likely be different if the patient's test result is less than or more than the target, asymmetric performance loss modeling is a future research area.

3.4.3 Healthcare: Smaller-the-Better Type Quality Characteristic

Next let us consider a performance function for an *s-type* characteristic. *S-type* quality characteristics have an upper specification limit with an ideal target at zero. Below are three examples of in which *s-type* quality characteristics can be found in the health domain.

a. *Blood Glucose Levels*: Doctors monitor the average level of blood glucose over a window of two to three months utilizing a glycohemoglobin test. The test has a variety of names to include A1c, glycated hemoglobin, glycosylated hemoglobin, hemoglobin, and HbA1C. The test examines the average sugar levels over time within the blood stream. The results are given as a single value (percentage form), and are interpreted along a range. The higher the number, the higher the average blood glucose level over the time window. A normal test result is considered any value below 5.7 percent (Mayo Clinic, 2018). Prediabetes is present with an A1c result between 5.7 and 6.4 percent. A patient is considered diabetic with an A1c greater than 6.4 percent.

b. *Cancer Antigen 125 (CA 125):* CA 125 is a membrane-bound protein on the surface of cells and is released into blood. The test is used to monitor the status of cancer before, during, and after treatment. High levels of the protein have been linked with ovarian cancer, pelvic inflammatory disease, pancreatitis, and the first trimester of pregnancy. Sample measurements reading below 46 units per milliliter (U/ml) are considered normal (Mayo Clinic, 2018)

c. Antimitochondria Antibodies (AMA): In an autoimmune response, the body's immune system will attack healthy cells, tissues, and organs. An AMA test measures the amount of antibodies in the blood stream. The normal range for AMA is less than 1.0 units.

In healthcare *s-type* characteristics, it is desirable that the measurement reading from a laboratory test be as small as possible.

3.4.3.1 Healthcare: Step Function for Smaller-the-Better Type Characteristic

A performance function for an *s-type* characteristic estimates the growing performance loss as the measured characteristic deviates from the desired target range. The target range and boundaries for the zones are based on established reference intervals and decision limits. Like the performance function proposed for the *n-type* characteristic, the performance loss is due to mounting medical complications associated with the measured characteristic and increased risk for future health complications. Once the deviation surpasses an upper specification limit the loss value is set at a constant L. Figure 3.10 illustrates an *s-type* medical characteristic which spans three zones of health risk.

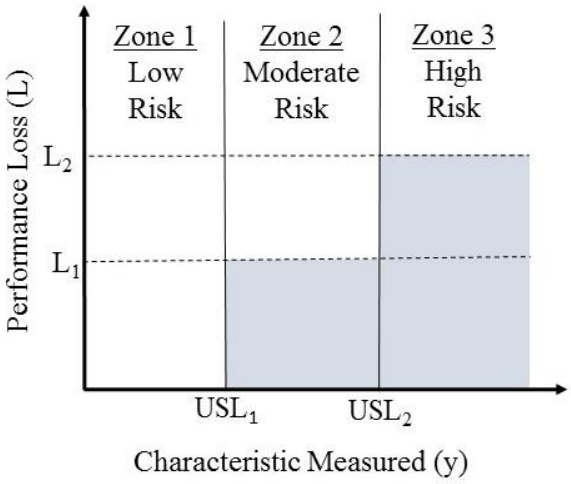


Figure 3.10. Performance Step Function for a Smaller-the-Better Type Characteristic.

The mathematical form the performance step function for the *s-type* characteristic would be as follows:

$$Performance(y) = \begin{cases} 0 & 0 \le y \le USL_1 \\ L_1 & USL_1 < y \le USL_2 \\ L_2 & USL_2 < y \end{cases}$$

The performance step function maps the value of a physical measurement to performance loss. The drawback of using a step function is that it is not sensitive to changes in a patient's measurement when the measurement remains within the same zone. As long as the measurement of the characteristic does not leave a zone, the performance loss will be the same for all measurements within the zone. The next section will propose an alternative to the *s-type* performance step function that allows for more sensitive means of articulating the performance degradation (health risk) of the patient.

3.4.3.2 Healthcare: Performance Function for Smaller-the-Better Type Characteristic

In this section, we will look at altering the *s-type* performance function to be more sensitive to changes in the health measurement. Figure 11 illustrates the *s-type* characteristic with three levels of risk. Zone 1 represents the low risk area and is defined as the normal range for the measured characteristic. Within this range, a patient's measurement falls within the published reference interval. As the patient's reading increases above the published range, the patient's risk for additional health complications increases. In zone 2, the increasing loss line is illustrative of the physical performance loss experienced by the patient. Once a patient's reading reaches zone 3, the patient is at high risk for additional health complications.

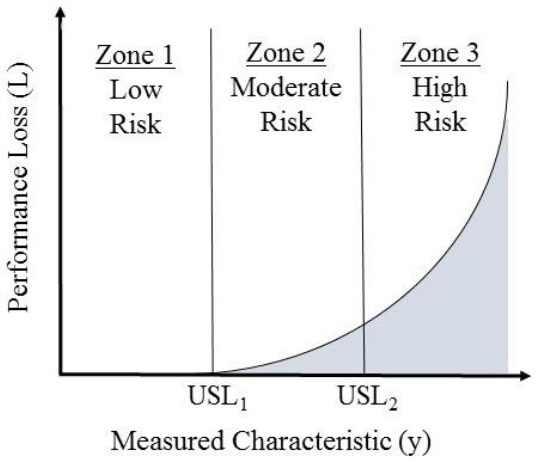


Figure 3.11. Performance Function for a Smaller-the-Better Type Characteristic

The mathematical form of the performance function for an *s-type* medical characteristic would be as follows:

$$Performance(y) = \begin{cases} 0 & 0 \le y \le USL_1 \\ k(y - USL_2)^2 & USL_1 < y \le USL_2 \\ L & USL_2 < y \end{cases}$$

3.4.4 Healthcare: Larger-the-Better Type Characteristic

The last type of health characteristic to be defined is the *l-type* medical characteristic. In this case, the patient's performance peaks when the measured value is as large as possible. *L-type* characteristics have a lower specification limit with an ideal target at infinity. Below are three examples of *l-type* characteristics found in the medical domain.

a. *High-density lipoproteins (HDL):* Lipoproteins help move lipids, fat molecules, around the body. HDL, one of the five major types of lipoproteins, helps to remove fat molecules from cells to the liver. A lipid panel is used to determine the amount of HDL in the body. A healthy amount of HDL is determined to be greater than 60 mg/dL. Patients with less than 40 mg/dL are considered to be at high risk for heart disease(Mayo Clinic, 2018). Patients with at least 60 mg/dL are not considered to be at risk for heart disease.

b. *Vo2 Max:* A practical example of a larger the better type quality characteristic in the healthcare field is the lung capacity of a patient measured as Vo2 max. The test determines cardiovascular and respiratory fitness. The greater the test score achieved by the patient indicates a greater level of fitness.

c. *Strength:* Another quality characteristic of the larger-the-better type could include muscular strength. There is no limit on the amount of strength that a person should have. A lower bound exists so that a person can perform "daily" tasks.

3.4.4.1 Healthcare: Step Function for Larger-the-Better Type Characteristic

A performance function for an *l-type* characteristic needs to be able to estimate the growing performance loss and increased health risk as the measurement deviates to the left of the target range. The target range and boundaries for the zones would be based off the reference intervals and established decision limits. Like the performance function proposed for the *n-type* characteristic, the zones reflect a varying level of performance loss due to mounting medical complications associated with the measured characteristic. Once the deviation surpasses a lower specification limit, the loss value is set at a constant *L*. Figure 3.12 shows an example of a step performance chart for an *l-type* medical characteristic. The example shows an increased risk to the patient's health and a loss of physical performance with a lower reading of the medical characteristic.

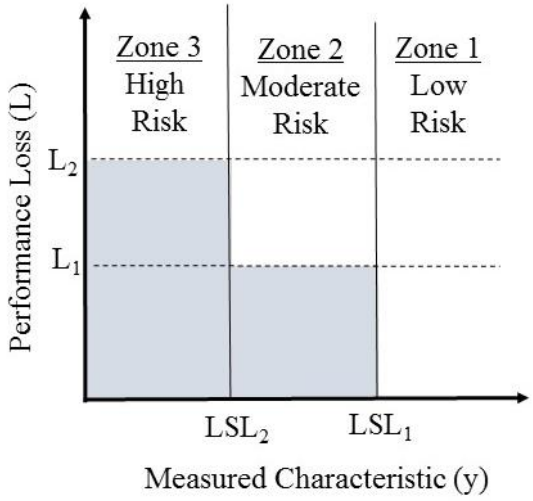


Figure 3.12. Performance Step Function: Larger-the-Better Characteristic.

The mathematical form of the performance step function for an *l-type* characteristic would be as follows:

$$Performance(y) = \begin{cases} 0 & LSL_1 \le y \\ L_1 & LSL_2 \le y < LSL_1 \\ L_2 & y < LSL_2 \end{cases}$$

As stated earlier, the performance step function is a good starting point for articulating physical performance of a patient based on a measurement. Like step functions used within manufacturing, the step loss function is not sensitive to changes in a characteristics measurement if the measurement remains within a defined window. For example at the low end of zone 2, the patient's situation is more precarious than at the high end of zone 2 because of the greater probability of moving into zone 3. Using the step function, the loss value does not adequately convey that risk to the medical professional. The next section will propose an alternative that allows for more sensitive means of articulating the performance degradation (health risk) of the patient.

3.4.4.2 Healthcare: Performance Function for Larger-the-Better Type Characteristic

The step function discussed earlier is not sensitive changes in the measured physical characteristic within a zone's boundaries. By altering the form of the performance function from a step function to a continuous function, the provider may better associate relationship between the measured characteristic and the increased health risk to the patient. Figure 3.13 illustrates the performance loss of an *l-type* medical characteristic across three levels of risk. Zone 1 is the low risk area and encompasses the characteristic's published reference interval. Within this range a patient's measurement

falls within the published "normal" ranges. As the patient's reading decreases below the published "normal" range, the patient has an increased risk for additional health complications. In zone 2, the line is illustrative of the physical loss experienced by the patient. Once a patient's measurement crosses into zone 3, the patient is considered high risk for a decreased physical performance.

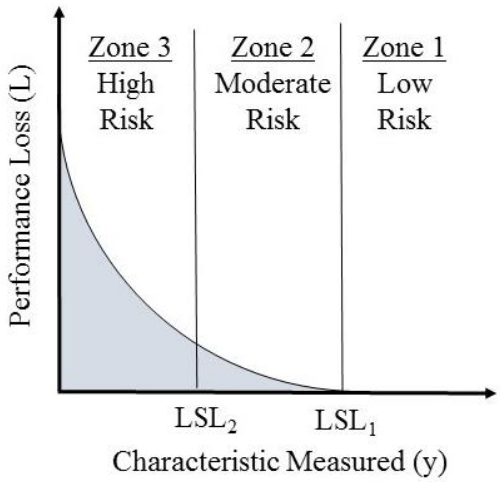


Figure 3.13. Performance Function for Larger-the-Better Type Characteristic.

The mathematical form of the performance function for an *l-type* medical characteristic would be as follows:

$$Performance(y) = \begin{cases} 0 & LSL_1 \le y \\ \dfrac{k}{y^2} & LSL_1 < y \le LSL_1 \\ L & y < LSL_2 \end{cases}$$

The above illustrations are merely starting points for further research in this area. The most important aspect of this conceptual work is the mapping of a physical measurement to associated health risk and the consideration for how the loss

measurement could be used by medical professionals. The potential value in this work is that it provides a way to articulate the trade space of risk within the medical field.

3.5 Proposed Bivariate Performance Functions

3.5.1 Applications of Bivariate Performance Functions

Both the traditional step loss function and the quality loss function described in Section 4 examine performance loss for a single quality characteristic. Given that there are possible trade-offs between the different types of characteristics, better insights might be obtained from looking at the region of interest for two loss characteristics and the resultant mapping of the measurements to a performance valuation. Since the body is a complex system, it would be more appropriate to assess treatment options based on sets of analyte values instead of in isolation. The use of bivariate performance functions might aid medical professionals in evaluating treatment options for multiple symptoms. Bivariate performance loss functions potentially may provide an avenue for treating medical conditions affecting one physical measurement characteristic using treatments whose side effects are known to affect another measurable physical characteristic. With the performance loss function, it would be possible to estimate the total performance loss before the medication is prescribed. For example, select medicines used to relieve high blood pressure can affect the glucose level in the blood stream. In the case of individuals with borderline A1c readings, it would be prudent to assess whether using the blood pressure medication would push the patient's A1c past a decision limit. However, use of the medication might be useful if it did not push the total performance loss to an unacceptably high level. This research is particularly important given the rise of patients

diagnosed with multiple ailments and prescribed multiple drugs. The following sections lead the reader through investigations of three bivariate cases. These conceptual illustrations allow the reader to ponder the possibility of the impact of bivariate performance functions.

3.5.2 Investigation of Two Nominal Type Characteristics

The use of bivariate performance functions would help medical professionals effectively assess the impact of a treatment on more than one physical characteristics. Figure 3.14 illustrates the region of interest for two *n-type* characteristics, *A* and *B*. Ideally, the patient's lab results would fall within the recommended reference intervals, the area between the LSL_1 and USL_1. The area in which both characteristics' measurements are within the "normal" reference interval is named the ideal conformance region. This region is illustrated by the light grey square in the center of Figure 3.14. If either of the patient's lab result measurements for characteristic *A* or *B* exceeds the associated LSL_1 and USL_1 the patient's performance moves from the ideal conformance region into the acceptable conformance region. The acceptable conformance region is denoted by a darker shade of grey than the ideal conformance region. Within this area, the patient is at an increased risk for medical complications. Decision limits are illustrated in this instance as LSL_2 and USL_2. As stated earlier a decision limit is a universally accepted boundary for a specific analyte between "diseased" and "not diseased." While only a few decision limits currently exist, it is expected that more decision limits will be defined in the coming years. As either of the patient's lab result measurements for characteristic *A* or *B* exceeds the associated LSL_2 and USL_2 the patient

moves into the non-acceptable conformance region. Within this region, the patient will experience health complications. The diagram provides a method of visualizing increased risk to the patient or decreased performance for multiple characteristics. Treating the medical issue as a multivariate problem is both more realistic and safer for the patient.

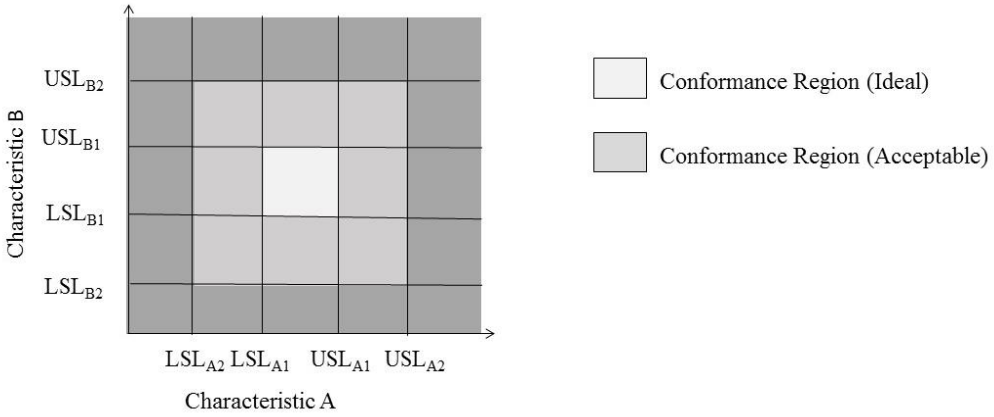


Figure 3.14. Conformance Region for Two Nominal Type Medical Characteristics.

Figure 3.14 illustrates a bivariate performance step function for two *n-type* characteristics. For this example, both *n-type* characteristics are symmetric. Like the performance step function described for the univariate case, loss is only incurred after the characteristic's measurement exceeds the first set of specification limits. For this example, the first set of specification limits is the upper and lower founds for the reference interval. The number of specification limits will depend on the number decision limits associated with the analyte.

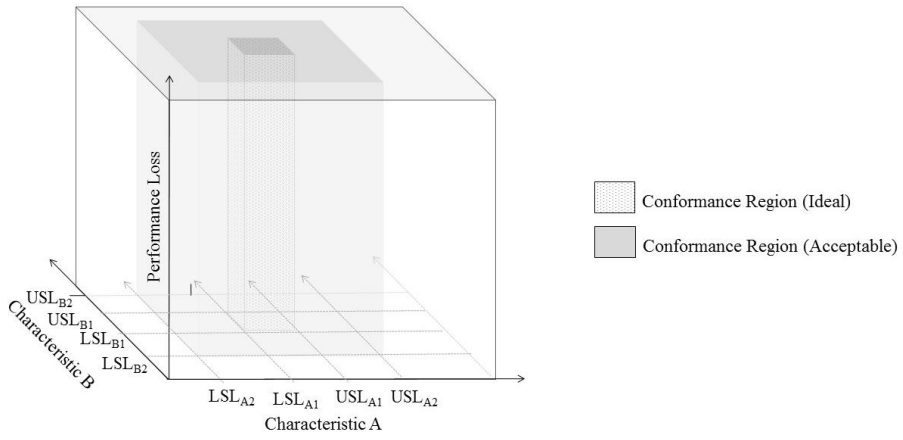


Figure 3.15. Bivariate Step Function for Two Nominal-the-Best Type Characteristics.

Figure 3.15 illustrates a performance function for two, symmetric *n-type* characteristics. Like the performance step function, the value of the performance function is zero if the reference intervals are not exceeded for either of the measured characteristics. Once either characteristic exceeds the upper or lower bound of the reference interval, the performance function takes on a value. The value of the performance loss is the distance from the plane connecting the axis for the values of characteristic *B* and characteristic *A* to the surface curve.

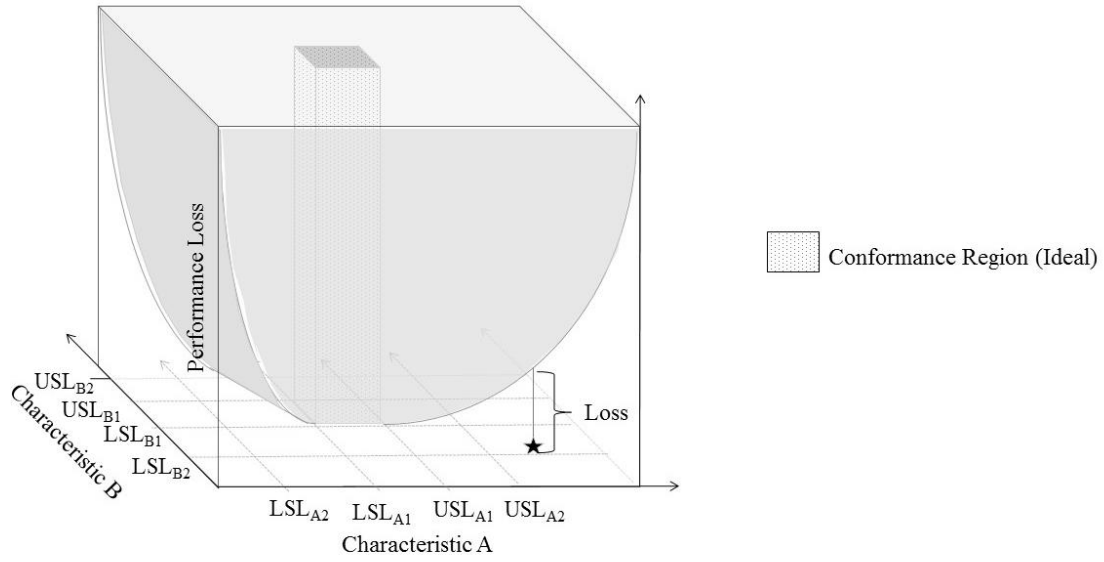


Figure 3.16. Bivariate Performance Function for Two Nominal-the-Best Type Characteristics.

