Managing the Risks and Potential of High-tech Innovations-in-use:

Predictive Analytic Modeling with Big Data and a Longitudinal Field Study

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Ujjal Kumar Mukherjee

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Adviser: Dr. Kingshuk K. Sinha

July, 2015

Acknowledgements

I would like to express my sincerest gratitude to my adviser Dr. Kingshuk K. Sinha for his tireless mentoring, and guidance throughout my doctoral studies. I consider myself really fortunate to have Dr. Kingshuk K. Sinha as my adviser. Without his invaluable feedback and incessant encouragement, this dissertation would not have seen the light of the day. I have been recipient of Dr. Sinha's unconditional support in several of my personal crisis situations during the doctoral studies. I am also very grateful to Dr. Snigdhansu Chatterjee for his support and guidance in matters of statistical methods. I would also like to thank the rest of my committee members, Dr. Enno Siemsen and Dr. Mili Mehrotra for their time, helpful suggestions and comments which have considerably improved my dissertation research. I would also like to acknowledge the invaluable contribution of all other faculty members in the Supply Chain and Operations department.

I would like to express a special thanks to my wife Ms. Amrita Mukherjee and son Aniruddha Mukherjee for their invaluable and unconditional support over the years. Without their support and encouragement my dissertation would not have been possible. I would like to thank my parents Ms. Shyamali Mukherjee and Mr. Dilip Kumar Mukherjee for their continuous care, encouragement and support, which I consider to be the greatest gift anyone has ever given me. I am deeply indebted to my brother, friend, philosopher and guide Chanchal Mukherjee for all his support and help over the years without which I would not have been able to do my dissertation.

Finally, I would like to acknowledge the contribution of my collaborators and friends many of whom have invaluable contribution towards the dissertation. Specially, I would like to thank Shoubhik Sinha who had tremendous tangible contribution towards the third study of my dissertation. Without Shoubhik's help in collecting and collating field data related to the third study, the study would not have been possible. I owe a special thanks to Scott Bosch for his support and guidance, without which the third study would not have been possible. I would like to thank Ravindra Kasturi for all his support in the data collection and data organization for the first two studies. Also, I would like to acknowledge the support of my friend and colleague George Ball and Suvrat Dhanorkar who have helped me at every step throughout the years in my studies and research. I am also deeply indebted to Anupam Agrawal for his help and guidance throughout the process of my Ph.D. dissertation. Last, but not the least, I would like to thank my close friend Bhupinder Singh Juneja for many useful discussions and debates on a variety of topics which has helped me better my dissertation in many ways.

Dedication

To my late grandmother Smt. Hemnalini Devi who has a profound influence on my growing up, my amazingly caring parents Ms. Shyamali Mukherjee and Mr. Dilip Kumar Mukherjee, my wonderful wife Ms. Amrita Mukherjee and my ever supporting son Aniruddha Mukherjee.

Abstract

Healthcare, like many other industry sectors, is increasingly becoming high tech innovation enabled. Success and growth of many firms in high tech industry segments such as medical device, automotive, electronics, telecommunication, and aerospace are dependent on rapid pace of technology innovation. High tech innovation provides functionalities and benefits that are not feasible otherwise. As an illustration, in the healthcare delivery segment, rapid innovation in three dimensional imaging has made the disease detection process much more accurate, fast and consistent as compared to the past. However, notwithstanding the potential benefits of high tech innovations, high tech innovations entail risks of failure while in use in the market. Failures of high tech innovations-in-use can cause severe harm to the users. As an illustration, a recent incident of failure of a cardio-vascular defibrillator caused severe injuries including several fatalities to many patients. Hence, firms and regulators need to manage the downside risks of failures of high tech innovations-in-use in a timely manner. Realizing the potential benefits of high tech innovation in many usage areas depend on how well firms and regulators can manage the potential downside risks of high tech innovations-in-use. Also, realizing the potential benefits of a high tech innovation-in-use depend on how well users can learn to use the high tech innovations.

In my dissertation, I investigate how firms can best manage the downside risks of high tech innovations-in-use as well as how users can realize the potential benefits of high tech innovations. The dissertation consists of three inter-related studies. The first two studies are aimed towards managing the downside risks of high tech innovations. The last study looks at a specific high tech innovation in healthcare delivery, namely, surgical robots and detail out a field study to understand the factors that lead to the realization of the benefits of high tech innovations in health care.

The first step in managing risks of failures of high tech innovations-in-use is to be able to detect signals of failure from the market. In the first study, we show that it is possible to use user feedback of adverse events related to medical devices to detect signals of device failures originating from either design failures, or supply chain failures or manufacturing process failures. Using text mining and machine learning based predictive analysis methods on a ‘big’ unstructured data-set of adverse events reported by users of medical devices, we show that firms can detect failures of medical devices with precision and consistency. We also identify that firms exhibit substantial judgment bias in interpreting and reacting to market signals of failures. Either they under-react or they over-react to market signals of failure under certain conditions. We use the theoretical lenses of signal detection and system neglect to setup the study and identify sources of judgment bias.

In the second study we extend the first study by identifying product related, firm related and industry related conditions under which firms are more likely to systematically under-react or over-react to market signals of failures of high tech innovations-in-use. An acknowledgement of these sources would help firms and regulators to bring in greater consistency in their detection and decision process. We integrate the theoretical perspectives of signal detection, system neglect and attention based view of firms to propose a framework of judgment bias in the context of detection of failures of high tech innovations-in-use from user reports of adverse events.

In the third study, we undertake a field research in a large multispecialty hospital in the United States to investigate factors that lead to development of technology capability in healthcare delivery in the context of usage of a surgical robot, namely, da Vinci robot. We identify conditions related to surgeon and team learning that lead to improved usage of the robot. More importantly, we show that with surgeon and team learning, technology mediation can help reduce surgical outcome variation in spite of input heterogeneity in the form of surgeons' experience and skill heterogeneity, patient heterogeneity and team heterogeneity.

Subject Areas: *Management of Technology and Innovation, Supply Chain Analytics and Health Care Management*

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to the Dissertation

1.1 Problem Statement

Many firms in many industry segments such as medical device, healthcare, automotive and aerospace depend on rapid pace of high tech innovation for success and growth. High tech innovations serve several purposes. Firstly, high tech innovations make functionalities available that are not possible otherwise. Secondly, high tech innovations help standardize and improve services such as healthcare delivery. Lastly, high tech innovations improve efficiency and effectiveness of existing functionalities. Notwithstanding the many benefits of high tech innovations, there are also growing concerns about the downside risks, especially since failures of a high tech innovations-in-use can cause injury or death to users. Several recent incidents of high tech innovation failures have attracted public and media attention because of their potential risks to users from failures of those innovations. By way of illustrative example, failure of General Motors cars has been linked to a number of deaths. Such failures have caused severe loss of brand equity and loss of revenue to General Motors in recent years. Just in the year 2014 Toyota has recalled upwards of 2.5 million vehicle units. In the medical device industry, incidences of failures of products, especially new products, have increased considerably in the last decade. In spite of the best efforts of firms and governmental regulatory agencies, high tech innovations are continuing to fail while in use in the marketplace. The failures of high tech innovations-in-use are evidenced as increasing number of product recalls in almost every industry segment. These failures, if not managed well in a timely manner, tend to take away much of the potential and promised benefits of high tech innovations. Motivated by these real world problems, the first part of my dissertation looks at answering the question: *how can firms avoid or minimize the downside risks of failures of innovations-in-use?* The first two chapters of my dissertation look at how can firms and regulators proactively minimize the impact of potential failures of high tech innovations-in-use.

The second part of my dissertation looks at how can users realize the intended benefits of

high tech innovations. In many applications such as healthcare delivery, realizing the potential of a high tech innovation is dependent on users' ability to adopt and learn to use the high tech innovation. Specifically in the context of healthcare delivery, high tech innovations such as advanced imaging technology, advanced radiology, and advanced surgical technology have become enablers of high quality service delivery. Realizing the potential of high tech innovations is dependent on how effectively and efficiently users or groups of users can adopt and learn to use those innovations for improving the service delivery process. Motivated by this, the second part of my dissertation tries to answer the questions: *how can users realize the potential of high tech innovations-in-use, especially after their launch and while use in the market place?* I try to answer this question through a detailed field study of a surgical robot, namely, the da Vinci surgical robot, at a large multispecialty hospital in the United States.

The rest of the dissertation is arranged as follows. In the remaining sections of the introduction I try to provide a brief summary of the three constituent studies. In chapter two, I describe in details the first study which looks at detection and prediction of failures of high tech innovations-in-use. In chapter three, I describe the details of sources of judgment bias in detecting failures of high tech innovations-in-use from user feedback on adverse events. In chapter four, I describe the details of the field study related to the adoption and usage of da Vinci surgical robot in a multi-specialty hospital in the United States. Finally in chapter five I describe the key takeaways and conclusions from this dissertation study and indicate potential directions of future research.

1.2 Study I: Predicting Failures of High Tech Innovations-in-Use: Application of Predictive Analytics to Big Data on Market Failures of Medical Devices

Most high tech firms today realize that failure of technological innovations while in use in the marketplace cannot be completely eliminated. Hence, there is a growing need for timely detection of early signals of such failures. By way of an illustrative example, a class of high tech innovations where there is much need for detection of early signals of failures while in use in the marketplace is medical devices. In the context of medical devices, it is imperative that the downside risks of potential failures in the marketplace be managed in a timely manner, since there are serious economic and social consequences of medical device recalls. Hence, the primary research questions we address in the first study are: *Can user generated market feedback provide credible signals for*

detection and prediction of failures of high tech innovations-in-use? If so, how can firms and regulators improve the precision and consistency of failure detection?

The theoretical lenses we have used to frame the research questions into testable hypotheses are signal detection and system neglect. Medical device industry is the empirical setting for this study. The primary data-set we have used for the empirical analysis is the Food and Drug Administration's (FDA) "Manufacturer and User facility Device Experience (MAUDE)" data-set. The MAUDE database represents "big data" with excess of three million data points generated through reports of adverse incidents involving the usage of medical devices. We analyzed the data through a combination of explanatory econometric methods (namely, regularized regression, generalized linear mixed models and generalized additive models) and machine learning based predictive analytic methods (namely, random forest, boosting, neural network and support vector machines) to develop a modeling framework that can ex ante estimate the risk of failure of high tech innovations-in-use.

The contributions of this study are the following: First, we demonstrate that early detection of failures of high tech innovations-in-use is possible by analyzing "big" and unstructured data on user level market feedback related to usage experience and adverse events. Next, we demonstrate that the precision of such a failure detection system can be significantly improved by incorporating product, firm and industry level factors. In doing so, we uncover a number of new relationships that can influence the performance of technological innovations while in use in the market. Finally, we present novel and nuanced insights into why firms often fail to correctly detect failure signals, and how the existence of judgment bias prevents the timely detection of failure signals. And, to that end, we identify factors that influence firms and decision makers to under-react to certain signals and over-react to certain other signals from the market.

1.3 Study II: Evaluating Judgment Bias in Detecting Failures of High Tech Innovations-in-Use: Analysis of Big Data on User Reported Adverse Events Related to Medical Devices

The motivation for the second study stems primarily from two sources. Firstly, the findings of the first study indicate that, in practice, firms exhibit significant judgment bias – i.e., over-reaction bias and under-reaction bias – in interpreting market signals related to failures of high tech innovations-in-use. Also, we find that there are systematic variations in the accentuation of one type of bias

over the other. Secondly, recent news media reports on several highly publicized product failures in the medical devices, aerospace and automotive industries provide face validity to the findings of the first study. Hence, in the second study, we address the following questions related to judgment bias of firms in interpreting market signals and detecting failures of high tech innovations-in-use: *Do firms exhibit systematic judgment bias in detecting failures of high tech innovations-in-use in the market? If so, what are the conditions under which firms exhibit judgment bias by way of over-reaction and under-reaction to market signals?*

By way of theoretical foundation for this study, we integrate three theoretical perspectives, i.e., the signal detection theory, system neglect hypothesis and the attention based view of the firm. This integration allows us to extend the basic system neglect framework by analyzing different sources of variance in failure signals. Also, from the perspective of an attention based view of the firm, we incorporate factors that lead firms to choose sub-optimal decision thresholds in detecting signals of failures of technological innovations while in use in the marketplace. In this study, we consider several factors related to markets, firms and products that are sources of judgment bias in an industry. The market related factors are market scope and competition. The firm related factors are firm size and firm diversity (i.e., geographic diversity and market segment diversity), firm focus (i.e., exploitation of existing technology versus exploration of new technology), and firm structure (i.e., partnerships, joint ventures and degree of outsourcing). The product related factors are technology life-cycle (i.e., maturity of the underlying technology), and usage of the product (i.e., the intended primary usage of a product).

As in the first study, medical device industry is the empirical setting for this study. We combined several data-sets to generate the variables for the study. In this study, too, the primary data-set is the Food and Drug Administration's (FDA) "Manufacturer and User facility Device Experience (MAUDE)" data-set which is a "big" data-set with excess of three million data points. We combined the MAUDE data with firm related data from Compustat and several firm sources. From the standpoint of research methods, we analyzed the "big" data-set for this study using a combination of predictive analytics (random forest ensemble model) and econometric methods (hierarchical linear models and generalized linear mixed models). The results indicate that significant and systematic variations exist in the nature of judgment bias in interpreting market signals related to technological innovation failures in the medical device industry.

The contributions of the second study are the following. From a theoretical standpoint, propose an integrated framework of judgment bias in detection of failures from user reported adverse

incidents related to high tech innovation-in-use. This integrated framework sheds light into how decision makers receive and interpret market signals related to high tech innovations-in-use, thereby advancing the literature on technology and innovation management, as well as the literature on behavioral decision making. From a practical standpoint, this study makes a contribution by identifying factors associated with firms and decision makers exhibiting under-reaction or over-reaction to market signals. The study is consequential to firms as well as regulators since it identifies the primary sources of judgment bias in detecting market failures. An acknowledgement of the sources will enable practitioners to deliberately account for some of the sources of bias while making critical product innovation related decisions such as continuous market surveillance, field testing and product planning.

1.4 Study III: Enabling Healthcare Delivery with High Tech Innovation: A Longitudinal Field Study of Robot-Assisted Surgery

High tech innovations are becoming enablers of service delivery. Realizing the potential of a high tech innovation is contingent on how efficiently and effectively individuals and groups of service providers learn to use the technological innovation for delivering a service. Specifically, realizing the potential of a technological innovation depends on the nature and rate of learning of individuals and groups to use the technological innovation, and the impact of individual and group learning on the overall usage of the technological innovation. We ground this study in the context of health care delivery since it is an operational context that is becoming increasingly enabled by high tech innovations. The primary questions we address in this study are the following: *What is the nature of individual and group learning mediated by a high tech innovation? How does individual and group learning mediated by a high tech innovation impact the usage of the innovation and realize the potential of high tech innovation?*

By way of theoretical foundation, the study draws on and synthesizes two streams of literature to develop testable hypotheses. Firstly, we draw on a stream of literature on new technology adoption, usage and learning in health care delivery, the context in which this study is grounded. To date, this stream of literature has focused on managing sources of input heterogeneity that leads to outcome variation in health care delivery. Secondly, we draw on an emerging stream of medical science literature focused on evaluating the "comparative effectiveness" of a high tech innovation in use on health care outcomes. We evaluate the potential of a high tech innovation by examining

the impact of the learning of individual and groups mediated by the innovation on: (i) increasing the usage of high tech innovation and (ii) reducing the variation of healthcare delivery outcomes. We posit that notwithstanding input (patient, surgeon and staff experience and skill) heterogeneity that, typically, is beyond the control of a health care delivery organization, learning mediated by high tech innovation will reduce outcome variation.

By way of research design, we conduct a longitudinal field study to investigate how the potential of a major high tech innovation – namely, surgical robot – can be realized to conduct surgeries. The setting of the field study is a major multi-specialty hospital in the United States. The high tech innovation that is subject of this study is a surgical robot, namely, da Vinci robot. We collected detailed data since the surgical robot was first adopted in the hospital in 2008 to until 2013, and was used to conduct a total of 1380 common robot-assisted surgeries, comprising of two types of urology procedures: pelviscopy and prostatectomy; and two types of OB/GYN procedures: hysterectomy and sacrocolpopexy. Through mixed effects econometric modeling, we estimated the relationship between the surgical process-time variation and experience of the surgeons measured by the cumulative number of robot-assisted surgeries performed by a surgeon. The model estimation results indicate that, as expected, surgeons' learning is significant with respect to cumulative experience. However, what is surprising is that the heterogeneity in surgeons' learning and performance is negligible in robot-assisted surgical procedures. This is a significant finding since the high tech innovation (da Vinci robot) dampens human skill based performance variation in performing specific surgeries. More generally, this implies that it is possible to minimize or eliminate the risks associated with human skill based health care delivery (namely, hand trembling of surgeons and fatigue) and transform health care delivery to become more reliable and replicable over a prolonged period of time, thereby increasing the health care delivery capacity.

The primary contribution of this study is in shifting the focus of the extant stream of literature from managing input heterogeneity to reduce outcome variation in healthcare to using technology mediation to reduce outcome variation in healthcare given a certain level of input heterogeneity. This shift in focus has significant practical implications since organizations that deliver health care have limited control on input heterogeneity such as the experience and skill of surgeons and staff. Finally, this field study is the very first of its kind to provide, from a user's standpoint, new and unique insights into (i) the nature of surgeon and surgical team learning with robotic surgery, and (ii) how surgeon and surgical team learning impact the usage of a surgical robot.

Chapter 2:

Predicting Failures of High Tech Innovations-in-Use: Application of Predictive Analytics to Big Data on Market Failures of Medical Devices

2.1 Introduction

Notwithstanding the promise of high tech innovations, there are also perils. In spite of the best efforts and intentions of firms and government regulators, high tech innovations continue to fail while in use in the marketplace. Failures of high tech innovations-in-use in the marketplace manifest as product recalls with significant economic costs, deaths, injury and tarnished reputation. Hence, it is imperative that efforts be directed at predicting failures of high tech innovations-in-use. In particular, the signals of failures of high tech innovations-in-use need to be detected and responded to timely and appropriately.

Fundamental to detecting signals of failures of high tech innovations-in-use in the marketplace is analyzing and interpreting user experienced adverse event reports on the innovations-in-use. Furthermore, inferring from news media and the academic literature, reactions of firms to user reports on adverse events can be biased, i.e. firms systematically under-react or over-react. Hence, the overarching questions that serve as the motivation for this study are: *Can failures of high tech innovations-in-use be predicted by user reports on adverse events related to the innovations-in-use? How do firms react to the user reports on adverse events related to high tech innovations-in-use?*

Our review of the relevant academic and practitioner literature suggests that in spite of the widespread acknowledgment of the significance of being able to predict failures of high tech innovations in use, the literature is largely devoid of studies on this topic. Instead, the published studies have been focused primarily on (i) predicting the success (but not the failure) of high tech

innovations-in-use or (ii) explaining (but not predicting) failures and successes of high tech innovations-in- use. The specific research questions that guide the execution of this study are:

- *Do adverse events reported by users of high tech innovations-in-use predict their failure? If so, how can the precision of such prediction be improved?*
- *Do firms exhibit judgment bias in predicting failures of high tech innovations-in-use? If so, do firms over-react or under-react?*

The above questions inform the identification of the relevant theoretical perspectives, development of the research design, and the choice of research methods for empirical analysis. The theoretical perspectives we use to frame and ground the research questions, and develop testable hypotheses are *signal detection* and *system neglect.* We develop three hypotheses related to prediction, precision of prediction, and consistency of prediction of failures of high tech innovations-in-use.

The high tech sector that serves as the empirical setting for this study is the U.S. medical device industry. This industry is an appropriate setting for the study since "U.S. medical device companies are highly regarded globally for their innovations and high technology products."[1] A recent McKinsey article points out how, in recent years, the medical device industry has experienced "transformative growth and innovation," how "the number and complexity of the devices have risen significantly," and how this industry "has delivered life-enhancing innovations" to the marketplace (Fuhr, George and Pai 2013, p. 1). The same McKinsey article also points out that the medical device industry's "transformational growth and innovation have placed new burdens on the quality systems" with "increasing likelihood of a quality event" with significant economic costs and negative publicity (Fuhr, George and Pai 2013, p. 1). This concern is echoed in a FDA report which indicates that "the annual number of medical device recalls increased by 97 percent from FY 2003 to FY 2012."[2] These recalls can be categorized into recalls of new products (pre-market approval [PMA]) or new versions of existing products (510K).

Each year the FDA receives several hundred thousand adverse event reports of suspected device-associated deaths, serious injuries and malfunctions. The Manufacturer and User Facility Device Experience (MAUDE) database houses the adverse event reports to the FDA by mandatory reporters (manufacturers, importers and device user facilities) and voluntary reporters such as

[1] http://selectusa.commerce.gov/industry-snapshots/medical-device-industry-united-states
[2] http://www.fda.gov/downloads/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDRH/CDRHTransparency/UCM388442.pdf

health care professionals, patients and consumers. MAUDE is a "big" database and a rich source of information with millions of data points generated through reports of adverse events involving the usage of medical devices. MAUDE is also an unstructured database, since the adverse event reports are uncodified text and format free; and a noisy database, since the reports submitted can be incomplete, inaccurate, untimely, unverified, or biased data. The Government Accountability Office (GAO) Report to the U.S. Senate states that the adverse event reports in the MAUDE database are "best used" to "capture qualitative snapshots of adverse events for a particular device or device type, such as the types of malfunctions or clinical events or both associated with the device" and for "*signal detection*, such as for identifying unexpected events associated with a particular device or device type" (GAO-12-816, page. 45).

In this study, we integrate the MAUDE database with other complementary databases to assemble the study database. Using predictive analytics we mine and analyze the data to build a modeling framework for signal detection to predict medical device recalls – i.e. failures of high tech innovations-in-use – and evaluate the existence of judgment bias in making such predictions.

The overarching contribution of this study towards advancing the technology and innovation management literature is in shedding light in to one of the least understood topics in the literature – namely, the prediction of failures of high tech innovation in use. Specifically:

- We demonstrate that the prediction of failures of high tech innovations-in-use is possible with sufficient lead time by analyzing big and unstructured data on user feedback on adverse events in the marketplace; and that the precision of such predictions can be significantly improved by accounting for time-varying covariates related to design, supply chain and manufacturing.

- We present novel and nuanced insights into why firms often fail to correctly detect market signals of failures of high tech innovations-in-use. In particular, we show that the existence of judgment bias leads firms to under-react or over-react to market signals in the form of user feedback on adverse events, and identify factors – such as (i) the severity of the adverse events, (ii) noise-to-signal ratio in the user feedback data stream on adverse events, and (iii) the interaction between (i) and (ii) – that lead firms to under-react or over-react. We find that firms under-react when there is high signal-to-noise ratio in the user feedback data stream on adverse events. The new insight from our study results is that under-reaction

changes to over-reaction under conditions of a high noise-to-signal ratio when the severity of adverse events is high. This change from under-reaction to over-reaction is indicative of the risk averseness of decision makers – i.e., when the severity of adverse events is high, decision makers overweight the risk of potential failures of a high tech innovations-in-use (namely, a medical device recall).

- This study – being the very first to develop a predictive analytic model with a big and unstructured database to predict failures of high tech innovations-in-use and evaluate judgment bias in making such predictions – opens up new empirical research possibilities of conducting technology and innovation management research. The possibilities include: (i) going beyond an orientation of innovation success to *innovation failure*, especially when the innovations are in use in the marketplace; (ii) going beyond an orientation of explanation to *prediction* while accounting for judgment bias; and (iii) going beyond the analysis of structured data-sets containing hundreds or thousands of observations to analysis of *unstructured data-sets* containing millions of observations or more.

The rest of the chapter is structured as follows. In section 2, we discuss the theoretical foundation of this study and the development of the study hypotheses. In section 3, we describe the empirical setting, the data sources and the data for the study. In section 4, we discuss the research design and the predictive analytic methods used in this study. We report the empirical analysis results in section 5. Finally, in section 6, we present our concluding remarks.

2.2 Theoretical Foundation and Hypotheses Development

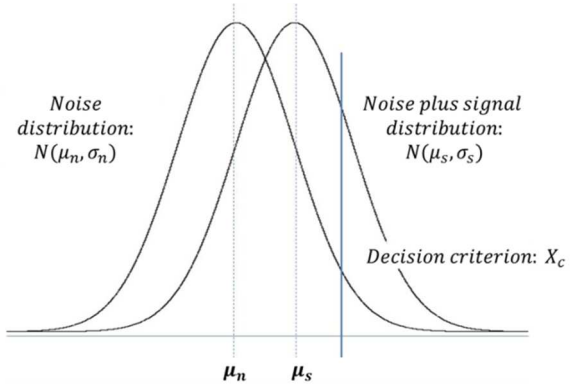
The relevant theoretical perspectives we identified to frame and ground the research questions stated in the earlier section 1, and develop testable hypotheses are *signal detection* and *system neglect.* We develop three hypotheses related to prediction, precision of prediction, and consistency of prediction of failures of high tech innovations-in-use.

2.2.1 Development of Hypothesis 1 (H1) on Prediction

We draw on signal detection theory (SDT) that originated in the field of psychophysics. SDT is used to analyze data stream from a man-made or a natural system. Specifically, SDT categorizes

the data stream into one which is generated either by a known process ("signal") or another which is obtained randomly by chance ("noise") (Salkind, 2007). The basic objective of SDT is to recognize from the input data stream (namely, user reported adverse events data stream related to a high tech innovation in use, as is the context of this study) the presence of a specific state of the system (namely, the failures of a high tech innovations-in-use). A detection method processes the signal and produces a judgment of whether a system level issue is present or not (Harvey, 1992; Wagner et. al., 2001). When the operating condition of a system is perfectly normal, the distribution of the data stream generated will be white noise. However, if a system level issue is present, then the distribution of the data stream will shift and will no longer be white noise. From the standpoint of signal detection theory (SDT), this distribution is the *combined noise and signal distribution.* The core idea of SDT is to be able to detect the presence of such a signal in the data stream. The detectability of the signal depends on the strength of the signal as well as the detection process operationalized by the choice of an appropriate decision threshold, as is depicted in Figure [2.1].

Figure 2.1 Depicting the Basic Concepts of the Signal Detection Theory (SDT)



In Figure [2.1], μ_n is the mean of the noise generating process under normal operating conditions and μ_s is the mean of the signal-plus-noise distribution. A key characteristic of a detection mechanism is sensitivity of the signal detection process expressed below by the "signal-to-noise ratio" (namely, Macmillan and Creelman, 2005; Simpson & Fitter, 1973; Swets, 1986a, 1986b), the inverse of which is the noise-to-signal ratio (which we use later in the chapter to integrate the theoretical perspectives of signal detection theory and system neglect, the other theoretical perspective we draw upon):

$$Sensitivity\ d_a = \frac{\mu_s - \mu_n}{\sqrt{\frac{\sigma_s^2 + \sigma_n^2}{2}}} \qquad \ldots [2.1]$$

For equal variance of noise and signal distribution the measure of sensitivity of the detection process will be:

$$d_a = \frac{\mu_s - \mu_n}{\sigma}, \qquad where \quad \sigma_s = \sigma_n = \sigma \qquad \ldots [2.2]$$

Hence, it can be seen that the sensitivity of the detection process depends on two factors, the signal strength $(\mu_s - \mu_n)$ and the signal-noise variance $\sqrt{\frac{\sigma_s^2+\sigma_n^2}{2}}$. Apart from the signal characteristics, the sensitivity of the detection outcome also depends on the *decision criterion* (X_c) which can be thought of as being analogous to a threshold criterion that represents individual variation across decision units. A decision unit's ability to detect a signal of failure from user feedback on adverse events is manifested through the choice of the optimal decision threshold. So, in summary, the sensitivity of signal detection depends on: (i) signal characteristics expressed through the noise-to-signal ratio and (ii) the decision unit's choice of the decision criterion (X_c).

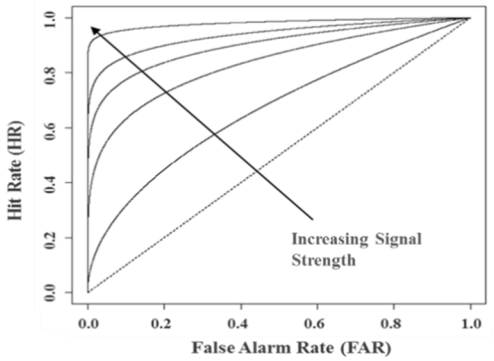
In SDT, two important measures of accuracy of the detection process are "Hit Rate" (HR), the percentage of the signals that have been detected correctly and the "False Alarm Rate" (FAR), the percentage of cases where a signal has been detected when no signal is present in the data. X_c represents a critical decision criterion in a high tech firm.

If the failure of a high tech innovation-in-use is not detected correctly (Miss Rate), then the firm is likely to incur a much higher cost in terms of market compensation or brand equity loss than it would have incurred if the failure was detected correctly (Hit Rate). On the other hand if a firm incorrectly detects a failure of high innovation-in-use where no problems exist (False Alarm Rate), it would unnecessarily increase its cost and harm its immediate stakes and brand equity with long term consequence. Hence, it is important for firms to balance the two costs in the long run, and this is at the heart of strategically managing the risk associated with a high tech innovation-in-use throughout its life-cycle.

In the SDT model, the above dynamics is graphically represented by "Receiver Operating Characteristics" (ROC) curve, which represents the trade-off between Hit Rate (HR) and False Alarm Rate (FAR). The objective of a detection process is to achieve a high hit rate while keeping the false alarm rate to be as low as possible. The accuracy of a signal detection process can be measured by the area under the curve (AUC) of a receiver operating characteristics (ROC) curve.