3.5.3 Investigation of a Nominal-the-Best Type & a Smaller-the-Better Type Characteristic

Figure 3.16 illustrates the region of interest of an *n-type* characteristic, *A*, and an *s-type* characteristic, *B*. Ideally, the patient's lab results would fall within the recommended reference intervals which are bounded in the illustration by LSL_1 and USL_1. This region is illustrated by the light grey rectangle in the center of the diagram. As the patient's measured characteristics exceed the reference interval limits, the patient moves into the acceptable conformance region, the next darker area. If the patient's values exceed the next set of limits, LSL_2 or USL_2, for either characteristic, the patient moves into the "non-acceptable" conformance region. Within this area, the patient has a diagnosed health complication.

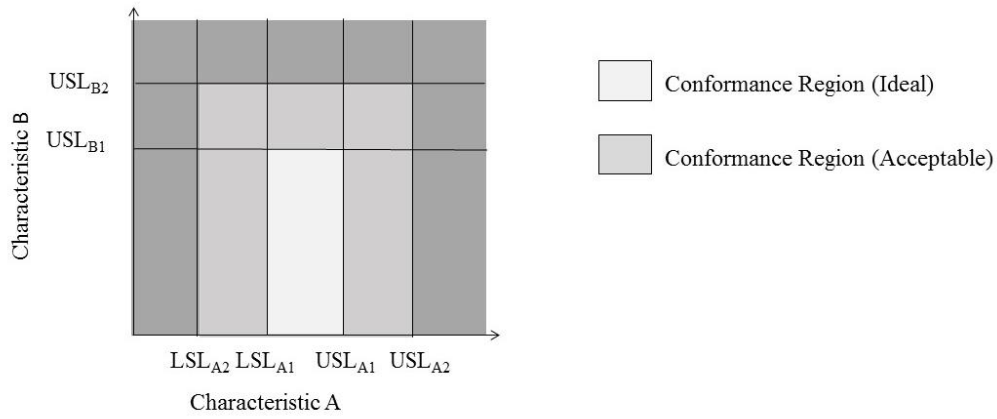


Figure 3.17. Conformance Region for a Nominal-the-Best Type & a Smaller-the-Better Type Characteristic.

Figure 3.17 below illustrates the performance function created by an *n-type* characteristic, *A*, and an *s-type* characteristic, *B*. Ideally, the patient's lab results would fall within the recommended reference intervals, the ideal conformance region bounded by LSL_1 and USL_1 for each characteristic. Within the ideal conformance region, the performance loss is zero. As the measurement for the *n-type* deviates above or below the target interval and as the measurement of the *s-type* characteristic gets larger, the value of the performance function increases. While the patient's values remain between the reference interval and the decision limit values, the patient is considered to be within the acceptable conformance region. Once the one of the patient's lab results indicate that a decision limit has been passed the patient moves into the non-acceptable conformance region. The figure illustrates that the value of the performance increases between the established limits. The goal of medical professionals is to treat a patient so that the overall performance function is minimized.

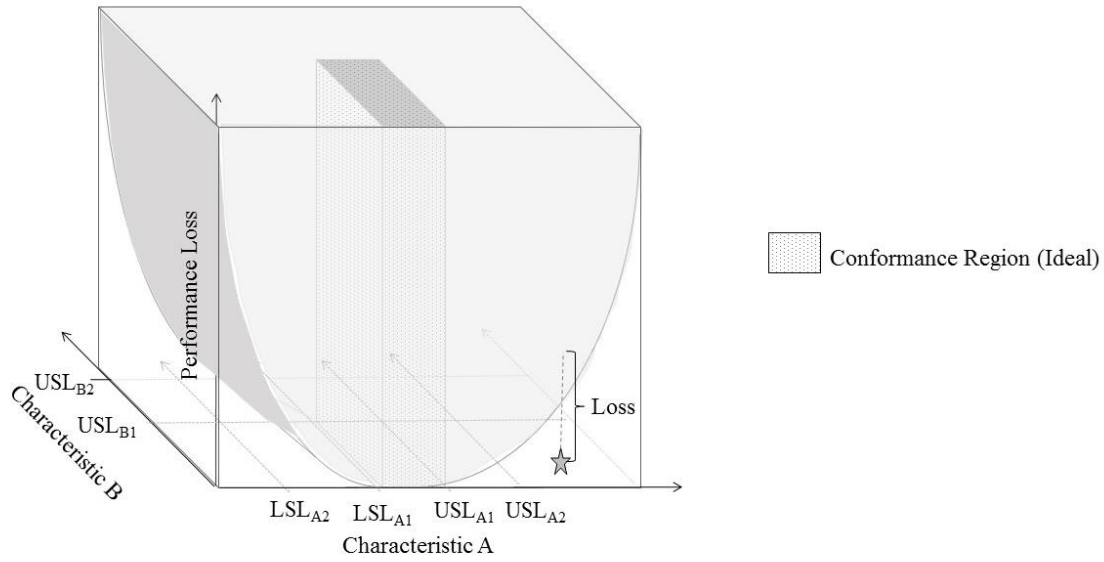


Figure 3.18. Bivariate Performance Function for a Nominal-the-Best Type & a Smaller-the-Better Characteristic.

3.5.4 Investigation of a Nominal-the-Best Type & a Larger-the-Better Type Characteristic

Figure 3.18 illustrates the region of interest for the interaction of an *n-type* characteristic, *A*, and an *l-type* characteristic, *B*. In the ideal conformance region, both measured characteristics are within the established "normal" region. This region is illustrated by the light grey rectangle in the center of the diagram. As either of values for characteristic *A* or *B* passes the first set of specification limits, the interaction between the characteristics enters the acceptable conformance region. The next set of specification limits, as shown on the diagram with a subscript 2, is associated decision limits with the measured characteristic. As either of values for characteristic *A* or *B* passes the second set of specification limits, the interaction between the characteristics enters the non-acceptable conformance region

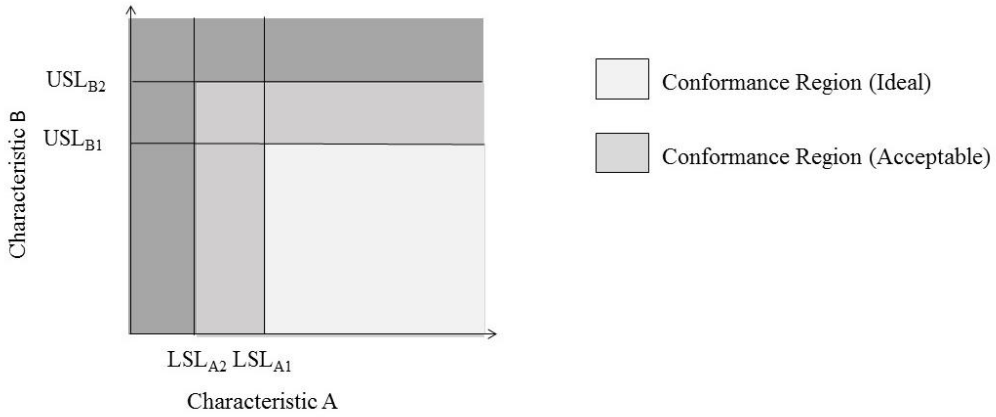


Figure 3.19. Conformance Region for a Smaller-the-Better Type & a Larger-the-Better Type Characteristic.

Figure 3.20 below illustrates the performance function created by an *l-type* characteristic, *A*, and an *s-type* characteristic, *B*. Ideally, the patient's lab results would fall within the recommended reference intervals, the ideal conformance region bounded by LSL_1 for characteristic *A* and USL_1 for characteristic *B*. Within the ideal conformance region, the performance loss is zero. As the measurement for the *l-type* characteristic gets smaller than LSL_1 or the measurement for the *s-type* characteristic increased above USL_1, the value of the performance function increases and the patient enters the acceptable conformance region. While the patient's values remain between the reference interval and the decision limit values, the patient is considered to be within the acceptable conformance region. Once one of the patient's lab results indicate that a decision limit has been passed the patient moves into the non-acceptable conformance region. The goal of medical professionals is to treat a patient so that the overall performance function is minimized.

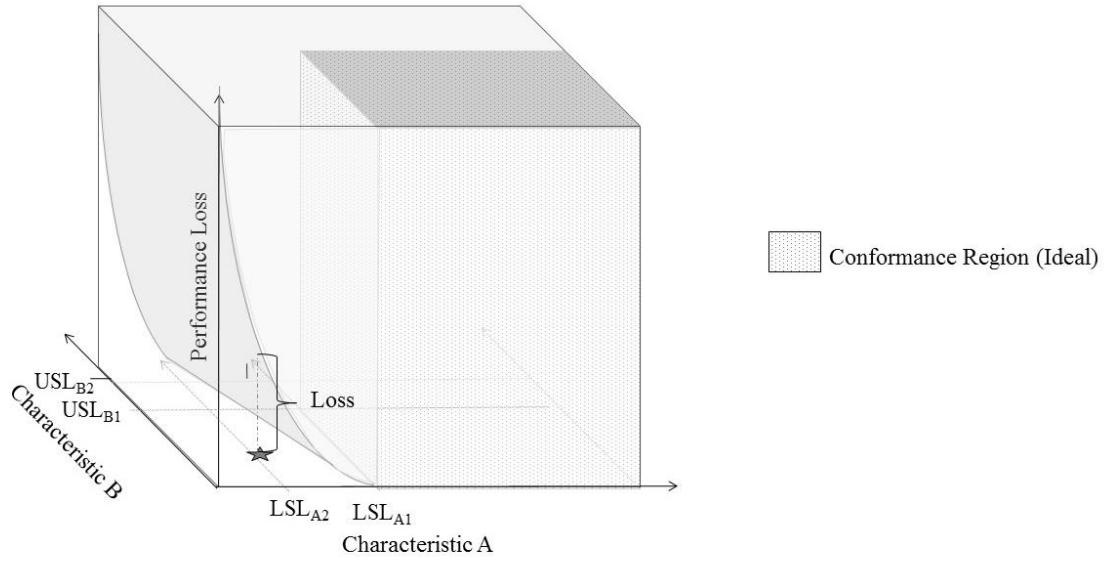


Figure 3.20. Bivariate Performance Function Conformance Region for a Smaller-the-Better Type & a Larger the Better Type Characteristic.

3.6 Conclusion

With the involvement of a diverse and innovative team that spans the entire scientific community, the promise of precision medicine has the best chance of becoming a reality. The purpose of this initiative is to assist medical professionals to more accurately predict the best prevention measures and treatment strategies for a specific disease for an identified group of people. While traditionally involved in solving manufacturing problems, quality engineers have a critical role to play in the development of precision medicine. The purpose of this paper was to initiate a dialogue about how to apply existing quality engineering methodologies to healthcare. Specifically, this paper looked at how to adapt the concept of quality loss functions first developed by Taguchi for use in valuing performance loss and increased risk for future medical complications given biometric measurements. The proposed performance function provides medical

professionals with a quantitative means of relating changes to physical measurements to an individual's overall health. As stated earlier, the most important step in adapting traditional quality engineering methodologies for use within healthcare is in identifying the defining the differences in the problem construction between manufacturing and healthcare. Each identified difference presents an opportunity for the quality engineer to either validate the universality of the methodology or to suggest an alternate methodology.

CHAPTER FOUR

DECISION MAKING IN HEALTHCARE USING ROBUST DESIGN WITH CONDITIONS-BASED SELECTION OF REGRESSION ESTIMATORS

This chapter has been published in Quality and Reliability Engineering International and should be cited as:

Pegues, K. K., Boylan, G. L., & Cho, B. R. (2017). Decision making in health care using robust parameter design with conditions-based selection of regression estimators. *Quality and Reliability Engineering International*, *33*(8), 2151-2169.

The foundation for the aforementioned publication was the joint work with Boylan, G. (2013).

Boylan, G. (2013). Robust Parameter Design in Complex Engineering Systems. (Doctoral dissertation). Retrieved from Tiger Prints, Clemson University.

4.1 Introduction

4.1.1 Research Motivation and Scope

Robust Parameter Design (*RPD*) is an engineering philosophy and statistical method used to determine the optimum conditions that bring the mean process performance towards the desired outcome target with minimum process variability. While the *RPD* methodology has been applied mainly in the manufacturing sector, we believe that this methodology has the potential for greater impact within the healthcare domain. Since tackling *RPD* problems for healthcare sector is a multistage effort, the purpose of this paper is to provide clarification on estimator selection when high variability and asymmetry dominate healthcare process outputs. In particular, a variety of alternative regression approaches are examined via experimental analysis and simulation to determine which methods produce the best solutions.

This chapter provides readers a clear, conditions-based approach for the application of RPD when the conditions of either asymmetry or a high degree of process variability cannot be ruled out. This work is an extension of previous efforts to examine situations in which the parameters needed for ordinary least regression (*OLS*) fail to hold. Data analysis can illuminate intrinsic process conditions that should inform selection of the regression estimation method. In a parallel paper Boylan and Cho (2012) examined distributional characteristics in the context of the four sample moments and investigated how variations in those moments affect the normal probability plot, focusing on the presence of skewness and kurtosis in the data under study. In that instance, residual-based assumptions supporting the use of *OLS* regression were assumed to hold to facilitate comparisons between the estimators considered. The paper also examined how the validity of assumptions associated with underlying populations impacts the resultant statistical analysis of the data. Many of the statistical procedures commonly utilized within quality engineering literature are based upon the assumption of normality. The assumption of normality, although helpful for tractability, may not reflect reality for healthcare applications. As the research community develops a deeper knowledge of medical conditions and underlying causes, a parallel effort within the engineering community should focus on developing better methodologies for dealing with non-normal distributions and asymmetry. The ramifications of utilizing faulty analysis can include increased cost to the organization, physical harm to the patient, or a combination thereof. Given that asymmetry naturally exists in situations involving smaller-the-better (*s-type*) and larger-the-better (*l-type*) healthcare quality characteristics, the use of a

normal distribution for modeling healthcare outcomes may not be appropriate when asymmetrical effects become amplified and are coupled with elevated degrees of process variability. A variety of alternative approaches to regression estimation exists and is documented in both statistics and regression-based literature. This chapter explores the performance of existing regression estimation techniques under varying process conditions with the aims of creating renewed interest in alternative approaches. To simulate asymmetric conditions, a skew normal distribution is integrated into the research. This distribution, also known as an asymmetric Gaussian curve, generalizes the normal distribution to allow for non-zero skewness.

The combination of experimental investigations and simulation allow us to evaluate which regression approach performs best in terms of producing the best results under examined conditions. A sequence for implementing this approach is portrayed in Figure 4.1. Monte Carlo simulation and numerical case studies are used in Section 4.3 to provide clarification as to which estimators should be considered in Phase Ib. In Section 4.2, a proposed healthcare specific methodology is developed using the skew normal distribution as the basis for modeling system attributes. In Section 4.3, the numerical demonstration provided is composed a case study with Monte Carlo simulation. Finally, in Section 4.4, the results are analyzed.

4.1.2 Robust Parameter Design: Development and Application

Although many researchers endorse the philosophical arguments behind Taguchi's (1986, 1987) original version of RPD methods his mathematical approaches have generated criticism. The differing viewpoints of the research community regarding

the validity of Taguchi's assumptions, the varying assessments as to the effectiveness of the approach, and the associated analytical methods are thoroughly documented by many researchers, including include Box (1988) and Tsui (1992). A review of the RPD literature since 1980 reveals that a majority of the work focused on alternative optimization techniques. The degree to which a response surface yields a "good fit" is contingent the correct identification of the prevailing conditions of the sampled data and the appropriateness of the method used to develop the fitted model. Researchers, by proceeding forward into the optimization phase of research, are endorsing that that the parameter estimates are sufficient for use, that estimates are obtained using appropriate estimators, and that prevalent conditions within the data support the chosen parameter estimator technique.

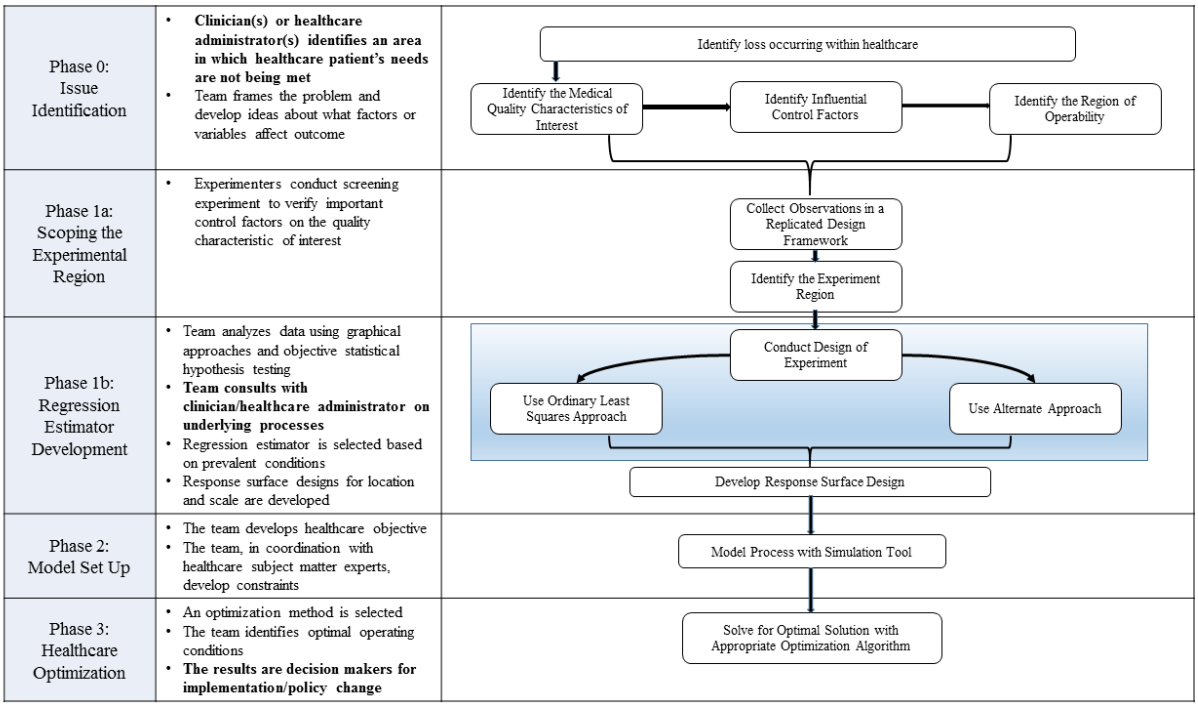


Figure 4.1. Methodology process map for healthcare applications.

A review of statistics-based literature highlights the variety of existing regression estimation methods. The volume of literature focused on alternative estimation approaches quantitatively demonstrates the need for better results than those found with *OLS*. Many of these alternatives were developed to overcome issues associated with outliers, contaminated data, non-normality in the responses and/or residuals, and heteroscedasticity. The spectrum of estimation methodologies includes data transformations, generalized linear models (*GLM*), the weighted least squares (*WLS*), and an assortment of various resistant and robust regression techniques to include least trimmed squares (*TLS*), least absolute deviation (*LAD*), M-estimation, M-M estimation, and S-estimation. A selection of these methods will be discussed in greater detail in Section 4.3.3.