Higher the AUC the more accurate is the signal detection. As the signal strength (represented by the difference of mean rates of noise-plus-signal and noise-only distributions) increases, the ROC becomes more favorable to correct detection of signals. Figure [2.2] shows the representative ROCs with increasing signal strength. However, the actual accuracy of detection is also determined by the right choice of the decision criterion.

Figure 2.2 Receiver Operating Characteristic (ROC) curves with increasing signal strength



The decision criterion X_c is influenced firstly by the relative frequency of the signal data received and most importantly by the payoff matrix, the relative cost of making errors in detection, i.e., FAR and the relative benefit of making correct detection, i.e., HR. These parameters influence the detection unit, a firm in our case, to use quite different decision criterion. We will discuss this in greater details later in this chapter when we develop the concepts of *judgment bias* using *system neglect theory.* For now, it would suffice to mention that the judgment bias in detecting failure signals is given by equation [2.3] (Macmillan and Creelman, 2005; Wickens, 2002).

$$Judgment\ Bias: \quad c = \frac{False\ Alarm\ Rate - Incorrect\ Rejection\ Rate}{2} \qquad \text{... [2.3]}$$

There are other ways of representing the detection judgment bias using only HR and FAR. However, it can be shown that the two representations are equivalent with some translation and scaling. We choose to use this definition of detection bias for ease of interpretation. With this definition, over-reaction likelihood increases for $c > 0$ and under-reaction likelihood increases for $c < 0$.

By observing the changes in distribution of the data stream on user reported adverse events, firms or regulatory agencies can detect and predict the failure of a high tech innovation in use. Let us denote some appropriate measure of the market signal by f_t. We argue that f_t is a consistent and significant predictor of failures of a high tech innovations-in-use, namely, a medical device

recall:

$$Prob(Failure) = \phi(f_t) + \eta \qquad \ldots [2.4]$$

where ϕ is a risk/hazard function and η is the prediction error. And, hence, we posit the following hypothesis:

> **Hypothesis 1 (H1)**: User reports on adverse events predict the failure of a high tech innovation-in-use.

2.2.2 Development of Hypothesis 2 (H2) on the Precision of Prediction

Adverse event reports by users of high tech innovations-in-use represent a steady stream of data on time-varying performance characteristics of innovations in the marketplace. Apart from the user generated data stream on adverse events, there are other time-varying covariates (factors) that are known to influence the performance of a high tech innovation-in-use in the marketplace. We reviewed the relevant academic and practitioner literature, including reports of regulatory agencies such as FDA, to identify the time-varying covariates that can be categorized into design, supply chain and manufacturing.

Design. Gokpinar et al. (2010) found that product design contributes significantly towards explaining the performance quality of products. In a related study explaining factors related to innovation failure, Gopiknar et al. (2013) show that the number of product development tasks within a given time-period contributes significantly towards explaining new product performance. In a predictive context, the number of products developed within a given time-period is likely to be more relevant. Hence, for our model building purpose, we include the following factors like the number of design changes in a time-period within specific product code, the number of design changes in a time-period across all product segments in a firm, and the number of new products introduced within a time-period across all product segments.

Supply Chain. Marucheck et al. (2011) identify several firm related factors that have the potential to influence new product failure risk, and address how global supply chains influence quality risk of new products, and pose the question if market failure data of new products can be used to predict and make timely recall decisions. Lyles, Flynn and Frohlich (2008) show that changes in supply chain structure can impact new product performance. Anand, Gray and Siemsen (2012) analyze data from pharmaceutical industry and show that changes in supply chain and firm structure significantly influence product quality.

Manufacturing. Thirumalai and Sinha (2011) found that the likelihood of recalls increases with a decreasing focus in manufacturing operations. They also found that the likelihood of recalls decreases with the cumulative recall experience of firms, suggesting a learning effect from the renewed attention and improvements to a firm's manufacturing operations triggered by recalls. In another study, Shah et al. (2013) investigated drivers of product recalls at the level of a manufacturing plant, with the recalls being categorized into design and manufacturing recalls. The results of this study indicate that increasing product variety increases design recalls; increasing product variety also increases manufacturing recalls when equipment utilization is high or operational focus is low; high equipment utilization has a negative effect on both design and manufacturing recalls; and operational focus reduces manufacturing recalls.

Accounting for the above three categories of time-varying covariates along with the adverse event data stream is likely to increase the precision of prediction of failure of a high tech innovation-in-use. Hence, we posit the following hypothesis:

> **Hypothesis 2 (H2)**: User reports on adverse events supplemented by time-varying covariates related to design, manufacturing and supply chain increase the precision of prediction of the failure of a high tech innovation-in-use.

2.2.3 Development of Hypothesis 3 (H3) on the Consistency of Prediction

Beyond precision, consistency or repeatability is a critical consideration in any prediction process. Presence of systematic judgment bias has the potential to adversely influence the consistency of prediction. Firms often exhibit judgment bias in reacting to market signals when making strategic decisions, where prediction is fundamental to decision making. Specifically, firms in the medical device industry – the high tech industry that serves as the empirical setting for this study – appear to exhibit judgment bias in the form of over-reaction or under-reaction to user reported adverse events related to high tech innovations-in-use in the marketplace. Such judgment bias has the potential to negatively impact a medical device firm's ability to predict the failure of a high tech innovation-in-use, i.e., medical device recall.

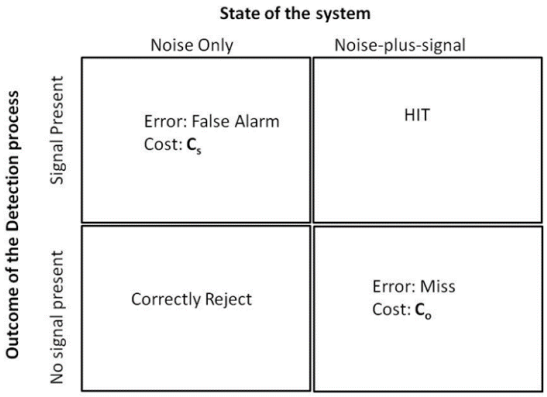
The theoretical perspective of system neglect is founded on the notion that the decision makers exhibit systematic under-reaction in environments of high instability and vice-versa. In this

study, we follow Massey and Wu (2005) to investigate if the data stream on adverse events indicates the presence of systematic judgment bias in detecting the market signals of failure of a high tech innovation-in-use that, in turn, may adversely influence the consistency of prediction. Systematic judgment bias in detection would either mean an over-reaction to a weak signal or an under-reaction to a strong signal. Massey and Wu (2005) experimentally studied the causes of over-reaction and under-reaction in detecting regime shifts. Their study showed that decision makers show systematic bias towards over-reaction in a relatively stable environment and a bias toward under-reaction in a relatively unstable environment. Further, Massey and Wu (2005) concluded that decision makers put excessive attention to the signal and less attention to the system that generates the signal probably due to the relative saliency of the signal with respect to the system parameters that generates the signal. Another significant study in the system neglect literature is Kremer et al. (2011) who investigated judgment bias and signal detection issues in the context of supply-chain forecasting. Kremer et al. (2011) found that decision makers systematically over-reacted to changes in a stable environment and under-reacted to changes in an unstable demand environment.

Now, we synthesize the concepts from the literatures on system neglect and signal detection theory to the third study hypothesis related to consistency of prediction of failures of high tech innovations-in-use. According to SDT, the decision makers reaction is captured through the decision criterion parameter, X_c (see Figure [2.1]). The decision criterion represents the propensity of a decision maker to over-react or under-react, given a signal-noise distribution. The question that would need to be answered is what would be a reasonable choice of the decision criterion for a given noise and signal-plus-noise distribution. Is there an optimality condition for the parameter X_c? Under what conditions would a decision maker adopt a high level for X_c versus a low level for X_c? A relatively high value for X_c would mean under-reaction to signals and vice-versa.

As has been mentioned earlier, there are costs associated with the decision process. If there is a signal in the data that is not detected, then a decision maker incurs an *opportunity cost*. Similarly, if there are no signals but the decision maker wrongly detects a signal, then the decision maker incurs a *sunk cost*. This scenario is pictorially depicted in Figure [2.3].

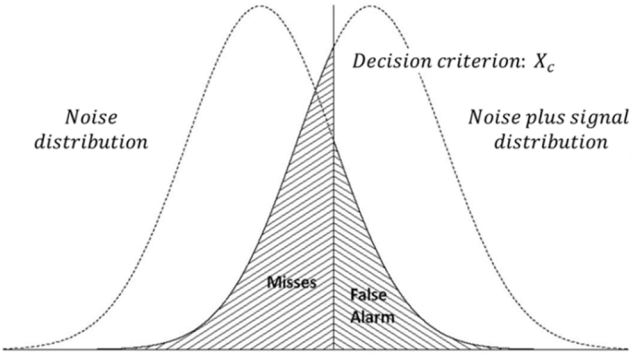
Figure 2.3 Errors of Detection with Associated Cost



A rational decision maker would like to minimize the expected costs associated with the two errors. We assume that there are no additional marginal costs associated with the correct detections. The decision criterion of the rational decision maker can be stated as:

$$\min_{X_c}\{C_o * P(Miss) + C_s * P(FA)\} \quad \ldots [2.5]$$

Figure 2.4 Operationalization of the Signal Detection Theory (SDT) Depicting the Effect of Decision Process (Decision Criterion)



The objective function for the optimization problem is:

$$\min_{X_c}\left\{C_o * \int_{-\infty}^{X_c} N(x|\mu_s, \sigma_s^2)dx + C_s * \int_{X_c}^{\infty} N(x|\,\mu_n, \sigma_n^2)dx\right\} \quad \ldots [2.6]$$

Normalizing the mean of the noise distribution to zero (WLOG), we further simplify the optimal value for the decision criterion as follows. See Appendix A.3 for proof.

$$X_c = \frac{\mu_s}{2} + \frac{\lambda\sigma^2}{\mu_s} \qquad \ldots [2.7]$$

where,

$$\lambda = \log(\beta) = \log\left(\frac{C_s}{C_o}\right) = \log\left(\frac{Sunk\ Cost}{Opportunity\ Cost}\right)$$

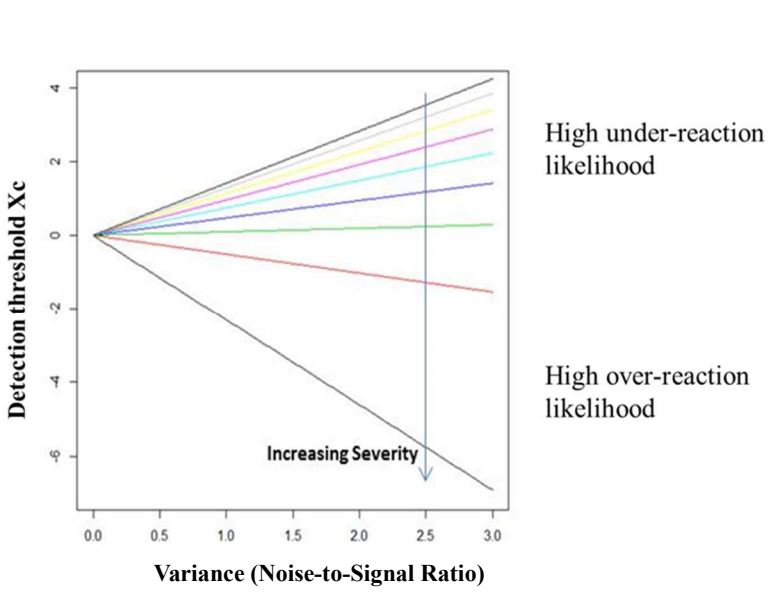
From the above relationship, we can conclude the following:

(i) Holding all other factors a constant, X_c increases as signal strength increases. In other words, as the detectability increases as the signal becomes more precise.

(ii) Again, holding all other factors a constant, the X_c increases as λ increases. Since the parameter $\lambda = log(Sunk\ Cost|False\ Alarm) - log(Opportunity\ Cost|Miss)$, as the sunk cost increases in relation to the opportunity cost, X_c increases. Similarly, as the opportunity cost increases in relation to the sunk cost, X_c decreases. However, pre-facto, neither the sunk cost not the opportunity cost is precisely known to any decision maker. Decision makers interpret values of this cost from the characteristics of a high tech innovations-in-use or the characteristics of an adverse event. As the severity of adverse events increases, the interpreted future cost estimates increase. Due to inherent risk averseness of decision makers, a disproportionately high cost may be assigned to adverse events with high severity measure. This would lead decision makers to choose a low decision threshold, which would result in over-reaction to market signals. On the other hand, a low severity measure may lead to choose a high decision threshold leading to under-reaction.

(iii) X_c increases (case of under-reaction) with system variance σ^2 when λ is positive, i.e., when the perceived sunk cost of a false detection is more than the perceived opportunity cost of no detection. However, X_c decreases (case of over-reaction) with system variance σ^2 when λ is negative, i.e., when the perceived future opportunity cost of no detection is more than the perceived sunk cost of false detection. As mentioned, higher severity leads to a perceived higher future opportunity cost which may be disproportionate to the actual cost given inherent risk averseness of decision makers. Thus, we may expect over-reaction bias to increase with increased severity and increased perceived opportunity cost of not detecting a credible signal. Incorporating this in the equation for decision criterion, we can depict the above conclusions pictorially in Figure [2.5].

Figure 2.5 Depicting the Operationalization of the Concept of System Neglect and the Resulting Biases



Building on the depiction in Figure [2.5], we posit the following set of hypotheses:

> **Hypothesis 3A (H3A)**: Noise-to-signal ratio (in H2) is associated with systematic bias in the prediction of failure of a high tech innovation-in-use. High noise-to-signal ratio is associated with under-reaction bias and low noise-to-signal ratio is associated with over-reaction bias.
>
> **Hypothesis 3B (H3B)**: Severity of an innovation failure (in H2) is associated with systematic bias in the prediction of failure of a high tech innovation-in-use. High severity is associated with over-reaction bias and low severity is associated with under-reaction bias.

Note, we use noise-to-signal ratio, the inverse of signal-to-noise ratio, to indicate high variance of the noise-plus-signal distribution. Noise-to-signal ratio makes the interpretation of H3A and H3B direct and simple, and in keeping with the system neglect literature.

2.3 Empirical Setting and Data

As mentioned earlier, medical device industry is the empirical setting for this study. The data collection strategy for the study was guided by the data needed to test the three hypotheses posited in the earlier section. The obvious data needed to test the hypotheses are data on device recalls, adverse events data and data on the relevant time-varying covariates related to design, manufacturing and supply chain corresponding to the devices recalls in the study sample. In addition, we reviewed the extant academic and practitioner literature to identify variables, for which data would be needed, to inform the specification of the models to be estimated for testing the hypotheses.