To overcome non-normality of a data sample, researchers applied transformations on the response (Y) or utilized *GLMs*. As Ryan (2009) observed, the drawback is that the use of transformations can induce change in both the distribution of data about the regression line and the vertical spacing of the observed values. Since the need to transform data may stem from a few influential observations, researchers are urged to use caution as they proceed. *GLMs*, another approach for dealing with non-normality, have received considerable attention as a practical alternative to transformations (Myers *et al*., 2002, Myers *et al*., 1997). The critical aspect of GLMs is the use of a smooth monotonic link function $g(\cdot) = EF(\mu_i, \phi)$ from any distribution within the exponential family. In a sense, $g(\cdot)$ acts like a transformation and link functions transform $E[Y_i]$ rather than y_i

itself. Therefore, $g(\cdot)$ transforms the systematic part of a model without altering the distribution of the associated random variation.

Published papers have aimed to determine the limitations of estimation approaches and performed limited comparisons between alternatives. Several research efforts examined the performance differences between select estimation approaches (Koutrouvelis *et al*., 2000; Bera *et al*, 2002). Other research efforts investigated robust estimators (Muhlbauer *et al*., 2009; Hamada *et al*., 1997). These works focused primarily on statistical measures (relative efficiencies, breakdowns, and robustness) to establish the superiority of one method relative to other methods. Interestingly, whereas *GLMs* have seen attention in the RPD literature, comparatively few RPD-specific efforts have explored the various resistant regression methods as viable alternatives for determining optimal solutions. Among those that have, the process conditions examined focus predominantly on outliers, non-normal (but symmetric) response distributions, and unbalanced data sets. Table 4.1 summarizes several of the more contemporary research efforts found in the RPD literature, which are addressed in some detail in the following paragraphs.

Table 4.1. Summary of works examining estimator selection in *RPD* problems.

Researchers	Year	Conditions Examined	Regression Estimators Compared
Simpson & Montgomery[34]	1998	Outliers under normality	OLS, M, most-B robust, LTS, S, M-M, various versions of Generalized M
Lee & Nelder[35]	2003	Non-constant variance and non-identity (Gaussian) link functions	GLM
Cho & Park[36]	2005	Unbalanced datasets	OLS vs. WLS
Ch'ng *et al*.[37]	2005	Non-normal responses and outliers	OLS vs. M-M
Robinson *et al.*[38]	2006	non-normal (gamma) and batch-to-batch variation (random block design)	GLMM (gamma with log link)
Lee *et al.*[39]	2007	Outliers, non-normal symmetric distributions	OLS vs. M-M
Goethals & Cho[40]	2011	Heteroscedastic conditions and unbalanced data	OLS vs. WLS

Regarding GLM*s*, Lee and Nelder (2003) examined their use as a generalization of data transformation and *RSM* approaches that allowed for "arbitrary variance and link functions." In a more recent effort, Robinson *et al.* (2006) examined generalized linear mixed models (*GLMM*) in an *RPD* context to address the non-normality encountered with a resistivity quality characteristic by using the known distribution for the response (Gamma) combined with a log link. While the results in each of these works clearly demonstrated the potential benefits of utilizing *GLM*s, they were not necessarily comparative studies.

The remaining works shown in Table 2.1 pertain to more direct comparisons between traditional and robust regression approaches. Simpson and Montgomery (1998) examined alternative regression techniques when dealing with outliers within normally-

distributed data. However, this study focused more on statistical estimator performance measures such as efficiency, consistency, and breakdown points rather than optimal *RPD* solutions obtained through application. Cho and Park (2005) considered *RPD* solutions in the case of unbalanced data and proposed the integration of a *WLS* approach. The proposed weighting scheme was based upon the quantity of observations at each design point and value at the design point is inversely proportional to the variance associated with the response surface functions obtained for the process parameters. In the interest of finding better optimal settings in dual-response surface optimization problems when non-normal conditions and/or outliers exist, Ch'ng *et al*. (2005) compared *OLS* to the *M-M* robust estimation technique developed by Yohai (1987). In the examination of estimators in *RPD* involving contaminated data, Lee *et al*. (2007) also included a comparison of the *OLS* method to the *M-M* regression technique. It is worth noting that the non-normal conditions examined in both of these cases focused on symmetric distributions. Goethals and Cho (2011) extended the work of Cho and Park (2005) to the optimal process target problem. Their work considered heteroscedastic conditions in addition to the unbalanced data case.

4.2 Potential Applications of RPD in Healthcare Environments

In an ideal world, medical treatment for a particular issue will have a beneficial impact on patients. For each instance, a patient would discover a problem, would then be diagnosed, a treatment plan outlined, and the patient is cured of the initial recognized ailment. This simplified scenario is not today's reality. Healthcare is a complicated,

complex system in which a large number of factors to include a patient's health, current intuitional knowledge regarding complaints, resource allocation, provider skill, the limitation of known developed protocols as well as other factors which play a role in final outcomes. The combination of compelling emotional medical narratives and the increasing costs of healthcare has spurred governmental agencies, public and private business ventures, and research and development teams for both industry and academia to focus on decreasing the cost (emotional and financial) of healthcare to society at large. As written, the "cost of healthcare" denotes the sum total cost of healthcare for both treated and untreated diseases and preventive measures for all aspects of health.

With a common goal of improving healthcare, powerful stakeholders for the US healthcare system are seeking efficiencies that will improve care and reduce overall costs. Another change to the United States' medical system has been brought about by technological advances of the computer age. Medical records are in the process of being digitized and the large swaths of data for medical procedures and vital statistics are available for research. The large data repositories by controlled agencies such as the National Institute of Health have enabled research teams to start dissecting complicated issues such as cancer treatment and diabetes management. Research teams are also working to improve other aspects of healthcare from operating room efficiency to prosthetic development. At regional and local levels, management teams are working to make hospital systems operate more efficiently. These parallel, and for most cases, unlinked efforts improve healthcare in small ways on a daily basis. Small changes

improve lives, but the question of how healthcare policies, both nationally and locally, should be established so that each patient has the optimum chance of recovery remains.

To the best of our knowledge, it is believed that one of existing methodologies that could significantly impact patient recovery is design of experiments (DOE). The application of DOE is not new in the healthcare industry and its professional workforce has long understood important roles of carefully-designed experiments. The most widely-used DOE tools are perhaps full factorial designs and fractional factorial designs with several factors, each at two discrete levels, in order to study the effects of main factors and interactions between those factors on the response variable of interest. When the number of levels becomes more than two, the number of experimental runs required considerable increases; thus, it becomes less feasible from the perspective of costs incurred and resources available to complete the whole experiment. In addition, when those two-level factorial designs are used, one of the fundamental assumptions is that the effects between the two levels are linear. The advancement of precision medicine relies on the adoption of evidence-based practices and process validation. Both linear and quadratic effects of factors and interactions are often inherent within healthcare data. Capturing those effects can be effectively done by the central composite design and obtaining optimum conditions through the *RPD* process.

The methodology for *RPD* is broken down into two phases. The first phase includes the identification of the primary quality characteristic, influential control factors, possibly noise factors, and the experimental region of interest. Control factors are those that affect outcomes. In the case of a healthcare system, they include the amount of drug

that a patient is provided or the existence of known, and genetic traits of a patient. Noise factors are those factors that may not be controllable or very costly, if they can be controlled, from the point of view of the system. For a healthcare example, those could be a patient's diagnosis, the patient's satisfaction with the benefits provided by his or her insurance company, or the length of the patient's trip to the hospital. The idea is to find optimum conditions for the control factors for which changes along the range of noise factor values affect healthcare outcomes as little as possible. While in this phase, observations are collected in a replicated design framework and data analysis is performed. The outcome of this phase is an approved model for the quality characteristics of interest. These models are then used to determine optimum conditions which allow the system to reach close to a specified target with minimum variance.

A comprehensive literature review shows that the *RPD* concept has not been applied rigorously in healthcare environments. Some potential applications of *RPD* for healthcare decision making are outlined below.

Patient Adherence to Treatment Plans: It has been estimated that less than 60% of prescribed treatment plans are followed by patients. Why, if the patient has sought out medical advice, is the final treatment plan not being followed? Possible factors could include the treatment costs, insurance benefits, medical severity, social factors, and risk of being admitted. *RPD* could help doctors determine optimal conditions, or hospital/insurance policies, which would achieve a target adherence rate with minimum variability.

a. *Incentive for Preventative Treatments*. *RPD* could potentially provide insights on how to increase the use of preventive care services made available through insurance programs, community wellness programs, or direct government funding. What services or incentives need to be offered to reduce the impact of uncontrollable factors on the use of preventive care?

b. *Gauging Risk. RPD* could provide an avenue to make assessments on the amount of risk (variance) that stakeholders are willing to take on. If the quality characteristic being studied is the effectiveness of a treatment, then being on target means that the patient is being effectively treated. As it is further away from the target, greater costs will be incurred by the healthcare system. Also, a smaller variability implies a less risk that the health industry is willing to take on.

c. *Improved Resource Allocation*. During the first phase of *RPD*, stakeholders gain a better understanding of specific control factors that influence the quality characteristic of interest. By exploring the interactions, stakeholders will have a better idea of how to allocate scare resources to improve the overall quality of healthcare.

4.3 Proposed RPD Modeling and Optimization Procedures

4.3.1 Experimentation and Analysis

Consider a situation in which we need to obtain the optimal conditions for which a patient can safely undergo surgery without further compromising health. A clinician might be interested in determining the range for a patient's body composition, resting heart rate, and possible other mitigating factors that must be met prior to commencing the

procedure. To that end, consider a replication-based experiment conducted with the intent to find optimal factor settings, **x**=($x_1, x_2, \ldots, x_k$) that achieve the desired target outcome, with the least variability. In this case, the quality characteristic of interest, *Y*, is suspected to be influenced by a set of control factors or **x**. The research team will define an experiment region which is bounded by the minimum and maximum values for each of the control factors. The experiment consists of *m* x *n* trails where *m* is the number of replications for each specified design point, *n*. For each design point, the values of the control variables are set to predetermined levels. Let y_{qj} denote the j^{th} response at the q^{th} design point, where $q = 1, \ldots, n$ and $j = 1, \ldots, m$. The tabular layout of a replication-based design of experiment is depicted in Table 2.2.

Table 4.0.1 Design of Experiment

Design Point	$x_1, x_2, \ldots, x_k$	Replications	$\bar{y}$	s	γ
1	Control factor settings	$y_{11} \ldots\ldots y_{1j} \ldots\ldots y_{1m}$	$\bar{y}_1$	s_1	γ_1
⋮		⋮ ⋮	⋮	⋮	⋮
q		$y_{q1} \ldots\ldots y_{qj} \ldots\ldots y_{qm}$	$\bar{y}_q$	s_q	γ_q
⋮		⋮ ⋮	⋮	⋮	⋮
n		$y_{n1} \ldots\ldots y_{nj} \ldots\ldots y_{nm}$	$\bar{y}_n$	s_n	γ_n

Parameter estimates for the data are found through analysis of the data collected at each design point. The estimates include both sample mean, $\bar{y}$, and standard deviation, *s*. If the distribution is suspected to be non-normal, the sample skewness, γ, can also be calculated to account for the asymmetry in the responses. Parameter estimates at each design point for a sample are found using the following equations:

$$\bar{y}_q = \frac{\sum_{j=1}^{m} y_{qj}}{m}, \quad s = \sqrt{\frac{\sum_{j=1}^{m} (y_{qj} - \bar{y}_q)^2}{m-1}}, \text{ and } \quad \gamma = \frac{\frac{1}{m}\sum_{j=1}^{n} (y_{qj} - \bar{y}_q)^3}{\left(\frac{1}{m-1}\sum_{j=1}^{n} (y_{qj} - \bar{y}_q)^2\right)^{3/2}} \tag{1}$$

The next step is to develop response surface functions for the select parameters of interest that are valid throughout the experimental region. To start, a comprehensive data analysis of the sample responses and the residuals, the difference between the model and the sampled responses, is conducted to verify assumptions regarding the underlying distribution. The analysis of the responses should include an investigation of normality and variability. The analysis of the residuals should include verification of the assumptions of normality, homoscedasticity, and independence. Investigators have the option to use graphical methods, numerical methods, and formal normality tests. Two approaches are briefly explained in parts (*i*) and (*ii*) below:

(*i*) Assessment of Normality and Variability. To assess normality in a set of responses, the three most common tests are the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Shapiro-Wilk test. Generally, due to a variety of reasons including experimental costs, the sample sizes obtained in *RPD* experimentation typically small. Thus, for the Shapiro-Wilk test using the *W* statistic given by $W^* = \left[b/s\sqrt{p-1}\right)\Big]^2$ where $b = \sum_{i=1}^{\kappa} a_{p-i+1}\left[y_{p-i+1} - y_i\right]$ may be a reasonable choice for testing $H_0: Y \in N(\mu, \sigma^2)$ and $H_1: Y \notin N(\mu, \sigma^2)$. The term *p* is defined as the number of observations sorted in ascending order, κ is the largest integer that is less than or equal to *p/2*, and *s* denotes the sample standard deviation. For a given significance level α, tables are then used to reference the coefficients *a*, and the critical values W_α. If $W^* > W_\alpha$, then insufficient evidence

exists to reject the assumption of normality. One drawback to objective tests is that if the normal assumption is rejected, the test provides no additional information regarding the underlying distribution of the data. As discussed by Boylan and Cho (2012), graphical measures, such as normal probability plots, may be useful tools to overcome this shortfall and to provide more salient information about the data. For example, normal probability plots often illustrate whether the distribution is symmetric or asymmetric, the degree of positive or negative skewness, the degree of variability, and the degree of kurtosis.

Determining whether a process is highly variable is a more subjective assessment. According to Willinger *et al*.(2004), high variability may be loosely defined as a phenomenon by which a set of observations assumes values that vary over orders of magnitude, with most taking closely grouped values, a few assuming extreme values that deviate considerably from the first group with non-negligible probabilities, and intermediate observations occurring with appreciable frequencies. In general, a trademark of highly variable data is that the sample standard deviation is quite large. This result implies a "largely uninformative" sample mean that does not adequately describe the location of the bulk of the observed values. Using this concept, we classify a highly variable process as one in which the range of variability in the responses is noticeably large and where one or more of the responses lies more than three standard deviations ($+/- 3\sigma$) from the mean response.

(*ii)* Residual analysis. As with the responses, normality in the residuals may also be examined using graphical measures such as the normal probability plot. Additional

complementary, objective methods are provided by the Kolmogorov-Smirnov or the Shapiro-Wilk tests. To investigate independence, the Durbin-Watson test is usually sufficient to detect a lack of randomness in the residuals. Should remediation be necessary, two possible options are the addition of predictor variables or use transformations in the variables to eliminate interdependencies. Finally, heteroscedasticity, or non-constant variance, is often investigated graphically using a plot of the residuals against the fitted values, as well as objectively using either the Brown-Forsythe test, which is more robust to departures from normality in the data, or the Breusch-Pagan (B-P) test. The B-P test assumes independence and normality among the residuals. The test also assumes a relationship for the error variance σ_q^2among the *k* regression coefficients and *k*-1 predictor variables that in the form $log_e\sigma_q = \gamma_0 + \gamma_1 x_{q1} + \cdots + \gamma_{k-1} x_{q,k-1}$. As can be seen, the error variance fluctuates up or down with **x**, based on the sign of the associated coefficients. Constant error variance corresponds to the instance where constrained coefficients in response function equal 0, the alternative hypotheses H_0: $\gamma_1 = \cdots = \gamma_{k-1} = 0$ versusH_1: not all $\gamma_i = 0$ are tested using the statistic, $X_{BP}^2 = (SSR^*/2) \div (SSE/Nm)^2$, in which *Nm* denotes the total number of experimental observations, *SSR** is the regression sum of squares obtained by regressing the squared residuals, and *SSE* is the error sum of squares obtained for the full regression model. If $\chi_{BP}^2 > \chi_{(1-\alpha),k-1}^2$ then we reject *H0* and conclude that sufficient evidence exists to support non-constant variance. In processes with high variability or asymmetry in the responses, the assumption of constant variance in the residuals would

most likely not hold. This situation would necessitate the use of remedial measures, as outlined in Section 4.3.3.

4.3.2 Modeling Symmetry and Asymmetry

In traditional *RPD* applications, asymmetric conditions typically observed in the univariate *s*- and *l-type* problems are often modeled via a normal distribution. Ideally, it would be preferable to use a distributional model capable of supporting both the symmetry usually assumed in the *n-type* model, as well as the asymmetry of the *s*-, *l*-, and certain *n-type* models. This would become particularly important if extending the problem to the multi-response case. Although some common distributions, such as the gamma, Weibull, and unbounded Johnson distributions, can effectively portray processes with innate skewness, these distributions present challenges in modeling normality when small skewness exists.

Due to an inherent relationship to the normal distribution, the skew normal (*SN*) distribution provides a suitable alternative for modeling both symmetric and asymmetric situations. First introduced by O'Hagan and Leonhard (1976) and addressed more recently by Azzalini (1985), Azzalini and Dalla-Valle (1996), and Arellano-Valle *et al. (2004)*, the skew normal distribution extends the normal distribution by incorporating a third parameter, α, as a shape parameter to account for non-zero skewness. The probability density function for the skew normal relative to the normal distribution is given by:

$$f(x \mid \alpha) = 2\phi(x)\Phi(\alpha x), \quad x \in \mathbb{R},$$

where $\phi(x)$ and $\Phi(\alpha x)$ correspond to the probability density and cumulative distribution functions of the normal distribution, respectively. Recall that the normal probability density function for some random variable Z with parameters μ and σ^2 can be rewritten in terms of the standard normal density function

$$f_z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right) \tag{2}$$

We can easily extend this by adding location (ξ) and scale (ω) parameters to the density function, using the transformation x→(x-ξ)/ω. This yields:

$$f_x(x) = \frac{2}{\sigma}\left[\phi\left(\frac{x-\xi}{\omega}\right)\right]\Phi\left[\frac{\alpha(x-\xi)}{\sigma}\right] = \frac{2}{\omega}\left[\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-\xi)^2}{2\omega^2}\right\}\right] \cdot \frac{1}{2}\left[1+\operatorname{erf}\left\{\frac{\alpha(x-\xi)/\omega}{\sqrt{2}}\right\}\right]$$
$$= \frac{1}{\omega\sqrt{2\pi}}\exp\left\{-\frac{(x-\xi)^2}{2\omega^2}\right\} \cdot \frac{1}{2}\left[1+\operatorname{erf}\left\{\frac{\alpha(x-\xi)}{\omega\sqrt{2}}\right\}\right] \tag{3}$$

When $\alpha = 0$, the skew normal distribution reduces to the normal, making normality a special case of the $SN(\xi, \omega, \alpha)$ distribution. From Azzalini (1985), the mean and standard deviation of a $SN(\xi,\omega,\alpha)$ distribution are given by:

$$E[y] = \hat{\mu} = \xi + \omega\delta\sqrt{2/\pi} \tag{4}$$

$$s = \sqrt{\omega^2\left(1-2\delta^2/\pi\right)} \tag{5}$$

Where $\delta = \alpha/\alpha\sqrt{(1+\alpha^2)}$. The *SN* distribution is a relatively new distribution compared to the more commonly observed family of continuous distributions. Since it derives from the normal distribution which remains widely used for *n-type* characteristics, its extension

to *s*- and *l-type* characteristics, as well as certain instances of *n-type* characteristics, may help to overcome many -modeling complexities encountered in asymmetric situations.

As Goethals and Cho (2012) showed, modeling system properties with the skew normal distribution can be achieved by initially calculating estimates for the first three sample moments (mean, standard deviation, and skewness) for the q^{th} design point. In Table 2, the sample mean $\bar{y}_q$ and standard deviation s_q then correspond to the location (ξ_q) and scale (ω_q) parameters at the q^{th} design point. Thereafter, $\bar{y}_q$ and s_q estimates are used to derive estimates for the skew normal process mean and standard deviation by applying them to Equations (4) and (5) as follows:

$$\hat{\mu}_{q(SN)} = \bar{y}_q + s_q \delta_q \sqrt{2/\pi} \tag{6}$$

$$s_{q(SN)} = \sqrt{s_q^{\,2}\left(1 - 2\delta_q^2/\pi\right)} \tag{7}$$

Here, the parameter δ_q is estimated using the sample skew. In short, using an alternative formulation for sample skew provided by Azzalini (1985), an estimate for δ_q can be derived as follows:

$$\hat{\gamma}_q = \frac{4-\pi}{2}\frac{\left(\delta\sqrt{2/\pi}\right)^3}{\left(1-2\delta^2/\pi\right)^{3/2}} \quad \Rightarrow \quad \left|\hat{\delta}_q\right| = \sqrt{\frac{\pi}{2}\cdot\frac{\left|\hat{\gamma}_{3q}\right|^{2/3}}{\left|\hat{\gamma}_{3q}\right|^{2/3} + ((4-\pi)/2)^{2/3}}}$$

where the sign of $\hat{\gamma}_q$ determines the sign of δ_q. For the purposes of simulation, the estimate for δ_q may then be used to estimate the shape parameter directly by rearranging the previously stated relationship in the following way:

$$\delta_q = \frac{\alpha_q}{\sqrt{1+\alpha_q^2}} \quad \Rightarrow \quad \hat{\alpha}_q = \frac{\hat{\delta}_q}{\sqrt{1-\hat{\delta}_q^2}}$$

This step is necessary within the R environment as the shape parameter and sample skew are scaled differently in the context of the skew normal distribution. As Equations (6) and (7) suggest, the estimates for the process mean and standard deviation are influenced by the inclusion of the sample skew. By using this methodology, we ensure that inherent process skewness is accounted for in the final response surface estimates, and that actual process characteristics are more accurately represented.

4.3.3 Selecting an Appropriate Regression Estimator

Prior to completing a comprehensive analysis, a research team needs to evaluate regression estimation methods and decide on which one to implement. For many groups, *OLS* is chosen due to familiarly as well its ubiquity in past research efforts. To obtain an optimal *RPD* solution, we advocate that teams instead match the regression estimation method with the inherent underlying conditions of the dataset. Table 2.3 below lists the ten methods selected for analysis. In the subsequent paragraphs, a brief synopsis of each alternative method is provided.

Table 4.0.2. Regression Estimators examined as potential *RPD* alternatives.

	Methods for Determining Regression Estimator
Base Case	*OLS*
Alternatives for Comparison	1) *GLM* (gamma or inverse Gaussian model) 2) *OLS* (SN) 3) *WLS* ($\bar{y}$ and *s*) 4) *WLS* (median and *MAD*) 5) Least trimmed squares (*LTS*) 6) *S*-estimation 7) Least absolute deviation (LAD) 8) *M-M* estimation 9) *M*-estimation (Huber Proposal 2)

In early research, Kutner *et al*. (2003) stated that asymmetry and high variability would likely bring about non-constant variance. If this is the case, then *OLS* standard errors are potentially inaccurate and statistical inferences based on the results could be potentially misleading. Thus, an alternative method may provide a better estimate for the regression coefficient.

The first alternative approach considered is that of the generalized linear model (*GLM*). The *GLM* is the conventional regression approach for data sets exhibiting non-normality. When applying the *GLM* method, a practitioner needs to specify the linear predictors' distribution and select an appropriate link function, g(·). For the purposes of this research effort, the *GLM* method utilizes a gamma or an inverse Gaussian. These distributions are suitable for modeling varying degrees of asymmetry. Link functions transform the expected value of the response to the linear predictor and assume the form $g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$. By using *GLM*s, the selection of the link function is distinct from the distributional assumption. Although a wide variety of link functions exist, this research is limited to those listed in Table 2.4.

Table 4.3. Applicable link functions for the gamma and inverse Gaussian distributions.

Link Function	$\eta_i = g(\mu_i)$	Gamma	Inverse Gaussian
Identity	μ_i	X	X
Log	$\log_e \mu_i$	X	X
Inverse	μ_i^{-1}	X*	X
Inverse-square	μ_i^{-2}		X*

* denotes the default link used by the `glm()` function in R.

Process conditions influence the selection of both the distribution and link functions. Individual distributions have recommended link functions that typically produce preferable mathematical and numerical properties. Given a suspected

distribution, practitioners should be cautioned to take the underlying process conditions for a particular set of data into consideration prior to using the traditionally associated link functions. For example, in many applications, the gamma distribution is paired with the inverse link. Consider a case in which the response *y* increases somewhat linearly with control factor x_i. If the variance appears to increase with the square of the mean, then the gamma distribution could be paired with the identity link. When choosing the appropriate link response, researchers should select the *GLM* model that results in the lowest Akaike information criterion value, or in other words, the smallest residual deviance. This recommendation is based on the fact that the response deviance is similarly scaled for models created with the *GLM* method.

When dealing with asymmetric conditions, two factors bear considerable importance. First, depending on the degree of variability and skew, the mean will shift away from the central tendency of the distribution. Secondly, the standard deviation may not accurately describe the dispersion in the distribution as it tends to be significantly affected by the "play" in the skewed or long tail of the distribution. When the data contains outliers, the effects would worsen. For these reasons, an alternative method of weighted least squares (*WLS*) is also examined in this paper. The *WLS* method was among the first to tackle how to perform analysis taking residuals into account. The method of *WLS*, developed by Aitken in 1935, alters the value placed on data dependent on the deviance of the residual. If the residual of the q^{th} point is relatively small, then it is assigned a relatively large weight. On the other hand, if the q^{th} residual is large, the

impact of the outlier will be reduced utilizing a smaller associated weight. The *WLS* is described with the following equation:

$$\underset{\hat{\beta}}{\text{Minimize}} \sum_{q=1}^{n} w_q \varepsilon_q^{\,2} \tag{8}$$

where $\boldsymbol{\varepsilon_q}$ denotes the residual associated with the q^{th} design point given by $\boldsymbol{y_q - x_q \beta}$. The term y_q refers to the sample mean if there are multiple replications per design point. For cases in which the data is assumed symmetric, the mean is typically the "starting point" or the value what has the greatest probability of occurring. However, in cases where asymmetry is suspected, the two options for alternate "starting points" are median and median absolute deviation (*MAD*). To illustrate this point, consider the comparison of probability densities for samples drawn from a skew normal distribution and a normal distribution with the same mean shown in Figure 2.2. For the sample with a normal distribution, the mean value corresponds to the peak in the density function. For the sample with this particular skew normal distribution, however, the mean lies to the right of the preponderance of probabilities. Therefore, in this case, the mean does not have the greatest likelihood of occurrence. However, an alternate measure, the median, occurs close to the peak of the distribution. This example illustrates that while the mean certainly defines the population's central tendency for any distribution, the mean does not necessarily correspond to greatest likelihood if asymmetry is present.

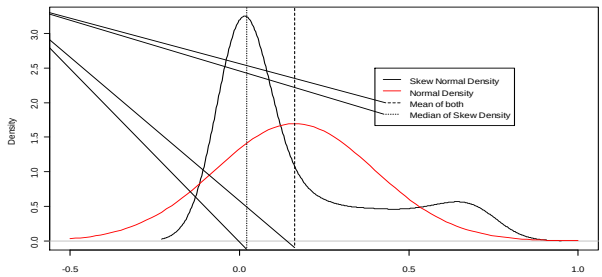


Figure 4.1. Comparison of normal and skew normal densities with the same sample mean.

Based on the conditions of asymmetry and variance examined in this paper, the weights for the *WLS* method are determined in a manner similar to that used by Goethals and Cho (2011). That is, observations possessing less variance receive greater weight. Denoting **ε** as the vector of residuals, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ in the general form of the standard regression model $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\beta\varepsilon}$ in which **Y** is the vector of responses, **X** is the design matrix, and **β** is the vector of estimated regression coefficients, in the case of non-constant error variances, we may rewrite the $n \times n$ covariance matrix as:

$$Var(\boldsymbol{\varepsilon}) = \begin{bmatrix} Var(\varepsilon_1) & 0 & \cdots & 0 \\ 0 & Var(\varepsilon_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Var(\varepsilon_n) \end{bmatrix}$$

Since $E(\varepsilon_q) = 0$ for each of *n* components of **ε**, the variance of the q^{th} component is equal to the expected value of the squared error of the *q*th component, or $Var(\varepsilon_q) = E(\varepsilon_q^2)$. In practice, a vector of squared residuals is used to estimate the error variance. By regressing the squared residuals against the predictors in **X**, the fitted values of the resultant variance vector, φ, establish the *n* design point weights, w_q. As the weights as inversely proportional to the error variance with the value of the weight, the

relationship is defined as $w_q = \frac{1}{\hat{\varphi}_q^2}$. Articulating the weights in this manner reduces the effects of high variability exerted by large residuals. If the error variance is large, then the associated weight would be comparatively small. Utilizing this method, the *WLS* estimator can be written as $\widetilde{\boldsymbol{\beta}}_q = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$. To ensure minimal model error, the procedure introduced by Goethals and Cho (2011) uses an iterative approach to reweight the model parameters-using subsequent estimation of the error variance. For this method, the algorithm either stops once either convergence is achieved or when the difference between the standard error for each of the estimated coefficients in $\boldsymbol{\beta}_w$is quite small relative to the standard errors obtained in the previous iteration.

The remaining regression alternatives listed in Table 2.3, alternatives 5-9, consist of variants of robust regression methods. These methods were developed to address the disproportional influence of outliers on the response surface functions. The term "robust" describes an estimator's ability to overcome an outlier's leverage on the generated estimate. Often, outlying responses are classified as anomalies or annotated as potentially contaminated. In the healthcare field, it is often hard to conclusively classify an outlier as an exception or an indicator of great importance. Robust estimators are particularly intriguing because they mitigate the leverage of extreme observations without discounting them altogether. The underlying methods for robust regression approaches are designed so that violations of supporting assumptions have little impact on the regression results. The four robust regression approaches analyzed for this study are *S*, *LAD*, *M-M*, and *M* estimations.

It could be argued that the simplest alternative for estimating robust regression coefficients the *LAD*. The introduction of the *OLS* method supplanted its use and interest in the method waned. A resurgence of in the use of *LAD* started when Karst (1958) suggested its suitability for use with data sets containing outliers in comparison to *OLS*. The *LAD* method minimizes the sum of the absolute values of the residuals, or errors between points generated by the regression function and corresponding data points:

$$\underset{\hat{\beta}}{\text{Minimize}} \sum_{q=1}^{n} \left|\varepsilon_q\right| \tag{9}$$

Although the *LAD* method proved more robust than *OLS*, significant outliers can still influence the resultant model. This particular shortcoming has motivated increased research in the search for more robust approaches.

One of the earlier efforts was made by Huber (1973) who introduced the *M*-estimation for regression. Mathematically, the method focuses on the residuals and takes the following form:

$$\underset{\hat{\beta}}{\text{Minimize}} \sum_{q=1}^{n} \rho\left(\frac{\varepsilon_q}{s}\right) + n\log s \tag{10}$$

where ρ = some symmetric function with a unique minimum at 0. If we presume s is known and set $\Psi = \rho'$ then the maximum likelihood estimator of the regression coefficients β solves the non-linear system of equations $\sum_{q=1}^{n} \boldsymbol{X}_q \Psi\left(\frac{\varepsilon_q}{s}\right) = 0$, where ψ represents Huber's bounded monotone ψ function. After some modification, this becomes $\sum_{q=1}^{n} \boldsymbol{X}_q \Psi\left(\frac{\varepsilon_q}{s}\right) = (n-p)\gamma$, where γ is selected for consistency at normality and the embedded tier-one estimates for location and scale are obtained using Huber's

Proposal 2 estimators, which result from solving the following equations simultaneously for μ and σ:[50]

$$\sum_{q=1}^{n}\psi\left(\frac{y_q-\mu}{\sigma}\right)=0 \quad \text{and} \quad \sum_{q=1}^{n}\psi^2\left(\frac{y_q-\mu}{\sigma}\right)=\eta$$

This method has proven to be a viable and efficient estimator that is robust to outliers in the response variable. However, it was also found to lack resistivity to outliers.

Rousseeuw (1984) proposed the least trimmed squares (*LTS*) method to overcome efficiency shortcomings with a previous method (least median of squares (*LMS*)). The objective in this approach involves minimizing the sum of squared residuals over a subset, *q*, of the complete set of *n* points:

$$\underset{\hat{\beta}}{\text{Minimize}} \sum_{q=1}^{q} \left|\varepsilon_q\right|^2_{q:n} \tag{11}$$

In short, the residuals are squared and then sorted in ascending order. Of the *n* residuals in the full set, the $(n - q)$ largest are "trimmed" so that only the residuals from the remaining *q* points are included in the regression. Thus, the $(n - q)$ largest points which are not used do not influence the fit. The result is a fit that retains the resistivity properties of the *LMS* method, and it is known to be more efficient.

Both the *LMS* and *LTS* methods involve the minimization of a robust measure of the scatter of the residuals. Introduced by Rousseeuw and Yohai (1984) as a means for performing robust regression in time series analysis, these methods find a plane or hyperplane that minimizes the scale *s* by obtaining the solution to:

$$\sum_{q=1}^{n} \chi\left(\frac{y_q - \mathbf{x}_q\hat{\boldsymbol{\beta}}}{c_0 s}\right) = (n-p)\varphi \tag{12}$$

In this context, p corresponds to the k-1 predictors and χ is typically denoted by the integral of Tukey's bisquare function given by

$$\chi(u) = \begin{cases} u^6 - 3u^4 + 3u^2 & |u| \le 1 \\ 1 & |u| > 1 \end{cases},$$

and c_0=1.548 and φ = 0.5 are selected for consistency at the normal distribution. This method is highly resistant to leverage points, robust to outliers in the response, and is often more efficient than the *LTS* method.

Yohai (1987) proposed the *M*-*M* estimator as an improved alternative that essentially blended earlier methods in order to retain the robustness, while gaining the efficiency of *M*-estimation. The *M*-*M* method proceeds in three stages. The first involves an initial estimation of regression coefficients. In the second, a highly robust and resistant *S*-estimate is computed that minimizes an *M*-estimate of the scale of the residuals. In the final stage, the estimated scale is then held constant, while a nearby *M*-estimate of the regression coefficients is determined.

4.4 Integrating the Estimators into the RPD Framework

Pursuant to the selection of a regression estimation approach based upon inherent process conditions, fitted response surface functions are then developed for the process location and scale. For the purposes of comparison, this is done for each of the alternative regression methods outlined in Section 2.3.3, using full second-order polynomial model.

Thus, the general form of the estimated response functions with k –1 predictor variables is expressed as:

$$\text{Location: } \hat{\mu}(\boldsymbol{x}) = \hat{\beta}_{\mu,0} + \boldsymbol{X}^T\hat{\boldsymbol{b}}_{\mu} + \boldsymbol{X}^T\hat{\boldsymbol{B}}_{\mu}\boldsymbol{X} + \varepsilon_{\mu} \quad (13)$$

$$\text{Scale: } \hat{s}(\boldsymbol{x}) = \hat{\beta}_{\sigma,0} + \boldsymbol{X}^T\hat{\boldsymbol{b}}_{\sigma} + \boldsymbol{X}^T\hat{\boldsymbol{B}}_{\sigma}\boldsymbol{X} + \varepsilon_{\sigma} \quad (14)$$

where $\underset{\mathbf{1\times k}}{\mathbf{X}^{\mathrm{T}}} = \begin{bmatrix} X_1 & X_2 & \cdots & X_{k-1} \end{bmatrix}$, $\underset{\mathbf{k\times 1}}{\hat{\mathbf{b}}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$, and $\underset{\mathbf{k\times k}}{\hat{\mathbf{B}}} = \begin{bmatrix} \beta_{11} & \beta_{12}/2 & \cdots & \beta_{1k}/2 \\ & \beta_{22} & \cdots & \beta_{2k}/2 \\ & & \ddots & \vdots \\ \textit{sym.} & & & \beta_{kk} \end{bmatrix}$,

where $\hat{\beta}_{\mu,0}$(and $\hat{\beta}_{\sigma,0}$), $\hat{b}_{\mu}$(and $\hat{b}_{\sigma}$), and $\hat{\boldsymbol{B}}_{\mu,0}$(and $\hat{\boldsymbol{B}}_{\sigma,0}$), reflect the estimates of the intercept, first-order, and second-order coefficients of the response surface functions for the location and scale, respectively. In addition, the terms ε_{μ} and ε_{σ} correspond to the residual error for the location and scale deviation, respectively. In order to investigate the performance of the fitted functions at estimating a response, a mean squared error (*MSE)*-based optimization scheme is used on either a spherical region of interest such that $\boldsymbol{x}'\boldsymbol{x} \leq \rho$, where ρ is the radius of the sphere, or a cubic region bounded by (-1, 1). Using this approach as a framework, each of the models delineated in Table 2.3 are evaluated on the *RPD* solution they produce.

4.4 Numerical Demonstration via Simulation

In this section, we examine a case using commonly-applied experimental data sets as bases for Monte Carlo simulation. The overarching purpose of the simulation is to determine the degree to which underlying process asymmetry and variability (and, consequently, non-constant variance) affect estimator performance in the context of *RPD* solutions, which should ultimately serve as a guideline for engineers and healthcare

professionals as to which estimators tend to perform best under a particular set of conditions.

This numerical study involves approximately normally-distributed data with moderately low variability in which all of the base assumptions concerning the data hold. An initial examination is intended to verify expectations regarding conventional approaches when assumed conditions hold. Subsequently, through experimentation, the impacts of increasing variability on estimator performance are examined. We examine four scenarios derived from combinations of high/low asymmetry with high/low variability to determine the effects on estimator selection. Within each case study, initial results are obtained from the base data and observations are then drawn to assess estimator performance. Thereafter, 1,000 iterations of each simulation scenario are conducted to facilitate performance trend analysis and assessments regarding estimator performance under the evaluated conditions. Simulations were developed in the statistical computing environment R version 2.14.1 (2012) which is open source software. For the purpose of estimator comparison, each simulation involves several key settings that are applied to each estimation model:

1) Using the actual experimental data, simulated data are derived using the skew normal approach.
2) Full second-order response surface functions are developed for the location (mean or median) and scale (standard deviation or *MAD*) response surface functions.

3) Optimization results are obtained for each estimation model using the *MSE*-based optimization scheme developed by Cho (1994) and Lin and Tu (1995).

Pursuant to (3) above, estimation approaches are then evaluated based on the optimal solutions they generate in terms of deviation from the established process target and variability in the result.

4.4.1 Case Study: Investigating the Effects of Variability on Estimator Performance

In this experiment, adapted from Phillips *et al*. (1995) and Shin *et al.*(2011), normally-distributed *n*-type quality characteristic is of interest. The control factors, X_1, X_2, and X_3 are known to influence the outcome Y with the desired target value $\tau = 57.5$. Using the original data obtained from Shin *et al.*(2011) and the procedure delineated in Section 2.3.2, five replicates were generated at each design point. The experimental framework displayed in Table 2.5 is a central composite design (*CCD)* comprised of eight factorial points, six axial points, and six center points, with the calculations for the mean, standard deviation, and skewness at each design point.