Specifically, we inferred from Marucheck et al. (2011) that the maturity of the underlying technology of a product, i.e., technology life-cycle, contributes significantly towards explaining product recalls. In the context of medical device firms, Thirumalai and Sinha (2011) found that as firms broaden their product scope – i.e., breadth and depth of products – the likelihood of recalls increases. Insights from the analytical models developed by Banker, Khosla and Sinha (1998) linking competitive intensity and quality suggest that an increase in competition intensity in an industry can lead to a lower average industry quality level. Given that the impact of competitive pressures can be evident at the level of an individual product as well as at the usage category level, in this study, we consider product level competition as well as the usage segment level competition as predictors of medical device recalls. Finally, our review of industry reports, such as FDA report on medical device recalls, surfaced factors that are known to explain systematic variations in device recalls. These variables are categorical in nature and include the following: regulation type [pre-market approval (PMA) – approval of new devices; and 510K – approval of new versions of existing devices]; device class, usage class, implant vs. non-implant; product class and manufacturer.

As mentioned earlier, the primary data-set we are using for this study is the Manufacturer and User Facility Device Experience (MAUDE) data-set of the Food and Drug Administration (FDA). Each year, FDA receives several hundred thousand adverse event reports of suspected device-associated deaths, serious injuries and malfunctions. The MAUDE database houses the adverse event reports to FDA by mandatory reporters (manufacturers, importers and device user facilities) and voluntary reporters such as health care professionals, patients and consumers. MAUDE is a “big” database and a rich source of information with millions of data points generated

through reports of adverse events involving the usage of medical devices. MAUDE is also an unstructured database, since the adverse event reports are uncodified text; and a noisy database, since the reports submitted can be incomplete, inaccurate, untimely, unverified, or biased data.

Below are two examples of user reports submitted to FDA. The reports are illustrative of the *ad verbatim* text in the reports, and, in turn, the unstructured nature of the data in the MAUDE database. The first report pertains to the death of a patient and the second report pertains to serious injury to a patient requiring hospital readmission and surgical intervention. To maintain neutrality and confidentiality, the user names and specific device brand names are not included (have been disguised) in the reports.

(a) ***A Death Report.*** *" ... Patient was in for laparoscopic roux-en-y gastric bypass, liver biopsy (this is normal. They all get a liver biopsy), and gastrostomy. Ten days later, the doctor informed me that this patient passed away over the weekend. Autopsy revealed failed staple lines in two places (pouch and distal stump). He stated that during the surgery he had no indication of a stapler problem. Four days after the initial surgery, the patient went to a different hospital with c/o abdominal pain and being clammy. The er called the doctor and he had three concerns: pe (pulmonary embolism), mi or a leak. The er doctor said they ruled out pe and mi. They then decided to do a ct to rule out a leak. This was the last that the surgeon heard from the doctors at the er. Apparently, the ct was read as negative. The patient was sent home. A few days later the patient was discovered dead in his home. Reportedly he passed away a day before he was discovered. The stapler device used was an ...xxx ... endoscopic cutter-straight ref ...xxx…. This device cuts and staples when loaded with staple product (it does not come pre-loaded. There are multiple sizes available to fit the device). The staples used in the case were …. ..."*

(b) ***A Hospitalization Report.*** *" ... As per morning report, a pt who had an iabp placement in the cath lab in ..., had to return within a short time for emergent removal. Fluoroscopy revealed that the balloon was in good position, but not inflating. As per cardiology, there was vascular calcification noted adjacent to the distal edge of the balloon which most likely damaged the balloon and caused a tear in the balloon and helium leak. Cardiology additionally noted*

that there was blood within the balloon confirming that it had ruptured. Actions taken: ruptured iabp catheter sequestered. Additional actions needed: incident report submitted. … xxx …to be notified and iabp catheter returned for inspection…"

Apart from the MAUDE and patient databases, we assembled data from several other sources. The main databases are: (i) the recall database; (ii) the 510(K) and PMA databases – where 510(K) is premarket notification, i.e., the approval process of new models of existing devices, and PMA is pre-market approval, i.e., the approval process of new devices; (iii) the facility registration database; and (iv) COMPUSTAT database from Wharton Research Data Services (WRDS). As mentioned earlier, the MAUDE database and the patient database comprises of data on adverse incidents reporting from manufacturers, distributors and user facilities like hospitals, dispensaries, clinics, healthcare professionals and individual patients. The patient database contains device reports from the field where patients were involved in adverse incidents. The patient reports are classified based on the actual or potential harm to patients involved using a multipoint classification scheme starting from as severe as death to no effect. Both the databases contain information on the products identified by a three letter acronym or product code, manufacturer name, date of event, date of reporting, identification of the reporting agency, location, text description of the event details including effects and remedial measures undertaken.

We collected data from 1998-2010 from all the databases. The adverse event and the patient database combined have upwards of three million data points. Apart from being large, another feature of this database is that the data is unstructured and has missing data. Much of the valuable information is in text format. FDA classifies devices primarily using two classification schemes, i.e., the usage class and the complexity class. There are 19 usage classes listed, as per the FDA classification, such as cardiovascular, surgical, anesthesiology, clinical chemistry, hematology and orthopedics. This classification scheme represents the intended usage of the device and also serves as a classification of devices for regulation and review. The second classification scheme is based on a combination of the complexity of a device and the potential risk associated with a device in case it fails. Based on this scheme, devices are classified into one of the three risk classes, i.e., Classes I, II and III. Class I devices represent the lowest potential risk for harm and are simpler in design than the other two classes. This class of devices is usually subject to only general controls from a regulation viewpoint. Class III devices pose the highest potential risk to patients in case of failure. Also, these devices are generally more complex in design as compared to the other two

classes. Examples of Class III devices include replacement heart valves, cardiac defibrillators, robotic surgical instruments and implanted cerebellar stimulator. These classes of devices are subject to general as well as several special controls based on the intended usage of the devices. In between the Class I and Class III are the Class II devices which are more complex than the Class I devices but are less complex than Class III devices. This class of devices is also subject to both special and general controls. This classification information is critical for our analysis, since the classification scheme controls for the complexity and risk class of devices. However, many of the databases like the recall database and the adverse event database either do not have this classification or many of the classification information are missing. Such gaps in the device classification pose a serious challenge in building the analysis models since it is critical to control for the complexity and the inherent risk measure of the devices.

According to Lynch (2008, p. 28), in an article published in the *Nature* journal, data can be "big by being of lasting significance." To that end, the MAUDE database along with the other relevant databases such as the recall database, the approval database and the surveillance database have the potential to reveal critical relationships between relevant variables which can inform the development of an ongoing monitoring and risk assessment tool for medical devices throughout their lifecycles.

2.4 Research Design and Methodological Foundation

The research design and the choice of research methods were guided by: (i) the hypotheses to be tested and (ii) study data-set to be analyzed that included big and unstructured data on adverse events related to medical devices. Given that the study hypotheses are related to prediction, precision of prediction, and consistency of prediction of failures of high tech innovations-in-use, predictive analytics is an appropriate methodological foundation for this study. Following Shmueli and Koppius (2010), we apply predictive analytics for building and assessing models aimed at making empirical predictions. This application involves two key considerations: (i) choice of empirical predictive methods that include statistical models, data mining algorithms (namely, classification trees and neural networks) and ensembles (i.e., averaging across multiple models); and (ii) evaluation of predictive accuracy of the estimated models.

As a methodological approach, predictive analytics makes it possible to identify complex relationships and patterns in a big and unstructured data-set, as is the data-set for this study. Predictive analytics is different from explanatory modeling in two significant ways: (i) whereas

explanatory models are based on underlying causal relationships between theoretical constructs, predictive models rely on associations between measurable variables, and (ii) whereas explanatory models seek to minimize model bias (i.e., specification error) to obtain the most accurate representation of the underlying theoretical model, predictive models seek to minimize the combination of model bias and sampling variance.

Notwithstanding the potential of predictive analytics, its application as a methodology in conducting technology and innovation management research is in nascent stages. Below are the steps in executing the research design of the present study where we demonstrate the application of predictive analytics to data-set that includes big and unstructured data-set on adverse events reported by users of medical devices to test hypotheses related to prediction, precision of prediction, and consistency of prediction of medical device recalls.

2.4.1 Classification of Devices

Data classification and organization represents an important step towards building a robust predictive model. While the data in the recall database, the adverse events and patient databases are unstructured and noisy, the data in the approval database are relatively more organized and complete. The approval database also contains fairly complete device classification information. Therefore, the task of classification for the data points where the classification information is not available is to primarily link the recall and the adverse event databases with the approval database. Linking would mean identification of a suitable primary key that is common between all the relevant databases especially in MAUDE and patient databases. However, no such codified key exists between all these databases other than the device name and manufacturer name combination. The problems with the device name and manufacturer name combination are that neither the product name nor the manufacturer name has a standard codified format in any of the database. We used the generic device names from the approval databases and the device names from the other databases to classify the databases and link the databases. However, since there is no direct way to perform classification, we used a modified text mining method to perform the classification task. Apart from the issue of data classification and linking, many of the key information related to several variables like severity, device failure type, causes for failure and device description are embedded in plain text paragraphs. Extraction and codification of such information also required text analytics.

Text classification is the task of identifying what class among a finite number of defined

classes a group of words belongs to (Madsen, Kauchak and Elkan, 2005). It is common to represent a string like a device name as a collection of words, or in text mining terminology a "bag-of-words." The basic algorithm we used is a Bayes classification approach using a Dirichlet distribution for the classification task (a modified Latent Dirichlet Allocation – LDA). For training the Naïve Bayes classifier, we used the device names from the approval database to create a frequency table. The rows of the table are the unique words from the device names and the columns of the table represent the different usage class of the devices. The frequency count for each word along the row would sum up to the total number of times that particular word appears in all the device classes. Individual values of the cells represent the number of times the particular word appears in a particular class. Let f_{ij} represent the number of times word w_i appear in class c_j. Naïve estimates of some of the probability values would be:

$$P(w = w_i|c = c_j) = \frac{f_{ij}}{\sum_i f_{ij}}; P(w = w_i) = \frac{\sum_j f_{ij}}{\sum_j \sum_i f_{ij}}; P(c = c_j) = \frac{\sum_i f_{ij}}{\sum_j \sum_i f_{ij}} \quad \text{... [2.8]}$$

The classification problem is to find out the probability that a device belongs to a class c_j given that it contains a set of words $W = \{w_1, \dots, w_k\}$. In text classification, the probability distribution of a word being in a specific class out of several possible classes is modeled as a multinomial or a Dirichlet distribution. A Dirichlet distribution has been proven to perform better in the presence of sparseness of word distribution. Also, a Dirichlet distribution is more appropriate when the class distributions are not even, i.e., when a few classes appear much more in the sample as compared to the rest of the classes. We assume that the probability that a word with frequency f_{ij} belongs to a specific class c_j is distributed according to the Dirichlet process $Dir\{f_{ij}|\theta_j\}$. The full conditional probability will then be given by the Bayes formula [9] below using a standard Dirichlet distribution for text classification (Madsen et al. 2005; Nigam et al., 2000; Blei et. al, 2003). See Appendix A.4 for proof.

$$P(c = c_j|w = \{w_1, \dots, w_k\}) = \frac{\left\{\frac{1}{B(\boldsymbol{f})}\prod_{i=1}^{k}\frac{\left(\frac{f_{ij}}{\sum_i f_{ij}}\right)^{f_{ij}}}{f_{ij}!}\right\}\frac{\sum_i f_{ij}}{\sum_j \sum_i f_{ij}}}{\prod_{i=1}^{k}\frac{\sum_j f_{ij}}{\sum_j \sum_i f_{ij}}} \quad \text{... [2.9]}$$

The 510K and the PMA approval databases generated more than 150,000 unique words. Many of the words were generic words such as "and," "the," and "of." To test the algorithm, we split the data into 80% train-set and 20% test-set. A first run of the Naïve Bayes classification

scheme led to a correct classification rate of 66%. Some words are more informative than others. Clearly, there are some key words which are important. To improve the accuracy of classification we needed to screen the words based on the information content. We calculated the entropy value for each unique word using the Shannon entropy equation [2.10]:

$$H(w_i) = -\sum_j \frac{f_{ij}}{\sum_i f_{ij}} \log\left(\frac{f_{ij}}{\sum_i f_{ij}}\right) \quad \ldots [2.10]$$

Words with high information content about class belongingness would have low entropy. So, we removed words from the train-set as well as the target classification-set using a threshold function λ such that all words with entropy value greater than the threshold parameter are not considered. Also, words with very low frequency count $\sum_i f_{ij}$ were removed from the list using a threshold value β. These are mostly words which are very specific to some device or firm. We chose the threshold values λ and β so as to minimize the classification error with the test set using a grid search algorithm in R (www.cran.r-project.org). After calibrating the algorithm with the train set and test set and obtaining the optimal calibration of threshold parameters λ and β, we trained the algorithm with the complete data from the approval database. Then the model was run on the target sets of device names from the recall, adverse event and patient database. For the adverse event and patient databases, only those data points were reclassified where the classification was not available in the original database. On the 20% hold-out sample or test-set we achieved $(92.5 \pm 1.4)\%$ classification accuracy. The data organization process and the resulting data-sets are presented in Table [2.1]. An algorithm for the classification method (A*lgorithm 1*) is presented in Appendix A.1.

Table 2.1 Source and Description of Databases for the Empirical Analysis

	Data Description	Data Source
1	Recalls	FDA Recall Database
2	Adverse Events (Event type, Date, Severity, Manufacturer, Device, Classification, Implant Code).	FDA manufacturer and user reported adverse event database (MAUDE) and patient database. About 3 Million usable data-points. Fairly unstructured with missing data and misclassification.
3	Device approval (Device class, device type, classification, date, manufacturer, etc.)	FDA Approval Database, Compustat database
4	Manufacturer per device class, models per device class	FDA approval database, Compustat database
5	Manufacturer location, Outsourced manufacturer location, Supplier Location	FDA user and manufacturer facility registration database; FDA supplemental approval text files
6	Supplier changes; manufacturing process changes; Manu*facturing location changes	FDA user and manufacturer facility registration database; FDA supplemental approval text files
7	Device classification	FDA approval database; Device classification algorithm.
8	Device approval / Supplemental Approval type, date and location	FDA 510(k) and PMA / PMA supplemental database

2.4.2 Variable Generation and Variable Description

2.4.2.1 Response Variable

The unit of analysis for the empirical investigation in this study is a failure of a high tech innovation-in-use – i.e., a medical device recall. Every medical device model in the database is identified by a unique identifier which is a combination of the device code and the firm code. By way of an illustrative example, a Medtronic defibrillator model is identified by the combination of the general device code for defibrillators and the firm code for Medtronic. We did not distinguish between versions of the same device since that information is already captured in the number of FDA approvals for the device. The response variable is a binary recall variable with the unit of time being a quarter (three months). If a device has been recalled within a quarter, then the value of the response variable is 1 for the quarter. Otherwise, the value of the response variable is zero.