Table 4.4. Experimental framework for Case Study A

Run	Coded Units: Cut Speed X_1	Cut Feed X_2	Cut Depth X_3	Observed Responses (simulated) Y_1	Y_2	Y_3	Y_4	Y_5	Metal Removal Rate (mm^3/min) $\bar{Y}$	s	γ
1	-1	-1	-1	44.6	52.5	57.4	52.4	57.8	53.2	3.82	0.19
2	1	-1	-1	63.9	60.3	64.7	65.8	67.5	62.9	3.51	-0.77
3	-1	1	-1	45.6	51.5	45.4	62.0	52.8	53.4	3.67	-0.24
4	1	1	-1	67.1	64.5	61.6	58.6	55.5	62.6	3.24	-0.32
5	-1	-1	1	59.4	55.6	51.4	57.7	59.5	57.3	3.10	-0.11
6	1	-1	1	67.6	64.6	64.3	71.8	67.4	67.9	4.31	-0.21
7	-1	1	1	65.5	60.8	60.5	57.2	55.6	59.8	4.47	0.46
8	1	1	1	67.4	66.5	71.8	68.2	72.0	67.8	3.21	-0.85
9	-1.682	0	0	58.2	56.1	61.3	65.0	47.3	59.1	4.73	-1.13
10	1.682	0	0	69.5	63.2	59.3	73.0	61.0	65.9	4.46	0.73
11	0	-1.682	0	63.2	60.4	59.0	61.0	65.8	60	3.55	-0.16
12	0	1.682	0	59.5	62.6	61.7	57.3	59.9	60.7	3.10	-0.17
13	0	0	-1.682	51.7	66.3	57.2	61.9	64.4	57.4	4.29	-1.13
14	0	0	1.682	65.3	66.1	61.4	72.5	64.2	63.2	5.04	1.32
15	0	0	0	60.3	56.5	64.1	61.1	60.5	59.2	3.87	-0.03
16	0	0	0	59.2	66.9	56.7	62.7	57.8	60.4	3.74	-1.18
17	0	0	0	58.5	59.0	61.2	56.4	57.2	59.1	3.95	-0.08
18	0	0	0	62.4	53.0	59.6	64.0	56.6	60.6	3.71	0.22
19	0	0	0	64.8	63.3	60.9	54.9	66.3	60.8	4.00	0.64
20	0	0	0	53.4	60.5	60.9	64.7	59.9	58.9	3.92	-0.51

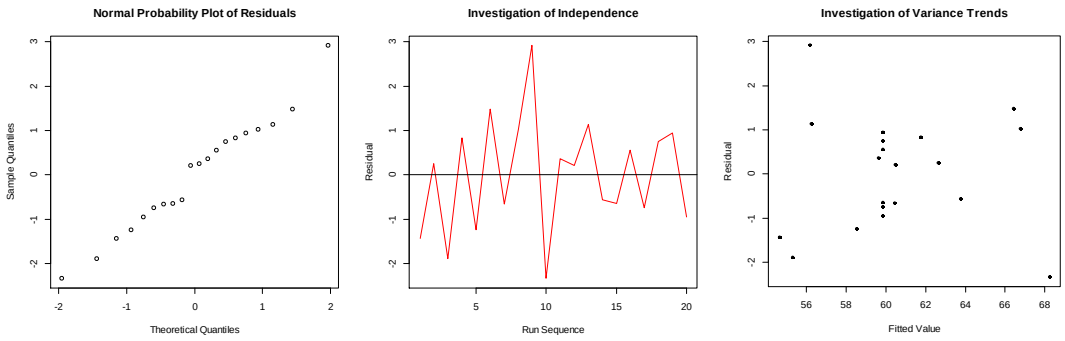
4.4.1.1 Preliminary Data Analysis

The initial graphical analysis of the responses suggests approximate symmetry and moderately low variability exist. Applying the Shapiro-Wilk test yields W^* =0.958 vs. W_α ≈0.951, and since $W > W_\alpha$, insufficient evidence exists to reject normality. Notwithstanding, the non-zero values in the γ column of Table 2.5 coupled with the few responses in Figure 2.3(a) that deviate from the reference line suggest that some asymmetry is, in fact, present. In terms of process variability, Figure 4.3(b) shows the deviations between the observations and the mean response to be quite small, and well within the 3σ threshold defined in Section 4.3.1.

(a) (b)

Figure 4.2. Assessing a) normality and b) variability in the responses.

After performing a preliminary regression for a full second-order model for the mean response $\bar{Y}$ using the OLS approach, a graphical analysis of the residuals (Figure 2.4) suggests that the assumptions of normality and independence hold, but that non-constant variance may exist. Yet, application of the Breusch-Pagan (B-P) hypothesis test yields χ^2_{BP}=16.1 < χ^2 (.95,9)=16.9 which suggests constant variance and thus disputes the deduction suggested by the plot in Figure 4(c). It is often the case with smaller sample sizes that the objective test results fail to capture the presence of non-constant variance.



(a) (b) (c)

Figure 4.3. Investigation of a) normality, b) independence, and c) variance in the residuals.

After performing a second-order regression for all 100 experimental observations, reiterating the Breusch-Pagan test yields $\chi^2_{BP} = 11.9$, which is clearly less than χ^2(.95,9)=.1691 and reinforces the initial test results. The revised residual plots shown in

Figure 5 are based on the full complement of 100 observations and illustrate the validity of the basic residual assumptions in this case.

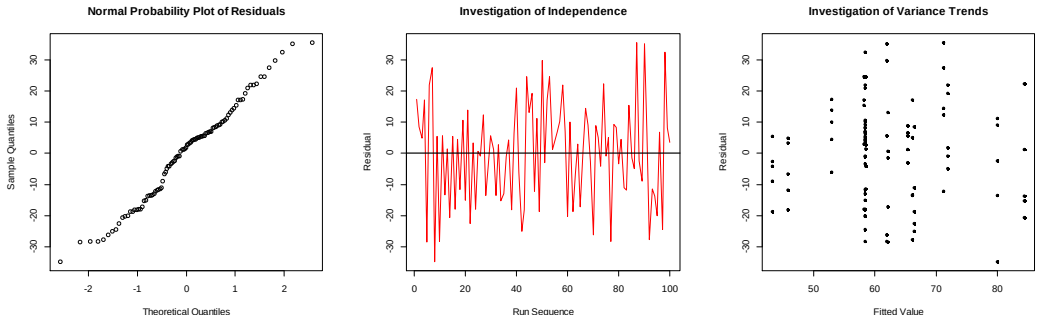


Figure 4.4. Residual analysis based on all 100 observations in the metal cutting study.

Taken together, the results of the data analysis suggest that the experimental data meet all requisite provisions for the application of *OLS* regression. This implies that *OLS* would be the best approach given that this method is known to produce the best linear unbiased estimates for the process location and scale, or dispersion, when these conditions hold.

4.4.1.2 Results Based on Original Experimental Data

We first performed a single run of the experiment to motivate the discussion on conditions-based estimator selection. Additionally, the run demonstrates the benefits of using the skew normal distribution to model system properties. The *OLS* method using traditional tier-one estimators under the assumption of zero skewness is also applied. The results in Figure 2.6 show the optimal operating conditions $\mathbf{x}^* = (x_1, x_2, x_3)$ obtained under each regression model using the *MSE*-based optimization scheme, the associated optimal process mean and standard deviation, and the resulting target bias and *MSE*. For the *GLM* approach, the Gaussian-identity (default) distribution-link combination was

used, which essentially mirrors the *OLS* counterpart and is appropriate when traditional assumptions hold. From the results in Table 4.6, a single run of the experiment suggests two things: first, accounting for even low degrees of asymmetry can produce better *RPD* solutions than the traditional approach to *OLS* estimation; and second, *OLS* regression (under the *SN* approach) is still suitable, although the use of the median-based *WLS* method can achieve superior results. However, recognizing that these solutions are estimates, it is therefore quite likely that subsequent implementations of the experiment could yield different sets of optimal coordinates.

Table 4.5. Regression and optimization results of a single run with five simulated observations.

	OLS (Traditional)	*OLS (SN)*	*WLS Mean/s*	*WLS Median/MAD*	*LTS*	*S*	*LAD*	*MM*	*Huber Prop 2*	*GLM*
x_1	-0.380	-1.682	-0.071	-0.515	0.056	0.023	-0.215	0.101	-1.682	-1.682
x_2	-1.682	-1.682	1.682	-1.682	-1.682	-1.682	1.682	-1.682	-1.682	-1.682
x_3	-0.170	1.014	-1.682	1.682	-1.682	-1.682	-1.682	-1.682	1.155	1.014
$\hat{\mu}(\mathbf{x}^*)$	57.172	57.615	58.825	57.540	57.466	57.490	58.430	57.459	57.586	57.615
bias	0.328	0.115	1.325	0.040	0.034	0.010	0.930	0.041	0.086	0.115
$\hat{\sigma}(\mathbf{x}^*)$	2.649	2.086	2.733	0.449	2.369	2.169	2.563	2.928	2.064	2.086
MSE	7.127	4.365	9.225	0.203	5.612	4.703	7.431	8.575	4.266	4.365

Moreover, the objective is to examine trends to develop a better sense of how the estimators perform on average, which cannot be achieved via a single run. Accordingly, we conducted 1,000 iterations of the simulation, generating fresh random data based on the original experiment at each iterate. At the end of each simulation run, the *MSE* and target bias were recorded for the optimal *RPD* solutions and then averaged across all iterations to observe trends. Table 4.7 contains the simulation results, along with the proportion of iterations in which a particular estimation approach yielded the smallest *MSE* and bias.

Table 4.6. Simulation results under low variability conditions.

	OLS (Traditional)	*OLS (SN)*	*WLS Mean/s*	*WLS Median/MAD*	*LTS*	*S*	*LAD*	*MM*	*Huber Prop 2*	*GLM*
Avg MSE	4.029	3.250	3.349	2.379	2.636	2.593	3.359	3.205	3.292	3.250
% Best MSE	4.90%	3.50 %	5.10%	30.30%	20.70 %	20.50 %	9.20 %	5.50 %	1.90 %	3.50 %
Avg Bias	0.300	0.266	0.260	0.243	0.205	0.200	0.220	0.220	0.259	0.266
% Best Bias	7.20%	4.80 %	6.70%	19.50%	23.10 %	19.90 %	9.70 %	7.70 %	3.90 %	4.80 %

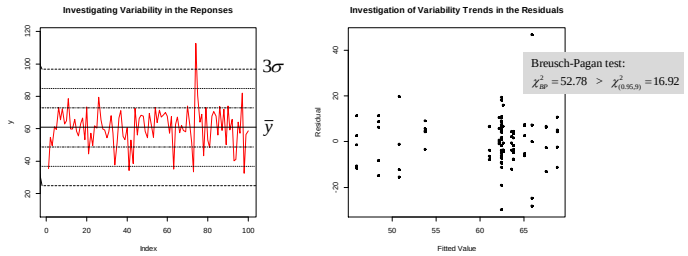
Noting that all nine of the alternative estimation approaches outperformed the traditional *OLS* approach in Table 4.7, it is clear that despite approximate symmetry/normality in the process data, there is enough inherent skewness to affect the optimization results. In the most basic sense, this is illustrated by comparing the first two columns in Table 4.7 (*OLS*-Traditional vs. *OLS-SN*), which suggests that by accounting for even slight levels of non-zero skewness, better *RPD* results can be obtained. Beyond this, the fact that median-based approaches (*WLS, MAD, LTS,* and *S* estimations) yielded better results in terms of both average performance and consistency suggests that these methods are preferable when any degree of asymmetry exists. The main reason that the *WLS* procedure produced better results on average is most likely from the result of down-weighting those observations with higher variability, thus demonstrating the viability of using that method to exert greater control over sources of process variation.

4.4.1.3 Investigating the Effects of High Variability Conditions

To examine the effects of variability, we incorporated a simple modification to the simulation that would induce a greater degree of variability in the process. Whereas before we used the sample standard deviation(*s*) in Table 4.5 to generate normal random variates in the base scenario, in this instance we randomly sampled an integer from a

range of 2 to 5 at each design point to serve as a factor that would then be multiplied by the original values for *s*. Hence, the variability at each design point would be increased by a factor of anywhere from 2 to 5 times.

The idea here is that simply multiplying the *s* vector by a single common factor would not have any impact on the results other than to scale them by that factor. That is, there would certainly be more variability in the responses, but the proportional change in each design point would be the same and would negate any real effects on the results as the underlying conditions regarding base assumptions would still hold. Thus, our objective is to inject variability not only horizontally within each design point, but also vertically across the vector of sample standard deviations. This would challenge system performance and very likely upend the underlying assumptions of response variability and heteroscedasticity. As the plots in Figure 4.6 show, this is precisely what occurs, as several observations exceed the 3σ threshold (Figure 4.6a), and the variability trends coupled with the Breusch-Pagan results in Figure 4.6b clearly suggest non-constant variance in the residuals. As noted previously, the presence of such conditions inhibits the use of *OLS* and suggests the need for either remedial measures or alternative estimation approaches.



(a) (b)

Figure 4.5. Analysis of responses (a) and residuals (b) under high-variability conditions.

To establish performance trends among the estimators, we performed 1,000 iterations of the high-variability scenario. In this instance, the *GLM* approach is modified to account for differences in the data. Specifically, further analysis of the residual data for response surface functions for both the mean and variation suggested the need to consider either a gamma or inverse Gaussian distribution to correct for non-constant variance. After preliminary modeling using the various distribution- link function combinations in the *GLM* approach, it is determined that the gamma-identity and inverse Gaussian-log combinations would produce the best fit for the mean and standard deviation response surface functions, respectively. Results for the MSE and bias were then averaged across all 1,000 iterations, and performance proportions were calculated to produce the results shown in Table 4.7.

Table 4.7. Simulation results under high variability conditions.

	OLS (Traditional)	*OLS (SN)*	*WLS Mean/s*	*WLS Median/MAD*	*LTS*	*S*	*LAD*	*MM*	*Huber Prop 2*	*GLM*
Avg MSE	35.604	30.398	28.261	20.194	19.714	20.277	32.284	28.514	31.253	3.643
% Best MSE	4.90%	3.70%	7.30%	18.30%	17.40%	16.10%	8.60%	5.60%	3.40%	16.00%
Avg Bias	2.225	2.079	2.087	1.790	1.579	1.694	2.089	2.037	2.112	0.166
% Best Bias	5.20%	2.80%	6.40%	15.10%	15.90%	14.40%	8.10%	4.90%	3.50%	24.80%

The results in Table 4.7 draw several insights. First, all nine of the alternatives once again produced a better result than the traditional *OLS* method, reinforcing the benefit of using the skew normal approach for modeling process asymmetry. Second, it is clear that the increased variation induces a change in estimator performance such that the *GLM* method using the gamma-identity and inverse Gaussian-log combinations

outperforms all others on average, both in terms of the resulting *MSE* and target bias. While the next-best performers (*LTS, WLS* (median/*MAD*), and *S*-estimation methods) performed relatively well, they all achieved an average *MSE* nearly six times larger than the *GLM* method.

Although differences in the generated data can be a contributing factor, the reasons behind these results can also be attributed to the increased likelihood of extreme observations in either tail. And if an extreme observation from one tail is not counter-balanced by an extreme point from the other, then the resulting sample could very well appear skewed, despite being generated from a normal distribution. Obviously, when the data are approximately normal, the mean and the median will have nearly the same value. However, as the data become skewed due to the occasion of one or more extreme observations, mean-based estimators deteriorate in their ability to provide the best estimate of central tendency due to the influenced of outlying data points. Similarly, the standard deviation no longer provides the best measure of the true dispersion in the distribution. The median and the *MAD*, on the other hand, retain their properties and are resistant to extreme observations, thereby making them preferable when such conditions exist. In addition to this, high variability typically will also induce non-constant variance, which invariably results in suboptimal solutions if *OLS* regression is applied. As the plots in Table 2.8 show, this is precisely what is occurring in this scenario, and serves to explain why the robust and *GLM* approaches perform well.

4.5 Summary of Findings

The numerical results in several key insights for solving the *RPD* problem in asymmetric and highly variable conditions, which are summarized in (i)-(iii) below:

(i) Most importantly, the simulation results across all scenarios clearly demonstrate that as process variability increases, alternative approaches to the traditional *OLS* method are not only necessary, but preferable. When coupled with asymmetric conditions, the effects become even more pronounced, particularly when the levels of both conditions are high. The key question is why. As previously discussed, once elevated degrees of variability and inherent asymmetry shift the data from assumed normality, the performance of traditional approaches to estimation suffers as a result of the influence exerted by extreme observations from the long tail of the skewed distribution. The alternative methods examined (namely the *GLM, S-, LTS,* and *WLS* (median/*MAD*)) tend to overcome those influences most effectively. As the results have shown, the *GLM* approach tended to perform very well, if not best, in all of the examined scenarios. But it is important to recognize that this is predicated on the identification of the right distribution-link combination, which is data-dependent and so constitutes another required step in the application of that particular method. However, viable alternatives to this are the *WLS* (median/*MAD), LTS*, and *S*-estimation methods, which also performed markedly better than traditional *OLS* and *WLS* approaches in high variability and high asymmetry-high variability situations. Thus, in view of the aims of this paper, the pressing question for healthcare professionals is which approach to use and when. Based

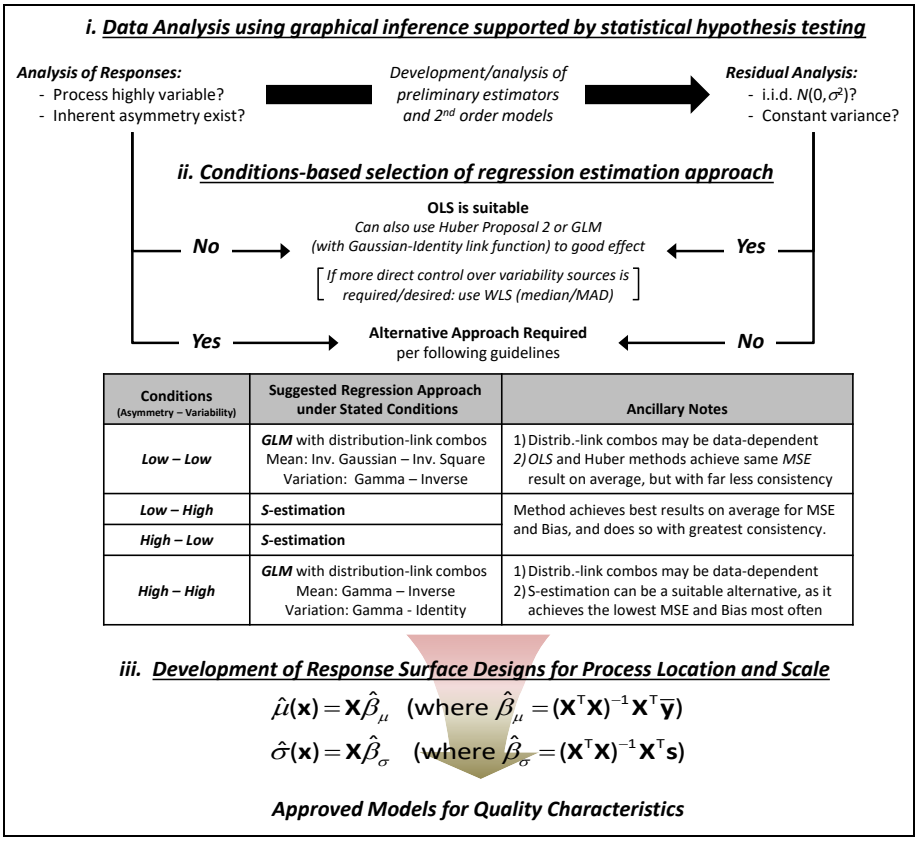
on the analysis of the presented results, the answer is depicted in Figure 4.7, which shows the modification to Phase 1b of the original process map from Figure 4.1.