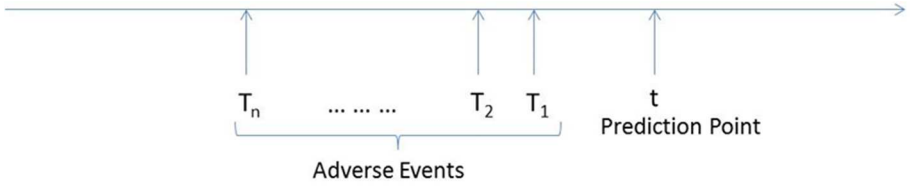
2.4.2.2 Predictor Variables

The first step in our problem formulation is to specify the correct signal variable that firms can use for detecting the failure of a high tech innovation-in-use and specify the failure to a source such as design, manufacturing or supply chain. To that end, let us consider the scheme of adverse event data distribution on a normal time-line shown in Figure [2.6]. A simple and intuitive measure of the variable would be the adverse event rate given by the following expression [2.11]:

$$Adverse\ Event\ Rate\ (R) = \frac{Number\ of\ Adverse\ Event\ Reportings\ (N)}{Time\ Period\ (T)} \qquad \ldots [2.11]$$

While simple and intuitive, the problems of using the above expression [2.11] are the following: Firstly, this variable ignores the influence of any reports outside the time-period even if the reports may fall very close to the boundary of the time-periods. Secondly, the effectiveness of R as a signal specification would be heavily influenced by the choice of the time-period, T. Increasing T would reduce variance of the detection process but would increase the bias of the detection process due to aggregation effect and vice-versa. Thirdly, within the time-period, T, all reports will have the same weights and the local distributions would be ignored in the specification of R.

Figure 2.6 Representative Timeline of Adverse Event Reporting



Similarly, the simple cumulative frequency count, as has been used in past studies (namely, Thirumalai and Sinha 2011) is not appropriate for predictive purpose. While cumulative count may be appropriate for an explanatory model, cumulative count can cause problems for a predictive model. The cumulative frequency count beyond any time t would either remain the same or increase with time, and, hence, would mean that the likelihood of failure would probably remain the same or monotonically increase with time. From a prediction standpoint, this is not informative. Instead, we can consider the histogram of blocked frequency count on the time-line at some time-

period. This measure of adverse events has a better likelihood of being more informative for a predictive model since it contains the information of the distribution of events albeit at a blocked level. However, this would be sensitive to the choice of time-interval for blocking. A narrow time-interval would have higher information on the time distribution of the events but would increase the variation of data distribution leading to a consistency issues with the prediction model. On the other hand choosing a wide interval would reduce the information content of the adverse event measure. These issues can be addressed by considering a weighting scheme for the frequency count which accounts for the relative timing and temporal concentration of the adverse event data. Here we use a kernel weighted frequency count (Kass-Hout and Zhang, 2010) of the adverse event data given by equation [2.12] as the primary predictor.

$$f_t = \left(\frac{1}{Nh}\right)\sum_{i=1}^{N} I_i * K\left(\frac{t - T_i}{h}\right) \quad \ldots [2.12]$$

In equation [2.12], h is the bandwidth of the kernel function. While there are many possible choices of a kernel function like the triangular, biweight, triweight, epanechikov and Gaussian kernels, we selected the Gaussian kernel function for its smoothness property as well its ease of interpretation. So, the first predictive variable for our model is the *Kernel.Weighted.Adverse.Events*. We also created lagged variables for the *Kernel.Weighted.Adverse.Events* with 100-day, 200-day and 300-day lags. This is because it is likely that, in reality, there would be a lag in information flow to the firm and its being able to detect the failure. The idea of a predictive model is to create variables that are likely candidates of being significantly predictive and then select the right variables from the data using an appropriate selection scheme.

While the above measure is informative in terms of the frequency of adverse events, it misses out information on the severity of the adverse event. To capture the severity of adverse events, we used the severity data from the FDA's patient database. FDA classifies the severity of adverse events based on the impact on patients involved in the case of an adverse event. The data on the severity of adverse events were available in codified format in the database, however, still a large part of this data were recorded in text format. We used a text key word search method to classify the events. The method used was very similar to the text classification algorithm described earlier in the context of the device classification. Next, following FDA's coding scheme, we coded the severity of adverse events on a 5-point ordinal scale with "no harm" = 0, "minor injury" = 1, "injury" = 2, "severe injury" = 3, and "death" = 4. Using this definition and denoting the severity by S_i we created the following two severity weighted adverse event variables, namely, a linear

weighted kernel and an exponential weighted kernel:

$$Linear.Severity.Weighted.Kernel\ f_t = \left(\frac{1}{Nh}\right)\sum_{i=1}^{N} S_i * I_i * K\left(\frac{t-T_i}{h}\right) \quad \text{... [2.13]}$$

$$Exponential.Severity.Weighted.Kernel\ f_t = \left(\frac{1}{Nh}\right)\sum_{i=1}^{N} \exp(S_i) * I_i * K\left(\frac{t-T_i}{h}\right) \text{... [2.14]}$$

Apart from the above, we also created kernel weighted variables for the three categories of time-varying covariates we have identified, namely, design, supply chain and manufacturing. We present a complete description of the variables in Table [2.2].

Table 2.2 List and Description of Variables Generated for Predictive Model Building

	Variables	Description
Primary Predictor: User Reported Adverse Events (MAUDE)		
1	Maude.Kernel.000	Simple adverse event kernel density estimate
2	Maude.Kernel.300	Simple adverse event kernel density estimate with 300 days lag
3	Exponential.Weighted.Kernel.000	Severity weighted (exponential weights) kernel density estimates of the adverse events
4	Exponential.Weighted.Kernel.300	Severity weighted (exponential weights) kernel density estimates of the adverse events with 300 days lag
Time-Varying Covariates related to the precision of prediction		
	Design	
5	Product.Dev.300	Number of design changes in a time-period (quarter) within a specific device category (kernel density estimate) of a firm with 300 days lag
6	Firm.Product.Dev.100 Firm.Product.Dev.300	Number of design changes in a time-period (quarter) across all device categories of a firm with 100 and 300 days lag
7	Firm.Innovation.100 Firm.Innovation.300	Number of new devices introduced by a firm within a time-period (quarter) across all device segments with 100 and 300 days lag
	Supply Chain	
8	Supply.Chain.Change.000 Supply.Chain.Change.100/300	Number of changes in supplier components registered within a time-period (quarter) for a device code with no lag, 100 days lag and 300 days lag (density estimate)
9	Firm Global Sourcing Index	Proportion of firm product (device) changes that included shifting sourcing abroad (outside US) for a time-period (quarter)
	Manufacturing	
10	Manufacturing.Change.000 Manufacturing.Change.100/300	Number of manufacturing process changes implemented within a time-period (quarter) for a device code with no lag, 100 days lag and 300 days lag (kernel density estimate)
Control Variables		
11	Technology.Life.Cycle	Number of substantially equivalent priors of a device code
12	Firm.product.Scope	Entropy Index of firm's number of models across device segments
13	Product.Competition	Number of competing models in a device segment
14	Industry Competition	Number of competing players within a usage class (Industry segment)
15	Regulation Type	Categorical variable of approval type of a device: PMA and 510K
16	Device Class	Categorical variable for complexity class of device Class I, II and III
17	Usage Class	Categorical variable of device usage classes (20 classes)
18	Implant	Binary variable (1= if the specific device is an implantable device; 0 = otherwise)
19	Product class	Product category control (732 different device codes in the study sample)
20	Manufacturer	Manufacturer control (105 manufacturers in the study sample)
21	Time	Age of the device in number of quarters

2.4.2.3 Variable Selection for Model Building

Variable selection is a key step in predictive model building. The criterion for selecting the right set of variables is predictive association of a variable with the response variable (Shmueli, 2010). From a number of likely predictors generated from data, a subset of variables are chosen to be included in the actual prediction stage. This is important from the point of view of predictive model parsimony – i.e., to avoid over-fitting and reduce prediction error. In a study on prediction of hospital readmission using patient level data, Kansagara et. al. (2011) demonstrated the importance of variable selection in achieving high predictive accuracy. Kansagara et. al. (2011) studied a number of statistical models for variable selection and identified that various methods of variable selection have relative strengths and weaknesses. Kansagara et. al. (2011) concluded that a triangulation of several methods can lead to a good variable subset selection strategy for predictive model building.

One of the most commonly used approaches to variable selection is a shrinkage method in the form of a linear model. This method is called the LASSO or the Least Absolute Shrinkage and Selection Operator (Zou, 2006; Tibshirani, 1996a). LASSO bounds the absolute sum of the coefficients to an upper limit. By doing so, usually some amount of bias is sacrificed but the model variance decreases improving the overall prediction accuracy of the predictive analytic models (Tibshirani 1996a & 1996b). Also, this shrinking process helps in the interpretation of results by identifying a subset of variables that are the strongest predictors of the response variable. While this method is efficient, LASSO has two limitations for our purposes. First, LASSO tends to over-shrink the selected subset of variables. This can lead to omission of some important variables for prediction purpose. While we do not want to include irrelevant variables, we also would not like to omit important ones. Second, the LASSO model uses a linear modeling setup. In predictive modeling, nonlinear relationships and variable interactions are important considerations. Another approach commonly used for variable selection is a step-wise regression using information criterion such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) to select the best subset at every step of the estimation method. The stepwise regression method we used is Generalized Linear Mixed Model (GLMM) estimation. While as an estimation method, GLMM is intuitive and simple, the model estimation does not always lead to the selection of a subset of variables that is globally optimal. The selection of the variable subset depends on the order in which the variables are included in the estimation model. Also, like LASSO, GLMM is a linear model. To account for the nonlinearities in the data, the variable selection method we used is the stepwise

Generalized Additive Models (GAM). GAM uses nonlinear additive splines to perform model estimation. The estimation of GAM also does not lead to the selection of a subset of variables that is globally optimal.

We followed the principle of triangulation to select the best subset of variables for predictive model building. This ensured parsimony as well as inclusion of the likely important predictive variables. Specifically, we selected the set union of the variables subsets selected by each of the three methods discussed above, namely, LASSO, GLMM and GAM. Note, the variable selection with LASSO was performed using a variant of LASSO called the adaptive LASSO which has been shown to have higher consistency properties in estimating model parameters than the ordinary LASSO (Zou 2006). Since the response variable is binary, we conducted L_1 penalization of the negative log-likelihood of the logistic response distribution. The results of the model selection stage are presented in Table [2.3]. As can be seen in Table [2.3], LASSO, GLMM and GAM selected a number of common variables and a few unique variables. The triangulation helped in removing unnecessary variables while selecting the variables that have either a linear or a non-linear predictive relationship with the response variable.

Table 2.3 Regression Model Estimation Results for Variable Selection

	LASSO	GLMM	GAM
Primary Predictor: User Reported Adverse Events (MAUDE)			
Maude.Kernel.000	0.239	0.602(0.015)*	7.335(0.001)***
Maude.Kernel.300	-	0.544(0.0002)**	8.694(0.002)**
Exponential.Weighted.Kernel.000	0.981	1.021(0.0001)***	-
Exponential.Weighted.Kernel.300	-0.691	-0.801(0.0001)***	8.767(0.000)***
Time-Varying Covariates related to the precision of prediction			
Design			
Product.Dev.300	-	1.568(0.001)***	5.161(0.000)***
Firm.Product.Dev.100	-0.369	-0.514(0.01)**	-
Firm.Product.Dev.300	-	0.491(0.02)*	8.867(0.005)**
Firm.Innovation.100	0.905	0.321(0.02)*	8.825(0.000)***
Firm.Innovation.300	-	0.307(0.003)**	1.000(0.0001)***
Supply Chain			
Firm Global Sourcing Index	0.209	0.298(0.02)**	8.842(0.000)***
Supply.Chain.Change.000	0.391	0.297(0.1)	-
Supply.Chain.Change.100	-	0.675(0.01)**	8.839(0.000)***
Supply.Chain.Change.300	-0.413	-0.686(0.004)**	8.885(0.0001)***
Manufacturing			
Manufacturing.Change.100	-	-	-
Manufacturing.Change.300	-	-	8.921(0.001)***
Control Variables			
Technology.Life.Cycle	-0.200	-0.354(0.001)***	8.359(0.000)***
Firm.Product.Scope	0.045	0.211(0.001)***	8.839(0.000)***
Product.Competition (Within Product Category)	1.746	2.961(0.000)***	8.618(0.001)***
Industry.Competition (Within Usage class)	0.044	0.033(0.000)***	8.842(0.005)**
Regulation.Type.PMA	S	3.435(0.000)***	S
Device.Class	S	S	S
Usage.Class	S	S	S
Implant	S	0.363(0.000)***	S
Product.Class (Random Effect)	NS	SD: 0.123(0.4)	-
Manufacturer (Random Effect)	NS	SD: 0.254(0.4)	NS
Time (Age of product)	0.0007	0.0002(0.001)***	8.921(0.000)***
Estimation sample properties			
Number of product groups	*732*	*732*	*732*
Number of devices	*5,873*	*5,873*	*5,873*
Sample size for estimation	*101,830*	*101,830*	*101,830*
Sample size for test	*25,461*	*25,461*	*25,461*

Notes:

1. *The table shows the slope parameter estimates with their respective p-values in braces.*
2. *Generalized additive models using smoothed splines do not have linear slopes. The reported parameters are estimated degrees of freedom for the polynomial estimates and hence, are not directly interpretable or comparable with the linear models.*

3. *All models were run with forward selection other than LASSO.*
4. *Errors were clustered on the product codes. Bootstrap based robust errors were used for the GLMM model.*
5. *The models were run with randomly selected subset of 80% of the total sample. Random sub-setting was clustered on the devices.*
6. *S: Significant and NS: Non-significant. Due to multiple parameters on factor variables, the estimates are not reported.*
7. *Significance codes: <0.001 '***', <0.01 '**', <0.05 '*', <0.1 '+'*

2.4.3 Predictive Model Building

In predictive model building, the choice of model through a model selection process is important and this process is distinctly different from explanatory model building in many ways (Shmueli, 2010; Chen et.al., 2012; Kansagara et. al., 2011; Zhu and Davidson, 2007; Shmueli and Koppius, 2010). Firstly, explanatory models require easy interpretability with explicit variable significances. In contrast, predictive modeling focuses on predictive accuracy of a hold-out validation sample separate from the data sample used for model estimation purpose. Secondly, unlike explanatory models, predictive models do not estimate just the expected mean of the response, but the complete response distribution over the data domain. Hence, in predictive modeling, where the priority is generating accurate and consistent predictions of new observations, the range of methods include not just statistical models (parametric or non-parametric nonlinear) but also data mining and machine learning based models (Shmueli, 2010; Breiman, 2001).

In light of the above discussion, the objective of our predictive model building effort was to achieve a high predictive accuracy while keeping the predictive method as simple as possible. Specifically, we started predictive modeling by estimating GLMM first, followed by GAM, then moving on to machine learning methods, and finally to ensemble class methods. We checked the improvement in predictive accuracy at every step. The machine learning methods we used are: Artificial Neural Network (ANN), Support Vector Machine (SVM) and Naïve Bayes models. Applications of these methods for predictive modeling are common in the academic and practitioner literatures. The ensemble class models we used are decision tree based Random Forest and boosting algorithms, both of which have been shown in literature to be accurate and robust for predictive modeling with unstructured data characterized by high noise and variable interdependence. The ensemble class methods combine output of several estimation models to generate predictions and hence, are robust to noise and variable interdependence.

In predictive analytics, the choice of the model is dependent on the data. There is no single

model that can perform equally well on all data-sets. The performance of the models is dependent on the nature of the data, data dimensions, presence of noise in data, and the complexity of interdependence of the variables. Hence, we present and compare prediction results of the models mentioned above and described in Appendix A.2 to choose the best predictive model. Specifically, we followed the succession of steps shown below to estimate the equation corresponding to each of the steps for all the prediction models. The ROC curves for all the four equations are generated and the areas under the curves are compared for evaluating the improvement in predictive accuracy at each of the successive steps. Following the literature, we randomly split the sample data into eighty percent train set for model estimation and twenty percent validation or test set for generation of the ROC curves for evaluating the predictive accuracy of the estimated models (cf. Verikas, Gelzinis and Bacauskiene 2011, Piao et al. 2012, Ghasemi and Tavakoli 2013)

$$Step\ 1\text{: } P(Recall) \sim F(Maude.Kernel, Control\ Variables)$$

$$Step\ 2\text{: } P(Recall) \sim F\begin{pmatrix} Maude.Kernel, Control\ Variables, \\ Design\ Covariates \end{pmatrix}$$

$$Step\ 3\text{: } P(Recall) \sim F\begin{pmatrix} Maude.Kernel, Control\ Variables, \\ Design\ Covariates, Supply\ Chain\ Covariates \end{pmatrix}$$

$$Step\ 4\text{: } P(Recall) \sim F\begin{pmatrix} Maude.Kernel, Control\ Variables. \\ Design\ Covariates, \\ Supply\ Chain\ Covariates. Manufacturing\ Covariates \end{pmatrix}$$

2.5 Results and Discussion

2.5.1 Testing Hypothesis 1 (H1) on Prediction

Table [2.3] reports the results of the models estimated for variable selection, namely, LASSO model, GLMM and GAM. While the purpose of estimating the LASSO model, GLMM and GAM was variable selection, a precursor to predictive model estimation, the results shed light on the significance of the variables from an explanatory modeling standpoint. For GLMM and GAM, the significance of the variables in the models and their p-values are reported. For the LASSO model, all variables (scaled) with non-zero coefficients are statistically significant and remain in the final subset of selected variables. For the spline based GAM, the coefficients reported are the mean degrees of freedoms of regression splines which are non-linear, and, hence, only the variable

significance is interpretable.

As can be seen in Table [2.3], the kernel density estimates of the user reported adverse events are significant predictors of medical device recalls. The GLMM estimation results indicate that the MAUDE kernel density estimate has a coefficient of 0.602 with a p-value of 0.015. The 300-day lagged MAUDE kernel density estimate (coefficient: 0.544, p-value: 0.0002) is more significant than the non-lagged MAUDE kernel density estimate in predicting medical device recalls. Also, the severity weighted MAUDE kernel density estimate is significant in predicting medical device recalls. The coefficient of the severity weighted MAUDE kernel density estimate (coefficient: 1.021, p-value: 0.0001) is more significant than just the MAUDE kernel density estimate (coefficient: 0.602, p-value: 0.015). Comparable results can be observed for LASSO in terms of coefficient estimates and GAM in terms of the relative statistical significance of the variables.

As discussed in the previous section, predictive modeling was initiated by estimating GLMM, followed by GAM, then moving on to machine learning methods, and finally to ensemble class methods. The machine learning methods we used are: Artificial Neural Network (ANN), Support Vector Machine (SVM) and Naïve Bayes models. The ensemble class models we used are decision tree based Random Forest and boosting algorithms. We used the area under the receiver operating characteristics curve as a measure for predictive accuracy of models on the hold-out test sample of data. Table [2.4] reports the area under the curve (AUC) values of the predictive models we estimated. In Table [2.4], the predictive accuracy of the initial equation estimated corresponds to the Step 1 of the predictive model estimation steps discussed in the previous section. Step 1 includes only the MAUDE variables and the control variables (see Table [2.1]). In Table [2.4], the predictive accuracy of the final equation corresponds to Step 4 of the predictive model estimation steps. Step 4 includes all the time varying covariates in addition to the MAUDE variables and the control variables. Figure [2.7] shows the ROC graphs corresponding to the final (Step 4) equation of each of the models estimated and whose AUC (i.e., predictive accuracy) is reported in Table [2.4].

Out of the predictive models estimated in this study, GLMM is the simplest and most easily interpretable. We see that even with GLMM we are able to achieve a mean AUC of 0.61 with a 95% bootstrapped confidence interval of (0.593-0.624) with the initial (Step 1) equation with the MAUDE variables and the control variables. As the predictive model estimation moved from linear to non-parametric, non-linear models such as the Naïve Bayes and the Support Vector Machine (SVM), the mean predictive accuracy improved in comparison with GLMM. However, the real improvement in the predictive accuracy occurred with the decision tree based ensemble model,

namely, Random Forest. The explanation for the improvement is the non-linear nature of the data and the likely interdependences between variables, specifically the interactions of MAUDE variables and time-varying covariates related to design, supply chain and manufacturing. Random Forest automatically accounts for variable interdependences during model fitting (training of the model). With Random Forest, the predictive accuracy with the hold-out test sample, measured as the mean AUC, is 0.82 with a 95% confidence interval of (0.815-824). As was just mentioned, this is a significant improvement in predictive accuracy compared the predictive accuracies of the linear and non-linear models that were estimated.

Using AUC as a measure of predictive accuracy, a random assignment prediction model – equivalent to a null model – would achieve a mean AUC of approximately 0.5. As can be seen in Table [2.4], the initial predictive accuracy of the estimated GLMM, a linear model, with 95% confidence interval is greater than predictive accuracy of the random null model. The initial predictive accuracy of the Random Forest ensemble model of 0.815 has a 95% lower confidence limit which is significantly greater than the predictive accuracy of the null model mean AUC. In sum, the user reported adverse events related to medical devices are a significant predictor of device recalls, lending support to H1.

Table 2.4 Predictive Accuracies of the Models Estimated in the Study

Method	Predictive Accuracy of the Initial (Step 1) Equation	Predictive Accuracy of the Final (Step 4) Equation
Least Absolute Shrinkage and Selection Operator (LASSO)	-	0.79
Generalized Linear Mixed Model (GLMM)	0.61 (0.593-0.624)	0.81 (0.765-0.823)
Generalized Additive Model (GAM)	0.64	0.86
Naïve Bayes	0.66	0.75
Artificial Neural Network (ANN)	0.65	0.85
Support Vector Machine (Radial Basis Function Kernel)	0.59	0.74
Boosting	0.59	0.78
Random Forest	0.82 (0.815-0.824)	0.95 (0.925-0.961)

Number of devices: 5,873

Train set size: 98,650

Test set size: 28,641

Total sample size (n) : 127,291

Notes:

1. *Model sensitivity is measured by numerical approximation of area under the receiver operator characteristics curve.*
2. *Confidence interval on mixed logit model is achieved by parametric boot-strap with repeated random sampling on devices.*
3. *Confidence interval on random forest is achieved by blocked boot-strap with repeated random sampling.*

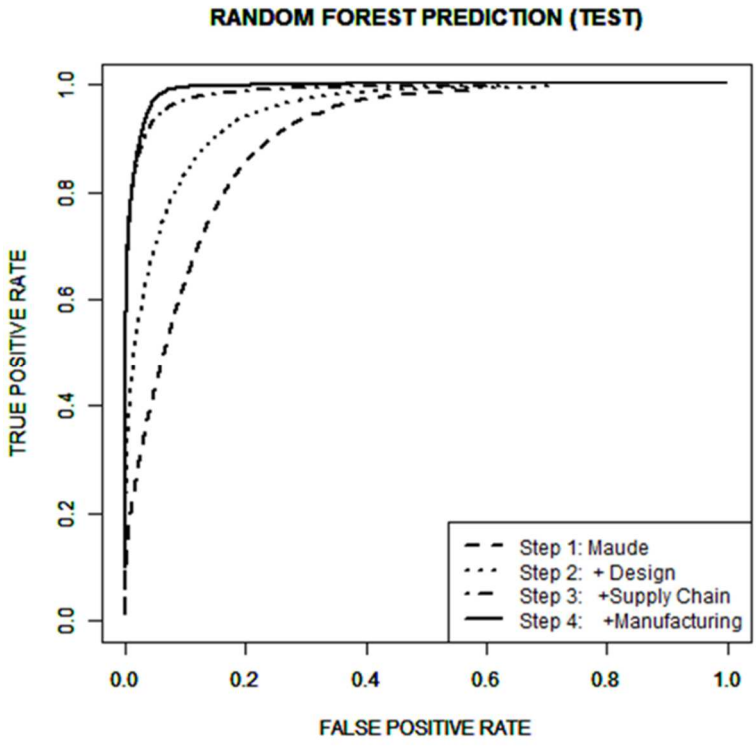
2.5.2 Testing Hypothesis 2 (H2) on the Precision of Prediction

Given that H2 posits that user reports on adverse events supplemented by time-varying covariates related to design, supply chain and manufacturing will improve the precision of prediction of failure of high tech innovations-in-use, we focus on the predictive accuracies of the final (Step 4) equations reported in Table [2.4]. As mentioned earlier, the final equation corresponds to Step 4 of the predictive model estimation steps. The final (Step 4) equation includes the time-varying covariates

related to design, supply chain and manufacturing in addition to the MAUDE variables and the control variables.

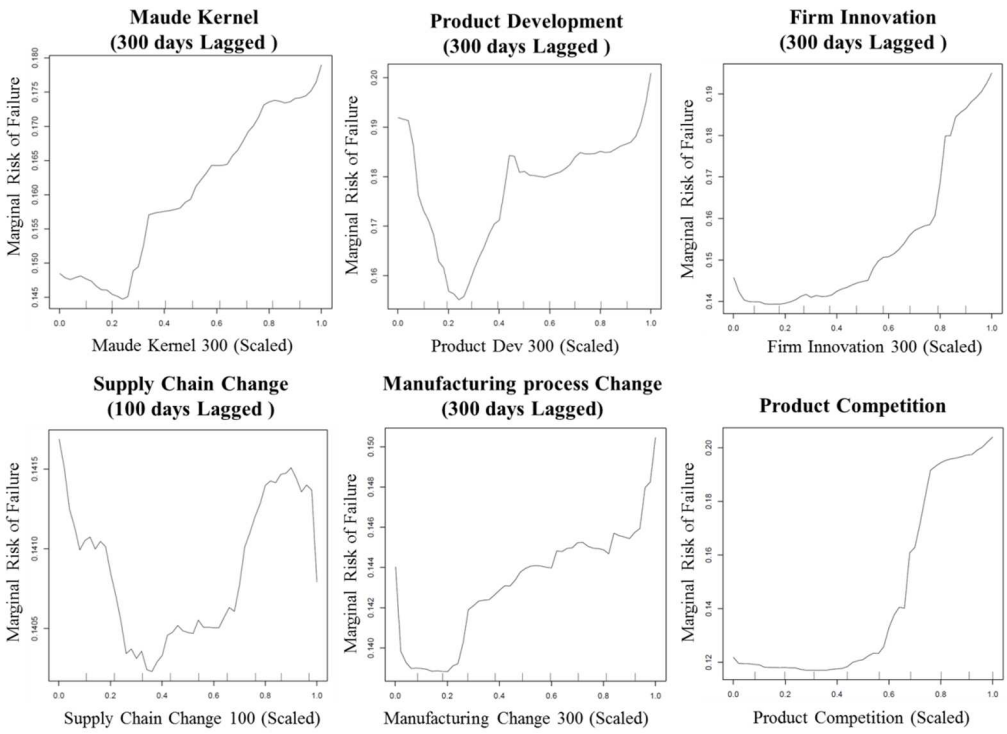
As the results in Table [2.4] indicate, the predictive accuracy of the final (Step 4) equation is greater than the predictive accuracy of the initial (Step 1) equation in all instances. Figure [2.7] depicts the improvements in the AUC or predictive accuracy from the initial (Step 1) equation to the final (Step 4) equation for the random forest model which gives the best predictive accuracy on the hold-out test sample. The ROC curve in Figure [2.7] also shows the intermediate steps after addition of the design and supply chain related covariates to the initial step where only the adverse event related primary predictors were present along with the categorical controls. Specifically, for GLMM, the mean AUC of the final (Step 4) equation has improved to 0.81 from 0.61 in the initial (Step 1) equation. We observe that the 95% bootstrapped confidence interval of the AUC of the final (Step 4) equation is (0.765-0.823) which is a significant improvement over the 95% bootstrapped confidence interval of AUC of the initial (Step 1) equation of (0.593-0.624). Statistically, the confidence intervals are significantly separated out with the initial (Step 1) equation's AUC upper confidence limit (0.624) being significantly lower than the final (Step 4) equation AUC's lower confidence limit (0.765). Similarly, for non-linear and non-parametric models such as the GAM, Naïve Bayes model, the Neural Network and the Support Vector Machine (SVM), the mean AUCs have improved considerably for the initial (Step 1) equation to the final (Step 4) equation. For the Random Forest ensemble model, the mean AUC of the final (Step 4) equation is 0.95 with a 95% bootstrapped confidence interval of (0.925-0.961). Taken together, these results indicate that when user feedback on adverse events contained in FDA's MAUDE database are supplemented by the time-varying covariates related to design, supply chain and manufacturing the accuracy of models for predicting medical device recalls is improved. In sum, these results lend support to H2.

Figure 2.7 Receiver Operating Characteristics (ROC) Curves for the Estimated Predictive Models



Since H2 is supported, we conducted additional analysis to delve into the potential contributions of individual time varying covariates on improving the precision of prediction of medical device recalls. It is conceivable that there would be practical implications of the insights gained from such analysis. For illustrative purposes, we used the Random Forest ensemble model that has the highest predictive accuracy (Step 4 equation) to generate marginal response curves (see Figure [2.8] below) for a select set of time-varying covariates and continuous control variables. These response curves provide us with a nuanced understanding of the nature of influence of specific variables in predicting medical device recalls.