Two additional points should be made. First, some might suggest that high variability should not pose an issue, as it could be overcome by simply increasing the sample size required for estimation. Added replication at each experimental design point could ameliorate potential issues and would be preferred. However, this is often not feasible due to time and cost constraints, as well as other limitations on resources required for experimentation. Second, the results obtained in the numerical example show performance trends rather than definitive conclusions as to the certainty of one estimator's performance versus another's. What they demonstrate is that when elevated degrees of process variability and asymmetry exist, estimator selection matters in terms of achieving the best *RPD* solutions. This echoes the importance of a detailed analysis in the early stages of experimentation to ascertain the degree to which such underlying conditions exist in the data, which in turn will influence the selection of the most appropriate estimation approach to use for response surface modeling and optimization.

(ii) The use of the skew normal distribution facilitates more accurate modeling of the inherent distributional properties associated with a particular set of data and, based on the results of the numerical example, can produce better *RPD* results in terms of minimal bias and variability. Most notably, because normality is a special case, the skew normal can very easily capture both symmetric and asymmetric properties, thereby accounting for the presence of either in process outputs. Thus, by using the first three moments to replicate experimental observations in the numerical example, we were able to more

accurately portray process characteristics. This is important, as elevated variability and asymmetry are, in reality, quite probable in many healthcare and medical processes. Hence, the use of the skew normal distribution provides the capability to model either situation simultaneously, thereby allowing for a more accurate accounting of innate system properties.

(iii) The ease and explicitness associated with the *OLS* approach has helped to solidify its position as the basis for regression estimation for more than two hundred years; and it continues to see the preponderance of use throughout the literature and in applied statistics texts. Moreover, what tend to steer engineers away from considering realistic process conditions (i.e., asymmetry) and many alternative estimation methods are the computational complexities associated with them. But with today's high-speed computing power and myriad readily-available software platforms such as R, the computational complexity of alternative estimation methods should no longer be avoided. As our results show, these methods can make a significant difference in the quality of the results achieved when certain conditions exist. But the reality is that these conditions actually exist more in practice than otherwise; and when they do, the necessary assumptions that underpin *OLS* regression no longer hold. If used in spite of this reality, the *OLS* method may likely yield suboptimal solutions.



Conditions (Asymmetry – Variability)	Suggested Regression Approach under Stated Conditions	Ancillary Notes
Low – Low	***GLM*** with distribution-link combos Mean: Inv. Gaussian – Inv. Square Variation: Gamma – Inverse	1) Distrib.-link combos may be data-dependent *2) OLS* and Huber methods achieve same *MSE* result on average, but with far less consistency
Low – High	***S*-estimation**	Method achieves best results on average for MSE and Bias, and does so with greatest consistency.
High – Low	***S*-estimation**	
High – High	***GLM*** with distribution-link combos Mean: Gamma – Inverse Variation: Gamma - Identity	1) Distrib.-link combos may be data-dependent 2) S-estimation can be a suitable alternative, as it achieves the lowest MSE and Bias most often

Figure 4.6. Conditions-based selection guidelines for regression estimation in asymmetric and/or high-variability process conditions.

4.6 Conclusion

High variability and asymmetry are conditions that occur quite often across a broad range of healthcare applications and it is believed that it should be given special consideration in the experimental process regarding the selection of appropriate approaches to response surface estimation. To that end, the focus of this paper has been to examine various alternatives to *OLS* regression in the *RPD* framework when such conditions prevail. The results and analysis demonstrate that, as process conditions evolve (i.e., variability and/or asymmetry increase), the estimator selection process should evolve, as well, to achieve the best solutions possible. In particular, the results

have shown that the *GLM*, *S*-, *LTS*, and median-based *WLS* methods tend to yield better *RPD* solutions. While such methods are fairly well-known in statistical circles, their use by healthcare professionals in robust parameter design applications is comparatively rare, as noted by Haenkamp *et. al* (2009):

> "The majority of past [robust parameter design] research has traditionally been carried out by statisticians targeting an audience with good insights on statistics. When, instead, targeting engineers with less statistical knowledge as the major audience, clearly other demands are put on guidelines and tools."

One such demand is a clearer understanding of which tools to use and when. The methodology and analysis offered in this paper should help to answer this need by providing healthcare professionals with some clarification as to which estimation approaches will tend to provide the best *RPD* solution when certain conditions exist. The analysis in this paper is based upon controlled experimentation, the replication of observations made on a specified quality characteristic of interest under highly variable and asymmetric conditions, and the implementation of the skew normal distribution to effectively model both symmetric and asymmetric instances. Future research may expand the investigation to include additional conditions, as well as processes that involve multiple quality characteristics of interest. Furthermore, the development of skew normal-based link functions for use with *GLM* approaches would also add benefit. In the end, proper accounting for the inherent conditions in the data will allow healthcare professionals to more accurately model the processes they endeavor to optimize, which

will invariably translate to better *RPD* solutions and more reliable recommendations to decision makers.

APPENDICES

A. Summary Table of Stated Healthcare Concerns and Associated Research Focus

APA Reference	Healthcare Concern	Research Focus
Acciaroli et al., 2018	How to use glycemic variability indices to classify subjects remains controversial	[1] Assess feasibility of using a glycemic variability index to distinguish between healthy individuals and those with impaired glucose tolerance (IGT) or T2DM [2]Assess feasibility of using a glycemic variability index to distinguish between individuals with IGT versus T2DM
Agarwal et al., 2016	[1] ML approaches to electronic phenotyping are limited by the scarcity of training datasets [2] Manual creation of training sets for ML approaches is time intensive	Investigate an alternative method to manual labeling to create training sets for statistical models of phenotypes
Allalou et al., 2016	[1] Gestational diabetes (GDM) affects 3-14% of pregnancies [2] 20-50% of women with GDM will develop T2DM within 5 years [3] Prediction of progression from GDM to T2DM critical for individual risk stratification	Develop a metabolomics signature to predict patient progression from gestational diabetes mellitus to T2DM
Anderson et al., 2016	25% of T2DM are undiagnosed due to inadequate screening	Assess whether electronic health record phenotyping could improve T2DM screening compared to conventional models
Anderson et al., 2016	Providing more detailed insights on factors that drive progression to DM would be valuable in characterizing and intervening on at-risk patients	Develop a prediction model ensemble for progress to prediabetes or T2DM using variables found within electronic health records
Basu et al., 2017	It is unclear how to best individualize glycemic targets	Identify characteristics of patients at high cardiovascular risk with decreased or increased mortality risk from glycemic therapy
Basu et al., 2017	There exists substantial mis-estimation of risks of diabetes complications using existing equations (RECODe)	Develop updated risk equations for complications of T2DM (RECODe)

APA Reference	Healthcare Concern	Research Focus
Cao et al., 2017	Molecular basis for the comorbidity of schizophrenia and T2DM is not completely understood	Determine molecular commonality between schizophrenia and glycemic markets of T2DM (identify a polygenic schizophrenia signature and explore its impact on T2DM
Casanova et al., 2016	Prediction for incident diabetes is based on limited variables	[1] Investigate relative performance of machine learning method such as RF for detecting diabetes in a high dimensional setting [2] Uncover potential predictors of diabetes
Chen et al., 2015	Performance of the recommended glomerular filtration rate (GFR)-estimating equations in T2DM population is inferior to the nondiabetic population; important for drug dosing	Develop new GFR-predicting models for use in Chinese patients with T2DM
Dagliati et al., 2017	ML algorithms can be embedded into data mining pipelines to extract knowledge from data	Predict the onset of retinopathy, neuropathy, or nephropathy at different time scenarios
Dong et al., 2017	Missing heritability is still a big problem for Genome-wide association studies ; susceptibility loci identified by GWAS only account for a limited proportion of the observed heritability of diseases	Development of more powerful methods to predict novel risk SNPs from the large amount of SNP data and regulatory features
EITanboly et al., 2017	Detecting early retinal changes in T2DM to give patients a chance to delay further complications is absent so far	Diagnosis of diabetic retinopathy (DR) using optical coherence tomography (OCT) images for T2DM patients
Farran et al., 2013	Efficient preventative strategies are needed to control risk factors for T2DM; use knowledge to on individual population or at population level to identify groups of high-risk patients	Build classification models and risk assessment tools for diabetes, hypertension and comorbidity using ML algorithms on data from Kuwait
Han et al., 2017	Need a system or process to stratify individuals according to disease risk for clinical disease prevention	Develop a risk stratification model of clinical disease to be used for interventions

APA Reference	Healthcare Concern	Research Focus
Hertroijs et al., 2017	Iplementation of precision medicine based solely on genomics has proven to be difficult for certain diseases; phenotyping approach to precision medicine is only sparsely adopted in evidence-based guidelines for diabetes treatment	[1] Identify subgroups of people with newly diagnosed T2DM with distinct glycemic trajectories [2] Predict trajectory membership using patient characteristics [3] Validate findings in different cohort of patients with T2DM
Kagawa et al., 2017	Existing phenotyping algorithms are not sufficiently accurate for screening and identifying clinical research subjects	Distinguish T2DM patients based on electronic health records; propose new metric to evaluate practicality of algorithms
Kim et al., 2017	Need to find markers for end-stage renal disease; because diabetic patients are likely to develop ESRD it is imperative to discover which elements of diabetic patient's medical problems lead to ESRD	Discover frequently appearing medical complications at various levels of kidney functions for two different subpopulations defined by ethnicity
Lee et al., 2016	No study has assessed the predictive power of phenotypes based on individual anthropometric measurements	Assess the association between the HW phenotype and T2DM in Korean adults
Lee et al., 2014	Prediction of type 2 diabetes using a combination of anthropometric measures remains a controversial issue	Predict the fasting plasma glucose status that is used in the diagnosis of type 2 diabetes by a combination of various measures of Korean adults
Leung et al., 2013	Diabetic kidney disease is rising in parallel to the growing epidemic of T2DM / rapid advancement of molecular tech, large datasets containing many genotypes and phenotypes; challenge is in synthesizing discoveries and translating them to clinical practice	Explore computation tools with a comprehensive data base on T2DM
Li et al., 2016	Growing risk to patient population with the use of Electronic health records; need to reconcile the preservation of patient privacy and the need to have sufficient data for modeling and decision making	Develop two adaptive distributed privacy preserving algorithms based on a distributed ensemble strategy

APA Reference	Healthcare Concern	Research Focus
Lo-Ciganic et al., 2015	Little empirical evidence to support medication adherence threshold levels as a predictor of health outcomes	Apply ML to examine how adherence to oral hypoglycemic medications is associated by avoidance of hospitalization; identify adherence thresholds for optimal discrimination of hospitalization risk
Lopez et al., 2017	Need to definitively link single nucleotide polymorphisms to disease development	Identify relevant SNPs to T2DM and build a decision support-tool for risk prediction
McCoy et al., 2017	Individualized diabetes management would benefit from prospectively identifying well-controlled patients at risk of losing glycemic control	Identify patters of H1bA1c change among patients with stable controlled diabetes
Moreno et al., 2017	Develop noninvasive method to test for T2DM	Screen for the presence of T2DM by means of the signal obtained by a pulse oximeter
Neugebauer et al, 2013	Clinical trials are unlikely to be launched for many comparative effectiveness research questions	Adaptation of a data adaptive estimation approach called Super Learning avoids reliance on arbitrary parametric assumptions in CER
Neugebauer et al., 2016	Contribute to the understanding of potential consequences of the choice of estimation for propensity scores in real world comparative effectiveness analysis	Use EHR data to evaluate the effects of four adaptive treatment intensification strategies (bias from incorrect parametric model specification)
Olivera et al., 2017	T2DM is a chronic disease associated with a wide range of serious health complications	Develop and validate predicative models for detecting undiagnosed diabetes
Ozery-Flato et al., 2016	Development of a tool that could automatically evaluate a patient's response to treatment, identify patients who are most likely experiencing problems, and focus physicians' attention on those patients who require it most would be extremely valuable	New approach for detection and analyzing patients with expected responses to antidiabetic drugs
Ozery-Flato et al., 2013	Only limited information is available on the predictors ofT2DM in the group of patients already diagnosed with metabolic syndrome	Investigate the predictive value of different biomarkers for the incidence of T2DM in patients with metabolic syndrome

APA Reference	Healthcare Concern	Research Focus
Peddinti et al., 2017	Predictive biomarkers are needed to allow physicians to identify and monitor individuals at high risk for T2DM	Systematically evaluate the predictive power of comprehensive metabolomics profiles to predict T2DM
Pedersen et al., 2016	Not all patients undergoing weight-loss surgery experience diabetic remission, but the mechanistic insights that cause the heterogeneous therapy results are not understood.	Combine clinical and genomic factors using heuristic methods to identify patients who may have a low likelihood in responding to bariatric surgery for improved glycemic control
Pimentel et al., 2016	Current methods of treating T2DM are inadequate therefore it is important to focus on prevention of the disease	Propose a new approach for T2DM based on EHR without using invasive techniques
Pazavian et al., 2015	Interventions can only be cost effective when the target population has a high likelihood of developing diabetes at the baseline	Develop a population-level risk prediction model for type 2 diabetes that can be used with health insurance claims or other readily available data
Pamezankhani et al., 2016	Most ML classified work well when the class distribution is evenly distributed, but class imbalance is prevalent in medical datasets	Evaluate the impact of synthetic minority oversampling technique (SMOTE) on the performance of probabilistic neutral network, naïve Bayes (NB), and decision tree (DT) classifiers for predicting diabetes
Sudharsan et al., 2015	Minimizing the occurrence of hypoglycemia is a challenging task since T2DM patients typically check on 1-2 self-monitored blood glucose levels per day	Develop a probabilistic model to predict an hypoglycemic event within the next 24 hours
Vyas et al., 2016	The number of revealed protein-protein interactions is limited compared to the available protein sequences of different organisms	Develop a model for discriminating disease proteins from non-disease proteins for T2DM
Zheng et al., 2017	Existing expert based identification algorithms often have a low recall rate and could miss valuable samples	Propose a data informed framework for identifying subjects with and without T2DM from EHR via feature engineering and machine learning

B. Summary Table of Machine Learning Techniques and Selected Dataset

APA Reference	Machine Learning Algorithms	Model Performance Metrics	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Acciaroli et al., 2018	logistic regression	accuracy, F1 score, precision, recall		Finland	Botnia Study Group	
Agarwal et al., 2016	L1 penalized logistic regression	accuracy, against rule-based definitions, positive predictive value	XPRESS	United States	Standford Children's Health and Stanford Healthcare	January 1994 - June 2013
Allalou et al., 2016	decision tree, J48 decision tree, Naïve Bayes, logistic regression	accuracy, area under the curve, F-score precision , specificity, sensitivity	R-studio, Waikato Environmen t for Knowledge Analysis workbench	United States	Study of Women, Infant Feeding, and Type 2 Diabetes Mellitus After GDM Pregnancy (SWIFT)	2008-2014
Anderson et al., 2016	logistic regression, random forest	accuracy, area under the curve, positive predictive value, negative predictive values, sensitivity, specificity	R	United States	Practice Fusion Diabetes Classification	2009-2012
Anderson et al., 2016	Bayesian posterior	area under the curve	Reverse Engineering and Forward Simulation (REFS)	United States	Humedica Electronic Health Records	2007-2012

APA Reference	Machine Learning Algorithms	Model Performance Metrics	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Basu et al., 2017	gradient forest, decision tree		R	United States and Canada	Action to Control Cardiovascular Risk in Diabetes (ACCORD)	2001-2009
Basu et al., 2017	elastic net regularizatio n to inform the Cox Hazards Model		R	United States and Canada	[1] Action to Control Cardiovascular Risk in Diabetes (ACCORD) [2]Diabetes Prevention Program Outcomes Study [3] Action for Health in Diabetes	[1] 2001-2009 [2] 1996-2001 [3] 2001-2012
Cao et al., 2017		accuracy, Nagelkreke's R^2	R		[1] GEO database: GSE53987, GSE21138, GSE35977, GSE12679, GSE38642, GSE3489, GSE36980 [2] GWAS Data: GO, KEGG, Panther, Reactome, Target Scan [3] Expression Atlas: GTEx	

APA Reference	Machine Learning Algorithms	Model Performance Metrics	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Casanova et al., 2016	random forest, logistic regression	area under the curve, standard panel - accuracy, sensitivity, specificity		United States	Jackson Heart Study, University of Mississippi Medical Center	
Chen et al., 2015	artificial neural networks (back propagation)		MATLAB 2011A	China	Third Affiliated Hospital of Sun Yatsen University	
Dagliati et al., 2017	logistic regression, naïve Bayes, support vector machines, random forest	accuracy, area under the curve, matthew's correlation coefficient, negative predictive value, positive predictive value, sensitivity, specificity, area under the curve		Italy	[1] Istituto di Ricovero e Cura a Carattere Scientifico - Research hospital [2] Istituto Clinico Scientifico Maugeri (ICSM)	

APA Reference	Machine Learning Algorithms	Model Performance Metrics	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Dong et al., 2017	decision tree, class analogy, random forest, support vector machines	F1 score, number of features, sensitivity, specificity, accuracy	R		[1] 1000 Genome Project [2] ENCODE; Roadmap Epigenomics Project, expression quantitative trait loci data in T2DC relevant issues from the GTEx database [3] genomic regions from the GERP++	
EITanboly et al., 2017	deep learning (compared with K-Star, K-Nearest, Random Tree, Random Forest)	Accuracy, area under the curve, sensitivity, specificity	Waikato Environ-ment for Knowledge Analysis workbench			
Farran et al., 2013	logistic regression, *k*-nearest neighbors, support vector machines	classification accuracy	MATLAB - MATrix LABoratory	Kuwait	Kuwait Health Network (KHN)	(12 years)
Han et al., 2017	*k* means variants (base, PSC, Seeded, COP, PCK, MPCK, Supervised, Constrained, FSCL), random forest	ratio of minimum to expected, sensitivity, specificity, standard deviation in cluster sizes, Youden index, standard deviation in cluster sizes	R	China	Chinese Hospital Data	(7 years)

APA Reference	Machine Learning Algorithm	Model Performance Metric	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Hertroijs et al., 2017	*k*-nearest neighbor, Fisher, Parzen, quadratic discriminant classifier, linear discriminant classifier, support vector machine, logistic regression, stacked support vector machine	Akaike Information Criterion, Bayesian Information Criterion, Lo-Mendel-Rubin-likelihood ratio test, predicted and observed trajectory, calibration slopes, sensitivity, specificity, positive predictive values, negative predictive values		Nether-lands	[1] Zwolle Outpatient Diabetes Project Integrating Available Care (ZODIAC) [2] ZIO, a regional care group	[1] 2006-2013 [2] 2009-2013
Kagawa et al., 2017	Support vector machine, PheKB	sensitivity, area under the curve, specificity, sensitivity	R (kernlab, ROCR, caret)	Japan	University of Tokyo Hospital	2009-2014
Kim et al., 2017	apriori				2012 Cerner database	
Lee et al., 2016	naïve Bayes, logistic regression	area under the curve	SPSS 19, Waikato Environment for Knowledge Analysis data mining tool	Korea	Korean Health and Genome Epidemiology Study Database	Nov 2006 - August 2013
Lee et al., 2014	naïve Bayes, logistic regression	area under the curve, sensitivity, specificity	SPSS 19, Waikato Environment for Knowledge Analysis data mining tool	Korea		