Figure 2.8 Marginal Response Curves from the Random Forest Model



2.5.3 Testing Hypothesis 3 (H3) on the Consistency of Prediction

Fundamental to testing H3 is the estimation of prediction bias. Hence, we estimated the prediction bias model using the Random Forest method with all the selected variables. Note, we chose Random Forest since it has the highest predictive accuracy as can be seen in Table [2.4]. We split the test data into subsets based on usage class and device class. There are 20 usage classes including those that are not classified as miscellaneous devices. There are three device classes based on complexity of technology. Hence, a total of 60 subclasses are feasible. However, out of these subgroups some classes had too few data. Leaving out the subgroups that have very few data points, we had 48 data subgroups in all. The prediction model bias was calculated using the bias measure in equation [2.3] – i.e., the mean difference between the two classification errors, namely, the false alarm rate and the incorrect rejection rate. However, the bias value can vary based on the decision threshold. The bias was calculated at the minimum bias point for the prediction model based on the test-sample prediction accuracy. The model was run on each subset multiple times to generate the

entire ROC profile and the bias at each point (1000 in all) was calculated. The minimum bias was stored for each of the sub-groups. It is worthwhile noting that for any device-firm unit when the actual failure rate in data was below the expected failure rate from data, we classified it as under-reaction and vice versa. Also, the input data variance from the kernel density estimation using the most efficient band-width was stored for each of the 48 usable subgroups of data. For each of the subgroups, the mean severity was calculated for all the events of the subgroups using an expectation of the input MAUDE data for the period under consideration for each data point. The following bias model [2.15] was estimated using Huber's M estimation method:

$$Bias = \beta_0 + \beta_1 * Variance + \beta_2 * Mean.Severity + \beta_{12}(Variance * Mean.Severity) \quad \ldots [2.15]$$

Note, Huber's M estimation (Huber 1964) is a robust regression estimation method that is relatively less sensitive to violations of the ordinary least square estimation assumptions of conditional independence, normality, spherical errors and exogeneity. Table [2.5] reports the results from the M estimation.

Table 2.5 Huber's M-Estimation of Judgment Bias

	(1)	(2)	(3)
Severity	1.621 (0.075)+	1.477 (0.11)	-1.907 (0.339)
Variance		0.017 (0.535)	-1.278 (0.05)*
Severity*Variance			2.557 (0.047)*

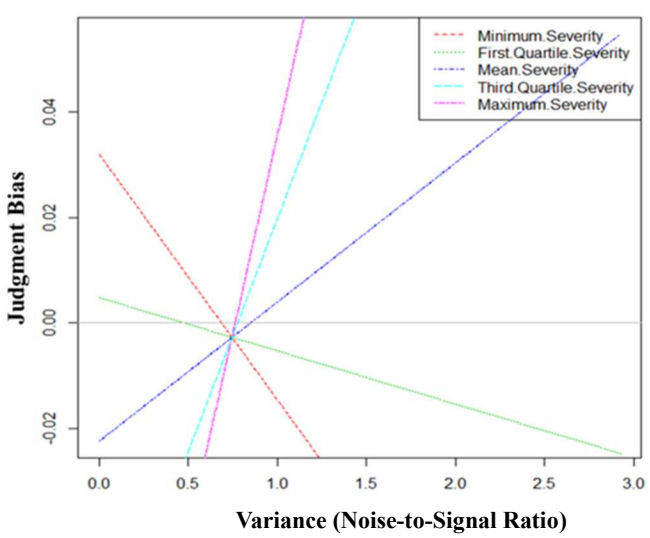
Note:

1. *Estimation was done using Huber's smoothed weight function*
2. *Robust regression is appropriate due to non-normality of error distributions from predictive models*
3. *Significance codes: <0.001 '***', <0.01 '**', <0.05 '*', <0.1 '+'*

The results in Table [2.5] indicate that while the main effects of variance (noise-to-signal ratio) and severity are not significant or marginally significant, the interaction effect of variance and severity is significant. This result can be interpreted by viewing the marginal response curves for the bias model shown in Figure [2.9]. The figure indicates that as severity increases, over-reaction bias increases with increased variance. However, as severity decreases the under-reaction

bias increases with increased variance. This result is consistent with the theoretical prediction of the bias model from the system neglect literature shown in Figure [2.5]. We conclude that there is significant judgment bias in reacting to user reported market signals of failures of high tech innovations-in-use and the bias is dependent on the perceived cost of failure and the variance of the noise-plus-signal data.

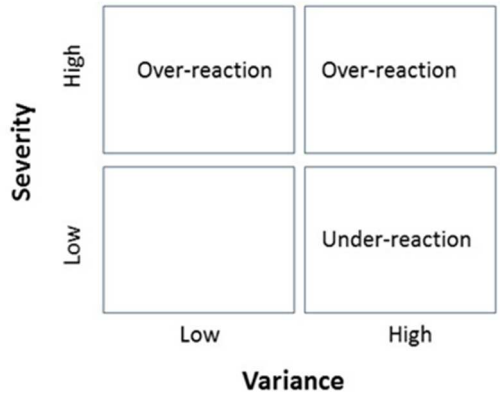
Figure 2.9 Marginal Response Curves from the Huber's M Estimate of the Bias Model



We now interpret these results in light of the relationships (direct effect) posited in Hypotheses 3A and 3B. Specifically, H3A posited that high noise-to-signal ratio (variance) is associated with under-reaction bias and vice versa; and H3B posited that high severity is associated with over-reaction bias and vice versa. The results of our analysis, however, provide novel and nuanced insights. As can be observed in Table [2.5] (column 3), there is support for the direct effect posited in H3A but not for H3B. We also uncover a significant interaction effect that provides insights into the dynamics of interpreting market signals in the form of user reported adverse events. From the standpoint of system neglect literature (cf. Massey and Wu, 2005; Kremer, Moritz and Siemsen, 2011), decision makers will under-react when there is high noise-to-signal ratio (variance). Our results support this point of view. However, the new insight that our results provide is that under-reaction changes to over-reaction in a high noise-to-signal ratio (variance) when the severity of adverse events is high. This change from under-reaction to over-reaction is indicative of the risk averseness of decision makers – i.e., when the severity of adverse events is high, decision makers

overweight the risk of potential failure of a high tech innovation-in-use (namely, a medical device recall). This result is pictorially depicted in Figure [2.10] that represents a modified system neglect framework pertaining to signal detection failures of failures of high tech innovation-in-use.

Figure 2.10 Modified System Neglect Framework based on Model Analysis



2.6 Robustness Checks and Extensions of the Predictive Model

We perform two robustness checks through two immediate extensions of the failure prediction models. First we build a prospective model by using past data to predict the future and check for the prediction accuracy. Second, we extend the prediction model to predict recall class using a two stage model.

2.6.1 Prospective Prediction Model

In this extension, we use past data on each product code to predict the future and then check the prediction accuracy on future recalls. For the purpose of this model building we use only 100 days and 300 days lagged variables from the list of final variables as described in table [2.2] for the final (step 4) equation. This is because the intention is to use past data to predict the future. Also, in the original prediction model, the lagged variables are a better predictor of recalls than the current variables. Hence, this step is useful in establishing robustness of the prediction models. We split data related to each product into a train set data and a test set data based on the temporal dimension of the sample related to the products. The total timeline of each product code has been split into an 80% prior sample (train sample) and 20% posterior sample (test sample). Products which do not

have recall events in the posterior sample have been removed from this model estimation for model testing purpose. 560 such products remained in the data. We only estimated the final (step 4) model with the averse event related primary predictor and design, supply chain and manufacturing related covariates. We used a random forest method with 10 fold cross validation to build the predictive model. Figure [2.11] shows the receiver operating characteristics of the model. The figure shows median ROC curve with an AUC value of 0.88 and the 95% bootstrapped confidence interval of the ROC curve with an AUC interval of 0.82-0.90. Table [2.6] shows the class prediction accuracy for one sample run of the random forest model (at the minimum bias decision threshold). This shows that market data related to adverse events combined with firms' information related to design, supply chain and manufacturing can be credibly used by firms to build prospective prediction models related to product failures with fairly high degree of confidence and accuracy. This can help firms behave proactively and strategically in managing failures of products and innovations already in use in the market.

Figure 2.11 Receiver Operating Characteristics (ROC) of Prospective Prediction Model

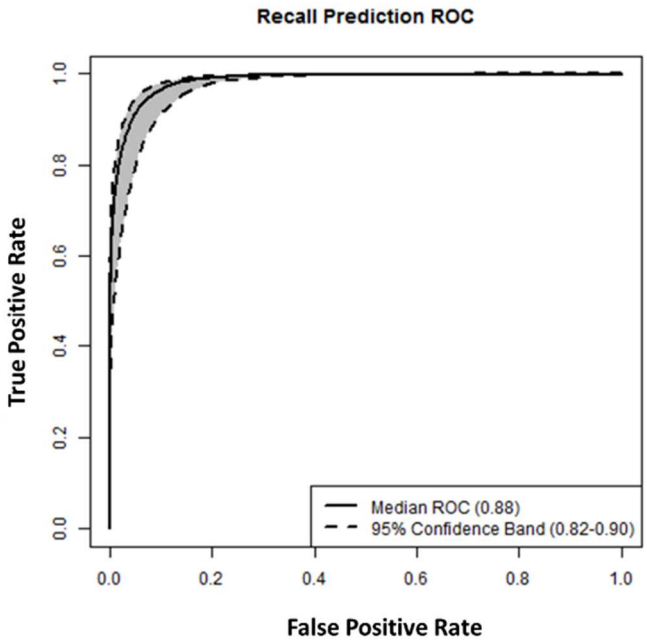


Table 2.6 Prediction Accuracy of the Prospective Prediction Model with Bootstrap (1000 runs) Standard Estimation Estimates

Actual \ Predicted	0	1
0	*0.91 (sd=0.04)****	*0.09*
1	*0.16*	*0.84 (sd=0.07)****

Overall Accuracy: 0.87 (sd=0.065)***

2.6.2 Prediction of Recall Class

Most of the recalls that happen in the medical device industry are class II recalls or the more severe class I recalls. Class III recalls are usually relatively few and are done for trivial issues like mis-labelling or packaging and distribution issues on relatively less critical medical device. Our study sample contains only data pertaining to class I and class II recalls. Having established that it is possible to use market feedback data related to adverse events along with relevant firm and supply chain related covariates to predict failures of products while in use, we tried to extend the model to predict recall class. We have used a two stage model to build the recall class prediction model as shown in system of equations [2.16] where the first step is the same as the step 4 of the original recall prediction model.

$$P(Recall) \sim F\begin{pmatrix} Maude.Kernel, Control\ Variables. \\ Design\ Covariates, \\ Supply\ Chain\ Covariates. Manufacturing\ Covariates \end{pmatrix}$$

$$P(Recall\ Class = 1 \mid True\ Positive) \sim F(Maude.Kernel, Relevant\ Covariates)$$

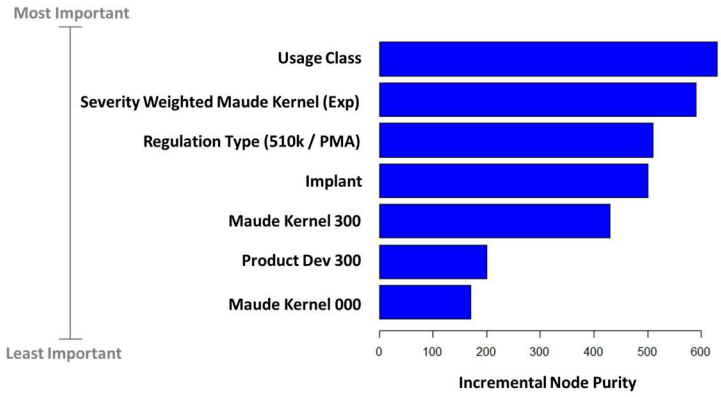
$$P(Recall\ Class = 2 \mid True\ Positive) \sim F(Maude.Kernel, Relevant\ Covariates)$$

… [2.16]

In the context of recall class prediction, selection of the right covariates is critical like the original prediction model. The importance order of the variables for recall prediction is not likely to the same for recall class prediction. We selected the right variables for recall class prediction as a subset of the variables that have been selected for the recall prediction model as described in

Table [2.2]. To do the selection, we first estimated a random forest model with all the variables in Table [2.2] with 10 fold cross validation and used the cross validation error to order the variables in terms of decreasing importance in predicting recall class. The variable ranking in random forest is done using a node purity measure which measures the predictive power of each variable at each node split of the ensemble of decision trees in a random forest ensemble model. The node purity is measured using a voting count of the number of splits and the related decrease in the cumulative cross validation prediction error. For this variable selection step, we built a random forest model with 1000 trees. In random forest model using a higher number of tree ensemble does not lead to over-fitting, rather using a higher number of trees lead to asymptotic convergence of the estimated model to the true data generating model. However, having extra variables which are not important in predicting recall class can lead to over-fitting and inconsistent prediction. Hence, important subset of variable selection is an important step both in terms of model parsimony as well as in terms of consistency of prediction. The important variables for recall class prediction is shown in Figure [2.12].

Figure 2.12 Variable Important Ranking for Recall Class Prediction



The four most important variables for recall class prediction are usage class, severity weighted adverse event density, regulation type (PMA/510k) and implant code. A closer look at the data and several illustrative decision trees (not shown here for sake of simplicity) from the random forest model reveals that class I recalls are dominant in a few usage classes like cardio-vascular, surgical, and orthopedic. The interaction of severity of adverse events lead to significant number of class I recalls. Regulation type is a significant predictor of recall types. Pre-market approval approvals for new products lead to higher class I recalls as compared to 510k approvals done for

new versions of existing products. This shows that new products has a higher likelihood of having class I recall as compared to new versions for existing recalls. Also, what is interesting is that the un-weighted density of adverse events is no more important relative to usage class, severity of adverse events, implant code and regulation type in predicting the severity of the recall. Including more variables reduces the node purity in the random forest model and does not add to prediction accuracy. Having identified the important variables for recall class prediction we estimated a recall class prediction model using 80% train set on the subset of the variables. We tested the model accuracy using the 20% hold-out test sample. The results of the test set prediction accuracy is shown in Table [2.7]. Table [2.7] shows that the overall prediction accuracy of recall class prediction is 0.938 with a standard deviation of 0.05 ($p<2.0e-16$). This shows that it is also possible to extend the recall prediction model to predict recall class with high degree of confidence. However, recall class specification is more dependent on the type and age of the product that is in use and the intended usage of the product. Market feedback on the severity of adverse events do have a significant role in determining the recall class, but the frequency of adverse events do not play a very significant role unlike in the case of recall prediction.