APA Reference	Machine Learning Algorithms	Model Performance Metrics	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Leung et al., 2013	decision tree, random forest, naïve Bayes, neural networks, partial least squares regression, support vector machine	ROC, sensitivity, specificity	R for ML, but SPSS for statistical analysis	China	Hong Kong Diabetes Registry	1 July 1994 - 30 June 1998
Li et al., 2016	AdaBoost (local learner)	area under the curve, F measure, sensitivity, precision			Practice Fusion Diabetes Classification Challenge (2002)	
Lo-Ciganic et al., 2015	random survival forest		SAS 9.3 and R	United States	Pennsylvania Medicaid program	2007-2011
Lopez et al., 2017	classify: random forest, support vector machine, logistic; predict: regression, *k*-nearest neighbor	area under the curve, prediction accuracy			Biomedical Research of Girona	
McCoy et al., 2017	random forest		R		OptumLabs Data Warehouse	2001-2013
Moreno et al., 2017	random forest, gradient boosting, linear discriminant analysis	area under the curve, sensitivity, septicity	MATLAB	Spain	6 Clinics around Barcelona	2013
Neugebauer et al., 2013	super learning				EHR from patients of four sites of the HMO research network consortium	January 2006 - June 2008

APA Reference	Machine Learning Algorithms	Model Performance Metrics	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Neugebauer et al, 2016	super learning				real-world comparative effectiveness research	
Olivera et al., 2017	logistic regression, artificial neural network, naïve Bayes, *k*-nearest neighbor, and random forest	accuracy, area under the curve, balanced accuracy, sensitivity, specificity,	R	Brazil	Longitudinal Study of Adult Health (ELSA-Brazil)	2008-2010
Ozery-Flato et al., 2016	*k*-nearest neighbor, support vector machine					
Ozery-Flato et al., 2013	logistic regression	area under the curve	MATLAB	Lithuani-a	Lithuanian High Cardiovascular Risk primary prevention program	2007-2011
Peddinti et al., 2017						
Pedersen et al., 2016	artificial neural network	Area under the curve, accuracy, integrated discrimination improvement	R, Plink	United States	CardioMetabochip, eMERGE, GIANT, DIAGRAM, MAGIC	
Pimentel et al., 2016	regularized least squares	area under the curve	R	Finland	Botnia Prospective Study (BPS)	1990-2000
Pazavian et al., 2015	logistic regression	area under the curve, sensitivity, specificity, positive predictive value		United States	cohort study of beneficiaries of Independence Blue Cross	2005-2013

APA Reference	Machine Learning Algorithms	Model Performance Metrics	Machine Learning Software	Dataset Origin	Dataset Name or Source	Dataset Time Frame
Pamezankhani et al., 2016	probabilistic neural networks, decision trees, naïve Bayes	accuracy, F-measure, precision, sensitivity, specificity, precision, Youden's index	Konstanz Information Miner (KNIME)	Iran	Tehran Lipid and Glucose Study (TLGS)	(12 years)
Sudharsan et al., 2015	random forest, support vector machine, *k*-nearest neighbor, naïve Bayes	accuracy, sensitivity, specificity			De-identified patient data from a clinical trial of patients with T2DM	(1 year)
Vyas et al., 2016	support vector machine	accuracy, area under the curve, precision, accuracy, recall	LibSVM	United States		
Zheng et al., 2017	*k*-nearest neighbor, naïve Bayes, decision tree, random forest, support vector machine, logistic regression	accuracy, precision, area under the curve, sensitivity, specificity	Weka	China	Regional distributed EHR in Shanghai China	2012-2014

REFERENCES

Acciaroli, G., Sparacino, G., Hakaste, L., Facchinetti, A., Di Nunzio, G. M., Palombit, A., Tuomi, T., Gabriel, R., Aranda, J., Vega, S. & Cobelli, C. (2018). Diabetes and prediabetes classification using glycemic variability indices from continuous glucose monitoring data. *Journal of diabetes science and technology*, *12*(1), 105-113.

Agarwal, V., Podchiyska, T., Banda, J. M., Goel, V., Leung, T. I., Minty, E. P., Sweeny, T., Gyang, E. & Shah, N. H. (2016). Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, *23*(6), 1166-1173.

Allalou, A., Nalla, A., Prentice, K. J., Liu, Y., Zhang, M., Dai, F. F., Ning, J. X., Osborne, L. R., Cox, B. J., Gunderson, E. P., & Wheeler, M. B. (2016). A predictive metabolic signature for the transition from gestational diabetes to type 2 diabetes. *Diabetes*, db151720.

American Association for Clinical Chemistry. (2017). Reference Ranges and What They Mean. Retrieved from https://labtestsonline.org/understanding/features/ref-ranges

American College of Cardiology. (2017). New ACC/AHA High Blood Pressure Guidelines Lower Definition of Hypertension. Retrieved from https://www.acc.org/latest-in-cardiology/articles/2017/11/08/11/47/mon-5pm-bp-guideline-aha-2017

American Diabetes Association. (2017). Classification and diagnosis of diabetes. *Diabetes Care*, *40*(Supplement 1), S11-S24.

American Diabetes Association. (2017). Standards of medical care in diabetes—2017 abridged for primary care providers. *Clinical Diabetes*, *35*(1), 5-26.

American Diabetes Association. (2013). Economic costs of diabetes in the US in 2012. *Diabetes care*, *36*(4), 1033-1046.

American Diabetes Association. (2015). The Cost of Diabetes. Retrieved from http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html?

American Heart Association News. (2017). Nearly half of US adults could now be classified with high blood pressure, under new definitions. Retrieved from https://www.acc.org/latest-in-cardiology/articles/2017/11/08/11/47/mon-5pm-bp-guideline-aha-2017

American Red Cross. (2017). History of Blood Transfusions. Retrieved from http://www.redcrossblood.org/learn-about-blood/blood-transfusions/history-blood-transfusions.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, *9*(7), 1545-1588.

Anderson, A. E., Kerr, W. T., Thames, A., Li, T., Xiao, J., & Cohen, M. S. (2016). Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study. *Journal of biomedical informatics*, *60*, 162-168.

Anderson, J. P., Parikh, J. R., Shenfeld, D. K., Ivanov, V., Marks, C., Church, B. W., Laramie, J. M., Mardekian, J., Piper, B. A., Willke, R. J., & Rublee, D. A. (2016). Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *Journal of diabetes science and technology*, *10*(1), 6-18.

Apache Incubator. (n.d.). Welcome to Apache SINGA. Retrieved from https://mahout.apache.org/

Appold, K. (2009). Determining Laboratory Reference Intervals: CLSI Guideline Makes the Task Manageable. *Laboratory Medicine,* 40 (2), 75-76. doi: 10.1309/LMEHV3HP39QOFJPA

Arellano-Valle, R.B., Gomez, H.W., & Quintana, F.A. (2004). A new class of skew normal distributions, *Communications in Statistics-Theory and Methods*; 33(7): 1465-1480. doi: 10.1081/STA-120037254

Ashley, E. A. (2015). The precision medicine initiative: a new national effort. *Jama*, *313*(21), 2119-2120.

Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. Lancet, 383(9911), 69–82. http://doi.org/10.1016/S0140-6736(13)60591-7

AWS. (2018). Machine Learning on AWS. Retrieved from https://aws.amazon.com/machine-learning/

Azzalini, A. (1985). A Class of Distributions Which Includes the Normal Ones, *Scandinavian Journal of Statistics*; 12(2), 171–178.Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew normal distribution, *Biometrika*; 83(4), 715-726. doi: https://doi.org/10.1016/S0047-259X(03)00131-3

Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. Biometrika 83 715–726. *Mathematical Reviews (MathSciNet): MR1440039 Zentralblatt MATH*, *885*.

Basu, S., Raghavan, S., Wexler, D. J., & Berkowitz, S. A. (2018). Characteristics Associated With Decreased or Increased Mortality Risk From Glycemic Therapy Among Patients With Type 2 Diabetes and High Cardiovascular Risk: Machine Learning Analysis of the ACCORD Trial. *Diabetes care*, *41*(3), 604-612.

Basu, S., Sussman, J. B., Berkowitz, S. A., Hayward, R. A., & Yudkin, J. S. (2017). Development and validation of Risk Equations for Complications Of type 2 Diabetes (RECODe) using individual participant data from randomised trials. *The Lancet Diabetes & Endocrinology*, *5*(10), 788-798.

Bera, A.K., & Bilias, Y. (2002). The MM, ME, ML, EL, EF, and GMM approaches to estimation: a synthesis, *Journal of Econometrics;* 107(1-2), 51-86. doi: 10.1016/S0304-4076(01)00113-0.

Berwick, D. M., Nolan, T. W., & Whittington, J. (2008). The triple aim: care, health, and cost. *Health affairs*, *27*(3), 759-769.

Boylan, G. (2013). Robust Parameter Design in Complex Engineering Systems. (Doctoral dissertation). Retrieved from Tiger Prints, Clemson University.

Boylan, G.L., & Cho, B.R. (2012). The Normal Probability Plot as a Tool for Understanding Data: A Shape Analysis from the Perspective of Skewness, Kurtosis, and Variability, *Quality and Reliability Engineering International*; 28: 249-264. Doi: 10.1002/qre.1241.

Box, G. (1988). Signal-to-noise ratios, performance criteria, and transformations, *Technometrics*; 30(1), 1-17. doi: 10.2307/1270311

Brewster, L. M., Mairuhu, G., Sturk, A., & van Montfrans, G.A. (2007). Distribution of creatine kinase in the general population: Implications for statin therapy, *American Heart Journal*, 154(4), 655-661. doi: 10.1016/j.ahj.2007.06.008

Breiman, L., & Cutler, A. (2007). Random forests-classification description. *Department of Statistics, Berkeley*, *2*.

Bughin, J., Hazen, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., Henke, N., and Trench, M. (2017). Artificial Intelligence: The Next Digital Frontier?. McKinsey Global Institute. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx.

Burt, V. L., Whelton, P., Roccella, E. J., Brown, C., Cutler, J. A., Higgins, M., & Labarthe, D. (1995). Prevalence of hypertension in the US adult population. *Hypertension*, *25*(3), 305-313.

Cao, H., Chen, J., Meyer-Lindenberg, A., & Schwarz, E. (2017). A polygenic score for schizophrenia predicts glycemic control. *Translational psychiatry*, 7(12), 1-9.

Carson JL, Grossman BJ, Kleinman S, Tinmouth AT, Marques MB, Fung MK, Holcomb JB, Illoh O, Kaplan LJ, Katz LM, Rao SV, Roback JD, Shander A, Tobian AA, Weinstein R, Swinton M, Djulbegovic B. (2012). Clinical Transfusion Medicine Committee of the, AABB. "Red Blood Cell Transfusion: A Clinical Practice Guideline From the AABB". *Annals of Internal Medicine.* 157: 49–58. doi:10.7326/0003-4819-157-1-201206190-00429. PMID 22751760.

Casanova, R., Saldana, S., Simpson, S. L., Lacy, M. E., Subauste, A. R., Blackshear, C., Wagenknecht, L., & Bertoni, A. G. (2016). Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning. *PloS one*, *11*(10), e0163942.

Centers for Disease Control and Prevention. (2018). *Diabetes Report Card 2017*. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services.

Centers for Disease Control and Prevention. (2016). High Blood Pressure Facts. Retrieved from https://www.cdc.gov/bloodpressure/facts.htm

Centers for Disease Control and Prevention. (2017). *National Diabetes Statistics Report, 2017*. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2017.

Center for Disease Control Newsroom. (2014). Up to 40 percent of annual deaths from each of five leading causes were preventable. Retrieved from https://www.cdc.gov/media/releases/2014/p0501-preventable-deaths.html.

Ceriotti, F. (2012). Establishing Pediatric Reference Intervals: A Challenging Task, *Clinical Chemistry*, 58(5), 808-810. doi: 10.1373/clinchem.2012.183483

Ceriotti, F. (2017). Quality specifications for the extra-analytical phase of laboratory testing: Reference intervals and decision limits, *Clinical Biochemistry*, 50(10-11), 595-598. doi: 10.1016/j.clinbiochem.2017.03.024

Ch'ng, C.K., Quah, S.H., & Low, H,C. (2005). The MM-Estimator in Response Surface Methodology, *Quality Engineering*; 17(4), 561-565. doi: 10.1080/08982110500225323

Chen, J., Tang, H., Huang, H., Lv, L., Wang, Y., Liu, X., & Lou, T. (2015). Development and validation of new glomerular filtration rate predicting models for Chinese patients with type 2 diabetes. *Journal of translational medicine*, *13*(1), 317.

Cho, B.R. Optimization issues in quality engineering, *Ph.D. diss*. University of Oklahoma: Norman, OK, 1994.

Cho, B.R., & Park, C. (2005). Robust design modeling and optimization with unbalanced data, *Computers & Industrial Engineering*; 48(2), 173-180. doi: 10.1016/j.cie.2005.01.004

ClinLab Navigator, LLC. (2017). Reference Ranges. Retrieved from http://www.clinlabnavigator.com/reference-ranges.html.

CLSI. (2008). *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline -Third Edition*. CLSI document EP28-A3c. Wayne, PA: Clinical and Laboratory Standards Institute.

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, *372*(9), 793-795.

Copeland, K.A., & Nelson, P.R. (1996). Dual response optimization via direct function minimization, *Journal of Quality Technology*, 28(3). 331-336.

Costa, N.R.P.(2010). Simultaneous optimization of mean and standard deviation, *Quality Engineering*, 22(3), 140-149. doi: 10.1080/08982110903394205.

Cox, L., & Peck, P. (2009). The Top 10 Medical Advances of the Decade. Medpage Today. Retrieved from https://www.medpagetoday.com/infectiousdisease/publichealth/17594.

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 59.

Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., & Bellazzi, R. (2017). Machine Learning Methods to Predict Diabetes Complications. *Journal of diabetes science and technology*, 1932296817706375.

Davidoff, F., Haynes, B., Sackett, D., & Smith, R. (1995). Evidence based medicine. *BMJ: British Medical Journal*, *310*(6987), 1085.

Davis, K. (2013). Statistical Brief #404: Expenditures for Hypertension among Adults Age 18 and Older, 2010: Estimates for the U.S. Civilian Noninstitutionalized Population. Retrieved from https://meps.ahrq.gov/data_files/publications/st404/stat404.shtml

DeJesus, R. S., Breitkopf, C. R., Rutten, L. J., Jacobson, D. J., Wilson, P. M., & Sauver, J. S. (2017). Incidence rate of prediabetes progression to diabetes: modeling an optimum target group for intervention. *Population health management*, *20*(3), 216-223.

Del Castillo, E. & Montgomery, D.C. (1993). A nonlinear programming solution to the dual response problem, *Journal of Quality Technology*, 25(3), 199-204.

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060-1089.

Dieleman, J. L., Baral, R., Birger, M., Bui, A. L., Bulchis, A., Chapin, A., Hamadvid, H., Horst, C., Johnson, E.K., Joseph, J., Lavado, R., Lomsadze, L., Reynolds, A., Squires, E., Cambell, M., DeCenso, B., Dicker, D., Flaxman, A., Gabert, R., Highfill, T., Naghavi, M., Nightingale, N., Templin, T., Tobias, T., Vos, T., & Murray, C.(2016). US spending on personal health care and public health, 1996-2013. *Jama*, *316*(24), 2627-2646.

Ding, R., Lin, D.K.J., & Wei, D. (2004). Dual Response Surface Optimization: A Weighted MSE Approach, *Quality Engineering*, 16(3), 377-385. Doi: 10.1081/QEN-120027940.

Dong, S. S., Guo, Y., Yao, S., Chen, Y. X., He, M. N., Zhang, Y. J. Chen, X., Chen, J., & Yang, T. L. (2017). Integrating regulatory features data for prediction of functional disease-associated SNPs. *Briefings in bioinformatics*.

Du, J., Park, Y.-T., Theera-Ampornpunt, N., McCullough, J. S., & Speedie, S. M. (2012). The use of count data models in biomedical informatics evaluation research. *Journal of the American Medical Informatics Association*, 19(1), 39–44. http://doi.org/10.1136/amiajnl-2011-000256.

ElTanboly, A., Ismail, M., Shalaby, A., Switala, A., El-Baz, A., Schaal, S.,Gimel'farb, G., & El-Azab, M. (2017). A computer-aided diagnostic system for detecting diabetic retinopathy in optical coherence tomography images. *Medical physics*, *44*(3), 914-923.

Farran, B., Channanath, A. M., Behbehani, K., & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ open*, *3*(5), e002457.

Fowler, M. J. (2008). Microvascular and macrovascular complications of diabetes. *Clinical diabetes*, *26*(2), 77-82.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics*, *15*(1-2), 107-143.

Goethals, P.L., & Cho, B.R. (2012). Designing the optimal process mean vector for mixed multiple quality characteristics, *IIE Transactions*, 44(11), 1002-1021. doi: 10.1080/0740817X.2012.655061

Goethals, P.L., & Cho, B.R. (2011). Solving the optimal process target problem using response surface designs in heteroscedastic conditions, *International Journal of Production Research*, 49(12), 3455-3478. doi: 10.1080/0740817X.2012.655061.

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on* (pp. 6645-6649). IEEE.

Hamada, M., & Nelder, J.A. (1997). Generalized linear models for quality improvement experiments, *Journal of Quality Technology*, 29(3), 292-304.

Han, L., Luo, S., Wang, H., Pan, L., Ma, X., & Zhang, T. (2017). An Intelligible Risk Stratification Model Based on Pairwise and Size Constrained Kmeans. *IEEE journal of biomedical and health informatics*, *21*(5), 1288-1296.

Harpaz, R., Chase, H. S., & Friedman, C. (2010, October). Mining multi-item drug adverse effect associations in spontaneous reporting systems. In *BMC bioinformatics* (Vol. 11, No. 9, p. S7). BioMed Central.

Harvard Health Publishing (2009). Medications for Treating Hypertension. Harvard Women's Health Watch. Retrieved from https://www.health.harvard.edu/heart-health/medications-for-treating-hypertension

Hasenkamp T., Arvidsson M., & Gremyr I. (2009) A review of practices for robust design methodology. Journal of Engineering Design, 20(6), 645-657, doi: 10.1080/09544820802275557

Health Sciences South Carolina (2017). Our Tools and Services. Retrieved from https://www.healthsciencessc.org/our-tools-services

Hertroijs, D. F., Elissen, A. M., Brouwers, M. C., Schaper, N. C., Köhler, S., Popa, M. C., Asteriadis, S., Hendriks, S. H., Bilo, H. J., & Ruwaard, D. (2017). A risk score including body mass index, glycated haemoglobin and triglycerides predicts future glycaemia control in people with type 2 diabetes. *Diabetes, Obesity and Metabolism*.

Horn, P.S., & Pesce, A.J. (2003). Reference intervals: an update. *Clinica Chimica Acta*, *334*(1-2), 5-23. Doi: 10.1016/S0009-8981(03)00133-5

Hothorn, T. (2018). CRAN Task View: Machine Learning. Retrieved from https://cran.r-project.org/web/views/MachineLearning.html

Huber, P.J. (1973). Robust Regression: Asymptotics, Conjectures, and Monte Carlo, *The Annals of Statistics,* 1(5), 799-821.

Hudson, K., Lifton, R., & Patrick-Lake, B. (2015). The precision medicine initiative cohort program—Building a Research Foundation for 21st Century Medicine. *Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, ed*.

Howard, A. (2012). Coding for Hypertension. For the Record 24(8), 28. Retrieved from http://www.fortherecordmag.com/archives/042312p28.shtml

H2O.ai. (2018). #1 Open-source machine learning platform for enterprises. Retrieved from https://www.h2o.ai/h2o/

Jameson, J. L., & Longo, D. L. (2015). Precision medicine—personalized, problematic, and promising. *Obstetrical & Gynecological Survey*, *70*(10), 612-614.

Jeong, I., Kim, K., & Chang, S.Y. (2005). Optimal Weighting of Bias and Variance in Dual Response Surface Optimization, *Journal of Quality Technology*, 37(3), 236-247.

Johnson, C. (2017, February 15). Why America's health-care spending is projected to soar over the next decade. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2017/02/15/u-s-health-care-spending-projected-to-soar-to-5-5-trillion-by-2025/?utm_term=.2672f2b90e9a

Kagawa, R., Kawazoe, Y., Ida, Y., Shinohara, E., Tanaka, K., Imai, T., & Ohe, K. (2017). Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach. *Journal of diabetes science and technology*, *11*(4), 791-799.

Kahn, R. (2003). Follow-up Report on the Diagnosis of Diabetes Mellitus: The Expert Committee on the Diagnosis and Classifications of Diabetes Mellitus, *Diabetes Care*, 26(11), 3160-3167.

Karst, O.J. (1958). Linear Curve Fitting Using Least Deviations, *Journal of the American Statistical Association*, 53(281), 118-32. doi: 10.2307/2282572

Katayev, A., Balciza, C., & Seccombe, D. (2010). Establishing Reference Intervals for Clinical Laboratory Test Results: Is There a Better Way? *American Journal of Clinical Pathology*, 133(2), 180-186. doi: 10.1309/AJCPN5BMTSF1CDYP

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal.* 15, 104-116.

Kibaya, R.S., Bautista, C.T., Sawe, F.K., Shaffer, D.N., Sateren, W.B., Scott, P.T., Nelson, L.M., Robb, M.L., Birx, D.L., & de Souza, M.S. (2008). Reference Ranges for the Clinical Laboratory Derived from a Rural Population in Kericho, Kenya. *PLOS One*, 3(10), e3327. doi: 10.1371/journal.pone.0003327

Kim, K., & Lin, D.K.J. (2006). Optimization of multiple responses considering both location and dispersion effects, *European Journal of Operations Research*; 169(1), 133-145. doi: 10.1016/j.ejor.2004.06.020

Kim, K., & Lin, D.K.J. (1998). Dual Response Surface Optimization: A Fuzzy Modeling Approach, *Journal of Quality Technology*, 30(1), 1-10.

Kim, Y.J., & Cho, B.R. (2002). Development of Priority-Based Robust Design, *Quality Engineering*, 14(3), 355-363. doi: 10.1081/QEN-120001874

Kim, Y. M., Kathuria, P., & Delen, D. (2017). Machine Learning to Compare Frequent Medical Problems of African American and Caucasian Diabetic Kidney Patients. *Healthcare informatics research*, *23*(4), 241-248.

Koksoy, O., & Yalcinoz, T. (2008). Robust Design using Pareto-type optimization: A genetic algorithm with arithmetic crossover, *Computers & Industrial Engineering*, 55(1), 208-218. doi: 10.1016/j.cie.2007.11.019

Koutrouvelis, I.A. & Canavos, G.C. (2000). A comparison of moment-based methods of estimation for the log Pearson type 3 distribution, *Journal of Hydrology*, 234(1-2), 71-81. doi: 10.1016/S0022-1694(00)00241-9

Kovach, J., & Cho, B.R. (2007). Development of a D-optimal robust design model for restricted experiments, *International Journal of Industrial Engineering – Theory Applications and Practice*, 14(2), 117-128.

Kovach, J., & Cho, B.R. (2008). Development of a multidisciplinary-multiresponse robust design optimization model, *Engineering Optimization*, 40(9), 805-819. doi: 10.1080/03052150802046304.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689-702.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, *30*(2-3), 195-215.

Kudyba, S. P. (2010). *Healthcare informatics: improving efficiency and productivity*. CRC Press.

Kutner, M.H., Nachtsheim, C.J., Neter, J., & Li, W. (2003). *Applied Linear Statistical Models* (5th edn*)*. McGraw-Hill: Boston, MA, 429.

Lee, B. J., & Kim, J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *IEEE journal of biomedical and health informatics*, *20*(1), 39-46.

Lee, B. J., Ku, B., Nam, J., Pham, D. D., & Kim, J. Y. (2014). Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. *IEEE journal of biomedical and health informatics*, *18*(2), 555-561.

Lee, Y., & Nelder, J.A. (2003). Robust Design via Generalized Linear Models, *Journal of Quality Technology*; 35(1) 2-12.

Lee, S.B., Park, C., & Cho, B.R. (2007). Development of a highly efficient and resistant robust design, *International Journal of Production Research*, 45(1) 157-167. doi: 10.1080/00207540600649202.

Leung, R. K., Wang, Y., Ma, R. C., Luk, A. O., Lam, V., Ng, M., So, W.Y., Tsui, S. K. W., & Chan, J. C. (2013). Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case–control cohort analysis. *BMC nephrology*, *14*(1), 162.

Lesko, L. J. (2007). Personalized medicine: elusive dream or imminent reality? *Clinical Pharmacology And Therapeutics*, *81*(6), 807-816.

Li, Y., Bai, C., & Reddy, C. K. (2016). A distributed ensemble approach for mining healthcare data under privacy constraints. *Information sciences*, *330*, 245-259.

Ligthart, S., van Herpt, T. T., Leening, M. J., Kavousi, M., Hofman, A., Stricker, B. H., van Hoek, M.m Sijbrands, E., Franco, O., & Dehghan, A. (2016). Lifetime risk of developing impaired glucose metabolism and eventual progression from prediabetes to type 2 diabetes: a prospective cohort study. *The lancet Diabetes & endocrinology*, *4*(1), 44-51.

Lin, D.K.J., & Tu, W. (1995). Dual Response Surface Optimization, *Journal of Quality Technology*, 27(1), 34-39.

Lo-Ciganic, W. H., Donohue, J. M., Thorpe, J. M., Perera, S., Thorpe, C. T., Marcum, Z. A., & Gellad, W. F. (2015). Using machine learning to examine medication adherence thresholds and risk of hospitalization. *Medical care*, *53*(8), 720.

López, B., Torrent-Fontbona, F., Viñas, R., & Fernández-Real, J. M. (2017). Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. *Artificial intelligence in medicine*.

Malka, R., Nathan, D. M., & Higgins, J. M. (2016). Mechanistic modeling of hemoglobin glycation and red blood cell kinetics enables personalized diabetes monitoring. *Science translational medicine*, *8*(359), 359ra130-359ra130.

Mahout. (2018). Mahout: For Creating Scalable Performant Machine Learning Applications. Retrieved from https://mahout.apache.org/

Marshall, W. J., & Bangert, S.K. (Eds.). (2008). *Clinical Biochemistry: Metabolic and Clinical Aspects, Second Edition*. Churchill Livingstone, Philadelphia, PA.

Martin, A. B., Hartman, M., Washington, B., Catlin, A., & National Health Expenditure Accounts Team. (2016). National health spending: faster growth in 2015 as coverage expands and utilization increases. *Health Affairs*, *36*(1), 166-176.

Mayo Clinic (2018). Type 1 Diabetes. Retrieved from https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/symptoms-causes/syc-20353011

May Clinic (2018). Type 2 Diabetes. Retrieved from https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193

Mayo Clinic. (2018g). Test ID: STSH. Retrieved from https://www.mayomedicallaboratories.com/test-catalog/Clinical+and+Interpretive/8939

McCoy, R. G., Ngufor, C., Van Houten, H. K., Caffo, B., & Shah, N. D. (2017). Trajectories of Glycemic Change in a National Cohort of Adults With Previously Controlled Type 2 Diabetes. *Medical care*, *55*(11), 956-964.

Merriam-Webster (2017). *Analyte*. Retrieved from https://www.merriam-webster.com/dictionary/analyte.

Moreno, E. M., Luján, M. J. A., Rusinol, M. T., Fernández, P. J., Manrique, P. N., Trivino, C. A., Miquel, M. P., Rodriguez, M. A., & Burguillos, M. J. G. (2017). Type 2 diabetes screening test by means of a pulse oximeter. *IEEE Transactions on Biomedical Engineering*, *64*(2), 341-351.

Muhlbauer, A., Spichtinger, P., & Lohmann, U. (2009). Application and Comparison of Robust Linear Regression Methods for Trend Estimation, *Journal of Applied Meteorology and Climatology*, 48(9), 1961-1970. doi: 10.1175/2009JAMC1851.1

Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002). *Generalized Linear Models: With Applications in Engineering and the Sciences.* Wiley: New York, 2002.

Myers, R.H., & Montgomery, D.C. (1997). A Tutorial on Generalized Linear Models, *Journal of Quality Technology*, 29(3), 274-291.

National Academy of Engineering. (2018). *Introduction to the Grand Challenges for Engineering*. Retrieved from http://www.engineeringchallenges.org/challenges/16091.aspx

National Center for Biotechnology Information. (2018, March 22). Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/.

National Center for Health Statistics. (2016). Health, United States, 2015: with special feature on racial and ethnic health disparities.

National Center for Health Statistics, & Centers for Medicare and Medicaid Services. (2010). ICD-10-CM official guidelines for coding and reporting. *Washington (DC): US GPO*.

National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.) Diabetes Tests & Diagnosis. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/tests-diagnosis

National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). Symptoms & Causes of Diabetes. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes

NET Framework. (2018). Accord.NET Framework. Retrieved from http://accord-framework.net/

Neugebauer, R., Fireman, B., Roy, J. A., Raebel, M. A., Nichols, G. A., & O'Connor, P. J. (2013). Super learning to hedge against incorrect inference from arbitrary parametric assumptions in marginal structural modeling. *Journal of clinical epidemiology*, *66*(8), S99-S109.

Neugebauer, R., Fireman, B., Roy, J. A., Raebel, M. A., Nichols, G. A., & O'Connor, P. J. (2013). Super learning to hedge against incorrect inference from arbitrary parametric assumptions in marginal structural modeling. *Journal of clinical epidemiology*, *66*(8), S99-S109.

Obama, B. (2015). President Obama's 2015 State of the Union Address. *WhiteHouse. gov, https://www. whitehouse. gov/sotu (accessed April 29, 2015). Notes*, *259*.

O'Hagan, A., & Leonhard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints, *Biometrika,* 63(1), 201-202. doi: 10.2307/2335105

Olivera, A. R., Roesler, V., Iochpe, C., Schmidt, M. I., Vigo, Á., Barreto, S. M., & Duncan, B. B. (2017). Comparison of machine-learning algorithms to build a

predictive model for detecting undiagnosed diabetes-ELSA-Brasil: accuracy study. *Sao Paulo Medical Journal*, *135*(3), 234-246.

Oryx 2. (n.d.). Overview. Retrieved from http://oryx.io/

Ozery-Flato, M., Ein-Dor, L., Parush-Shear-Yashuv, N., Aharonov, R., Neuvirth, H., Kohn, M. S., & Hu, J. (2016). Identifying and investigating unexpected response to treatment: a diabetes case study. *Big data*, *4*(3), 148-159.

Ozery-Flato, M., Parush, N., El-Hay, T., Visockienė, Ž., Ryliškytė, L., Badarienė, J., Solovjova, S., Kovaite, M., Navickas, R., & Laucevičius, A. (2013). Predictive models for type 2 diabetes onset in middle-aged subjects with the metabolic syndrome. *Diabetology & metabolic syndrome*, *5*(1), 36.

Paauw, D. (2017). Double the Dose of antihypertensive drugs? Internal Medicine News. Retrieved from http://www.mdedge.com/internalmedicinenews/article/138365/cardiology/double-dose-antihypertensive-meds

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine learning*, *27*(3), 313-331.

Pedersen, H. K., Gudmundsdottir, V., Pedersen, M. K., Brorsson, C., Brunak, S., & Gupta, R. (2016). Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. *NPJ genomic medicine*, *1*, 16035.

Pimentel, A., Carreiro, A. V., Ribeiro, R. T., & Gamboa, H. (2016). Screening diabetes mellitus 2 based on electronic health records using temporal features. Health informatics journal, 1460458216663023

Pegues, K.K., Boylan, G. L., & Cho, B. R. (2017). Decision making in health care using robust parameter design with conditions-based selection of regression estimators. *Quality and Reliability Engineering International*, *33*(8), 2151-2169.

Peddinti, G., Cobb, J., Yengo, L., Froguel, P., Kravić, J., Balkau, B., Tuomi, T., Aittokallio, T., & Groop, L. (2017). Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia*, *60*(9), 1740-1750.

Phillips, M.D., Madduri, K.S., & Cho, B.R. (1995). *Enhanced optimization strategies for robust design*, Fourth Industrial Engineering Research Conference, Nashville, TN.

Plebani, M. (2016). Harmonization in laboratory medicine: Requests, Samples, Measurements and Reports. *Critical reviews in Clinical Laboratory Sciences*, *53*(3), 184-196. doi: 10.3109/10408363.2015.1116851.

Pippitt, K., & Li, M. (2016). Diabetes mellitus: screening and diagnosis. *American Academy of Family Physicians*, 93(2), 103-109

Polio. (2017). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/features/poliofacts/.

Polio: Global Eradication Initiative. (n.d.). Retrieved from http://polioeradication.org/polio-today/history-of-polio/.

R Core Team. (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rachana, P. R., & Anuradha, H. V. (2014). Anti hypertensive prescribing patterns and cost analysis for primary hypertension: a retrospective study. *Journal of clinical and diagnostic research: JCDR*, *8*(9), HC19.

Rafey, M. (2013). Hypertension. Cleveland Clinic: Center for Continuing Education. Retrieved from http://www.clevelandclinicmeded.com/medicalpubs/diseasemanagement/nephrology/arterial-hypertension/

Ramezankhani, A., Pournik, O., Shahrabi, J., Azizi, F., Hadaegh, F., & Khalili, D. (2016). The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. *Medical decision making*, *36*(1), 137-144.

Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, *3*(4), 277-287.

Ricos, C., Cava, F., García-Lario, J. V., Hernandez, A., Iglesias, N., Jimenez, C.V., Minchinela, J., Perich, C., Domenech, M.V., & Alvarez, V. (2004). The reference change value: a proposal to interpret laboratory reports in serial testing based on biological variation. *Scandinavian Journal of Clinical and Laboratory Investigation*, *64*(3), 175-184. doi: 10.1080/00365510410004885

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). IBM.

Robinson, T.J., Borror, C.M., & Myers, R.H. (2004). Robust Parameter Design: A Review, *Quality and Reliability Engineering International*; 20(1), 81-101. doi: 10.1002/qre.602

Robinson, T.J., Wulff, S.S., Montgomery, D.C., & Khuri, A.I. (2006). Robust Parameter Design Using Generalized Linear Mixed Models, *Journal of Quality Technology*; 38(1), 66-75.

Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of American Statistical Association*; 79, 871-880. doi: 10.2307/2288718

Rousseeuw, P.J. & Yohai, V.J. (1984). Robust regression by means of S-estimators in Robust and Nonlinear Time Series Analysis. In *Lecture Notes in Statistics No. 26*, Franke J, Hardle, W and Martin D (eds). Springer: Berlin, Germany, 256-272.

Ryan, T.P. (2009). *Modern Regression Methods*, 2nd ed. Wiley: Hoboken, NJ, 2009.

Shaibu, A.B., & Cho, B,R. (2009). Another view of dual response surface modeling and optimization in robust parameter design, *International Journal of Advanced Manufacturing Technology*, 41(7), 631-641.

Shaibu, A.B, Cho, B.R., & Kovach, J. (2009). Development of a Censored Robust Design Model for Time-Oriented Quality Characteristics, *Quality and Reliability Engineering International*, 25(2), 181-197. doi: 10.1002/qre.961.

Shin, S., & Cho, B.R. (2008). Development of a sequential optimization procedure for robust design and tolerance design within a bi-objective paradigm, *Engineering Optimization*, 40(11), 989-1009.

Shin, S., & Cho, B.R. (2009). Studies on a bi-objective robust design optimization problem, *IIE Transactions*; 41: 957-968. doi: 10.1080/07408170902789084

Shin, S., Samanlioglu, F., Cho, B.R., & Wiecek, M.M. (2011). Computing trade-offs in robust design: Perspectives of the mean squared error, *Computers & Industrial Engineering*, 60, 248-255. doi: 10.1016/j.cie.2010.11.006

Shogun. (n.d.). Unified and efficient machine learning library. Retrieved from http://www.shogun-toolbox.org/

Simpson, J.R. & Montgomery, D.C. (1998). A performance-based assessment of robust regression methods, *Communication in Statistics – Simulation and Computation*, 27(4), 1031–1049. doi: 10.1080/03610919808813524

Skolnik, N. S., Beck, J. D., & Clark, M. (2000). Combination antihypertensive drugs: recommendations for use. *American family physician*, *61*(10), 3049-3056.

Sudharsan, B., Peeples, M., & Shomali, M. (2014). Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *Journal of diabetes science and technology*, *9*(1), 86-90.

Sunderman, F.W. (1975). Current Concepts of "Normal Values," "Reference Values," and "Discrimination Values" in Clinical Chemistry. *Clinical Chemistry*, 21(13), 1873-1877.

Tang, L.C., & Xu, K. (2002). A Unified Approach for Dual Response Surface Optimization, *Journal of Quality Technology*, 34(4), 437-447.

Taguchi, G. (1986). *Introduction to Quality Engineering*. UNIPUB/Kraus International: White Plains, NY.

Taguchi, G. (1987). *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost*. UNIPUB/Kraus International: White Plains, NY.

Taguchi, G., Chowdhry, S., & Wu, Y. (2004). *Taguchi's Quality Engineering Handbook,* Wiley, New York, NY.

Tsai, A. (2015, September). 6 Tests to Determine Diabetes Type. Retrieved from: http://www.diabetesforecast.org/2015/sep-oct/tests-to-determine-diabetes.html

Tsui, K. (1992). An Overview of Taguchi Method and Newly Developed Statistical Methods for Robust Design, *IIE Transactions*, 24(5), 44-57. doi: 10.1080/07408179208964244

Vining, G.G., & Myers, R.H. (1990). Combining Taguchi and Response Surface Philosophies: a Dual Response Approach, *Journal of Quality Technology*, 22, 38-45.

Vyas, R., Bapat, S., Jain, E., Karthikeyan, M., Tambe, S., & Kulkarni, B. D. (2016). Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis. *Computational biology and chemistry*, *65*, 37-44.

WEKA. (.n.d). Weka 3: Data Mining Software in Java. The University of Waikato. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

Willinger, W., Alderson, D., & Lun, L. (2004). *A Pragmatic Approach to Dealing with High-Variability in Network Measurements*, ACM SIGCOMM Conference on Internet Measurement, Taormina, Sicily, Italy. doi: 10.1145/1028788.1028800

World Health Organization. (2016). Global report on diabetes: World Health Organization.

Yohai, V.J. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression, *The Annals of Statistics,* 15(2), 642-656.

Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Yang, G., & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. *International journal of medical informatics*, *97*, 120-127.