Table 2.7 Prediction Accuracy of the Recall Class Prediction Model with Bootstrap (1000 runs) Standard Deviation Estimates

		Predicted	
		1	**2**
Actual	**1**	*0.963 (sd=0.02)****	*0.037*
	2	*0.116*	*0.884 (sd=0.043)****

Overall Accuracy: 0.938 (sd=0.05)***

2.7 Conclusion

2.7.1 An Overview

The motivation for this study stemmed from the growing recognition in the corporate sector, government, academia, and society-at-large that in spite of the best efforts of firms and

governmental regulatory agencies, high tech innovations are continuing to fail while in use in the marketplace. The failures of high tech innovations-in-use are evidenced as product recalls. So, while there has been growing excitement and enthusiasm about the upside potential – namely, life-saving and life-enhancing potential – of high tech innovations, there have been also growing concerns about the downside risk, especially since the failure of a high tech innovation-in-use can cause injury or death. This study was undertaken, in part, as a response to the call to action that efforts be directed at predicting failures of high tech innovations-in-use through early detection of signals of failures. We conducted a theoretically-grounded, empirical inquiry to address the questions: (i) Do adverse events reported by users of high tech innovations-in-use predict their failures? If so, how can the precision of such prediction be improved? (ii) Do firms exhibit a judgment bias in predicting failures of high tech innovations-in-use? If so, do firms over-react or under-react?

We used the U.S. medical device industry as the empirical setting for this study. This industry was an appropriate setting since it is an industry well-known and highly regarded for introducing high tech innovations in the marketplace that have the potential to save lives or enhance the quality of lives. Also, this industry is being increasingly challenged with rapidly growing number of failures of high tech innovations-in-use evidenced as device recalls. Using big and unstructured data on user reported adverse events related to medical devices (FDA's MAUDE data) along with data from other complementary databases, we applied predictive analytics to build a modeling framework for signal detection to predict medical device recalls, and evaluate the existence of judgment bias in making such predictions.

2.7.2 Contributions

By way of contributing towards advancing the technology and innovation management literature, this chapter has demonstrated that the prediction of failures of high tech innovations-in-use is possible with sufficient lead time by analyzing big and unstructured data on user feedback on adverse events in the marketplace; and that the precision of such predictions can be significantly improved by accounting for time-varying covariates related to design, supply chain and manufacturing. We also show that we can predict the recall class considerably consistently using product related and adverse event severity related data.

Next, the study has provided novel and nuanced insights into why firms often fail to correctly detect market signals of failures of high tech innovations-in-use. In particular, we showed

that the existence of judgment bias leading firms to under-react or over-react to market signals in the form of user feedback on adverse events, and identified factors – such as (i) the severity of the adverse events, (ii) noise-to-signal ratio in the user feedback data stream on adverse events, and (iii) the interaction between (i) and (ii) – that influence firms to under-react or over-react. Consistent with the past literature, we found that firms under-react when there is high noise-to-signal ratio in the user feedback data stream on adverse events. However, the new insight from our study results is that under-reaction changes to over-reaction under conditions of a high noise-to-signal ratio when the severity of adverse events is high. This change from under-reaction to over-reaction is indicative of the risk averseness of decision makers – i.e., when the severity of adverse events is high, decision makers overweight the risk of potential failure of a high tech innovation-in-use (namely, a medical device recall).

Finally, this study being the very first – to the best of our knowledge – to have applied predictive analytics to develop models with a big and unstructured database to predict failures of high tech innovations-in-use and evaluate judgment bias, opens up new empirical research possibilities of conducting technology and innovation management research. The new possibilities include: (i) going beyond an orientation of innovation success to innovation failure, especially when the innovations are in use in the marketplace; (ii) going beyond an orientation of explanation to prediction while accounting for judgment bias; and (iii) going beyond the analysis of structured data-sets containing hundreds or thousands of observations to analysis of unstructured data-sets containing millions of observations or more.

2.7.3 Practice and Policy Implications

Given that the failures of high tech innovations-in-use in the marketplace have significant economic and social costs, this study's contributions have implications for practice and policy. Since the medical device recalls serve as the empirical context of our investigation, the United States Government Accountability Office's (GAO) report to the United States Senate (GAO-11-468, June 2011) helps to put the contributions of this study into perspective. This GAO report calls for the use of market usage data on medical devices to improve the oversight and surveillance of such devices, and the device recall process. Towards that end, this study shows how the big and unstructured MAUDE database containing user reports on adverse events related to medical devices can be mined and analyzed for timely detection of market signals and prediction of medical device recalls. Further, the study identifies design, supply chain and manufacturing covariates, and

the databases from which the data on the covariates can be collected and analyzed along with the MAUDE data to improve the precision of prediction of medical device recalls.

As can be seen from the marginal response curves presented in section 5 (Figure [2.8]), this study makes it possible to estimate the marginal impact of individual variables in the predictive analytic models on the likelihood of device recalls. Firms can act upon individual risk factors depending on their marginal effects in ways that can reduce the risk of recalls, thereby improve the quality and reliability of medical devices after-market launch. For example, insights from the study results can inform the sequencing of activities for a medical device manufacturing firm – namely, stabilizing the design of a medical device before undertaking changes in supply chains or manufacturing processes. Such sequencing of activities is consequential since it is not uncommon for firms pressed by the need to speed up market launch of medical devices as well as to improve the profitability of devices to undertake multiple competing activities simultaneously.

Another key contribution of this study that has implications for practice and policy is the finding that significant judgment bias exists in decision making pertaining to medical device recalls. While some of the judgment bias may appear to be a reflection of a firm's resource constraints, the empirical support we found in this study for the existence of judgment bias provides medical device firms a basis to acknowledge that such bias exists – a necessary condition for engaging in efforts to reduce bias and improving decision making related to medical device recalls.

Given the resource constraints of a firm or regulatory agency, setting priorities and allocating resources for the surveillance of medical devices while in use in the marketplace are both consequential and challenging. The predictive analytic models developed in this study illustrates how such decision making can be informed based on the analysis of relevant and available data, thereby addressing to the calls of regulatory agencies – namely, FDA and GAO – to make the market surveillance process of medical devices data-driven and objective.

2.7.4 Future Research Directions

The execution of the research design for this study required a methodical approach: (i) programming data extraction, (ii) data organization, (iii) variable generation using modified latent Dirichlet models for text mining, and (iv) estimation of predictive analytic models – i.e., statistical (linear) models (LASSO, Generalized Linear Mixed Logistic Model, Generalized Additive Model), data mining models (Artificial Neural Network Classifier, Naïve Bayes Estimator), and machine learning ensemble models (Boosting Classifier and Random Forest Classifier) using the R and

Python computing platforms. This methodical approach to model estimation is foundational to pursuing a future direction of research specifically aimed at developing a decision support system to automate the detection of signals and prediction of medical device recalls. Another more general direction of future research would be to develop decision support systems to automate the detection of signals and prediction of failures of high tech innovations-in-use, namely, "Internet of Things" – i.e., sensors and actuators in equipment and devices linked through wired and wireless networks that churn out, on a real time, large volumes of data on the functioning of the equipment and devices to computers for analysis.

Given that in this study we found firms exhibit judgment bias, by way of under-reaction or over-reaction, in detecting market signals and predicting medical device recalls (failures of high tech innovations-in-use), a future direction of research would be to conduct field experiments in a medical device firm (a high tech firm) to investigate (i) conditions that cause under-reaction versus over-reaction and (ii) interventions that can reduce judgment bias.

In closing, this study is among the very first to demonstrate the application of predictive analytics to analyze big and unstructured data to yield novel insights for managing high tech innovations-in-use in the marketplace. With new types of data-sets becoming available today that are rich in detail and combine information of many types (namely, temporal, cross-sectional, geographical, and textual) on a large number of observations and with high level of granularity, predictive analytics is a methodological approach that has the potential to detect new patterns and behaviors and help uncover new causal mechanisms. We believe that this study will encourage further research using predictive analytics on the new types of data-sets yielding new insights, and will thereby advance the theory and practice of technology and innovation management.

Chapter 3:

Evaluating Sources of Judgment Bias in Detecting Failures of High Tech Innovations-in-Use: Analysis of Big Data on User Reported Adverse Events Related to Medical Devices

3.1 Introduction

Increasingly, high tech innovations are becoming key to the competitive success of firms across industry sectors. Notwithstanding the upside potential, high tech innovations entail risks of failures and adverse effects to users while in use in the market. Failures of high tech innovations-in-use have several downside risks for the innovating firms such as regulatory sanctions, loss of profit and revenue, and loss of market equity. Failures of high tech innovations-in-use entail risks of harm and injury to users, and economic loss to the society in general. Hence, firms and regulatory agencies collect market feedback from users in several forms and through multiple channels. These user generated market feedback on high tech innovations-in-use can act as a source of signal detection to identify potential failures of products either due to design issues or due to manufacturing and supply chain glitches. The Government Accountability Office (GAO) in a report on medical devices (GAO-12-816, 2012) has identified the need to use more precise measures like Unique Device Identification (UDI) to analyze market generated signals to identify potential failures of devices while in use in the market. Here is an excerpt from the aforementioned report:

"One initiative is the Unique Device Identification effort for the post-market surveillance of devices, which, according to FDA, will allow the agency to aggregate adverse event reports in order to more accurately analyze them when conducting signal analyses. The initiative will also allow FDA to identify specific devices included in adverse event reports, allowing for more rapid and effective corrective actions that can focus on specific devices" (GAO-12-816, 2012, p. 33).

In spite of the best efforts of firms and the best intentions of regulatory agencies, firms often fail to react to signals of potential innovation failures from user feedback. Either firms are too slow to react to user feedback on adverse events related to high tech innovations-in-use or sometimes they react too soon causing avoidable inconvenience to the firm and the user alike. This seems to indicate the presence of judgment bias in firms when it comes to reacting to market signals related to failures of high tech innovations-in-use. Many of the incidences of severe under-reaction to market signals have captured media and public attention. A few illustrative examples from automotive, aerospace and medical device industry segments of such incidences that have attracted media attention are stated here. In a media report[3] by New York Times (www.nytimes.com, Sept 14, 2014), it was highlighted that regulatory agencies were slow to react to reports of market failures of several automotive products. An excerpt from the report states:

> *"The Times analyzed agency correspondence, regulatory documents and public databases and interviewed congressional and executive branch investigators, former agency employees and auto safety experts. It found that in many of the major vehicle safety issues of recent years — including unintended acceleration in Toyotas, fires in Jeep fuel tanks and air bag ruptures in Hondas, as well as the G.M. ignition defect — the agency did not take a leading role until well after the problems had reached a crisis level, safety advocates had sounded alarms and motorists were injured or died … The analysis by The Times found that before the recalls the agency had received more than 5,000 complaints about the ignition problems, including more than 2,000 about unexpected stalling, in the models G.M. eventually recalled for an ignition defect that could lead to stalling…*". (www.nytimes.com, Sept. 14, 2014).

In a related subsequent media report[4], it was stated that:

> *"… (after the media attention) automakers are cleaning up years of defects that previously went undetected or ignored …*" (www.nytimes.com, Dec. 30, 2014).

---

[3] http://www.nytimes.com/2014/09/15/business/regulator-slow-to-respond-to-deadly-vehicle-defects.html

[4] http://www.nytimes.com/2014/12/31/business/a-year-of-record-recalls-galvanizes-auto-industry-into-action.html

In another report[5] on the medical device industry it was pointed out that a manufacturer of hip and knee joints had received reports of failures of their products long before they actually reacted to such failures. These are just a few of similar incidents reported in the media that seems to indicate judgement bias by way of under-reaction in the decision making of firms. While under-reaction of firms to market signals spur media attention, over-reaction to user-generated signals do not seem to spur similar media attention. However, our interactions with managers of several medical device manufacturers in the United States seems to indicate the prevalence of over-reaction bias to market signals in the form of user feedback on high tech innovations-in-use. This conjecture was also supported by our cursory review of a few user reports on adverse events in the context of the medical device industry, the empirical setting our study. Hence, the questions that serve as the motivation for this study are:

I. *Do firms systematically exhibit a judgment bias in interpreting market signals related to failures of high tech innovations-in-use? What factors influence the judgment bias, if present? and*
II. *If the firms do exhibit judgment bias in interpreting market signals related to failures of high tech innovations-in-use, what factors influence under-reaction bias and what factors influence over-reaction bias?*

The empirical context of the study is the medical device industry. The data-set that is central to conducting this study is a user-reported adverse event database called the Manufacturer and User-Facility Device Experience (MAUDE) collected by the Food and Drug Administration (FDA) of the United States. Each year FDA receives several hundred thousands of adverse event reports pertaining to either new medical devices (referred to as pre-market approvals [PMA]) or new versions of existing devices (referred to as 510K approvals). In this study, we mine this big data-set to generate insights related to the presence of judgment bias in firms related to detection of failures of high tech innovations-in-use. While the MAUDE database is unstructured, it is fine-grained in terms of capturing user experience on an almost real time and ongoing basis. In an article on use of big data in management, Haas and Pentland (2014, p. 321) have observed that "… the defining parameter of big data is the fine-grained nature of data itself, thereby shifting focus away from the number of participants to the granular information about the individual." In the same

[5] http://www.nytimes.com/2013/01/23/business/jj-study-suggested-hip-device-could-fail-in-thousands-more.html

article, the authors, Haas and Pentland, have observed that this granular nature of big data, e.g., user reported adverse events, can be a means for analyzing decision making.

Similarly, Lynch (2008, p. 28) observed that big data is "big" by way of "being of lasting significance" in terms of the context and the insights that the data can potentially provide. The context that we analyze is of lasting significance for several reasons. The rapidly increasing cost of health care in the United States and globally is well recognized. A substantial part of the health care cost is due to the cost of medical devices and equipment. This cost is increasing rapidly since health care delivery is becoming increasingly enabled by high tech innovations. Medical device recalls (i.e., failures of high tech innovations-in-use) contribute significantly towards this increasing health care delivery cost. Decision bias pertaining to management of high tech innovations-in-use in health care delivery is a serious concern for the innovation user community, the manufacturers and the regulators. Apart from direct cost of health care delivery, failures of medical devices has the potential to cause severe harm to individual patients that can range from injuries, hospitalizations to deaths. Hence, it is important to understand the sources of judgement bias in detecting medical device recalls (i.e., failures of high tech innovations-in-use). Acknowledgement of the sources of judgment bias, if any, will improve the post-market surveillance and make detection of failures of high tech innovations-in-use in health care more evidence based, timely, and predictive. Timely detection of failure risk will reduce cost and improve effectiveness of health care delivery. With health care delivery is becoming increasingly enabled by high tech innovations, accounting for judgment bias in detecting failures of high tech innovations-in-use will contribute towards addressing the grand challenge of ensuring safe and cost-effective health care delivery.

In a comprehensive McKinsey Global Institute report, Manyika et al., (2011) have identified the transformative potential of big data in the health care sector and have noted that the next wave of efficiency gain and cost containment in health care sector will depend on how manufacturers, healthcare delivery organizations, regulators and the Government is able to harness the potential of big data in health care that are becoming increasingly available from several sources including the patients. The same McKinsey report recognizes that the next level of competitiveness for manufacturers and health care delivery organizations would come from how effectively and efficiently those firms are able to utilize information from the users to improve their operations and how well those firms are able to make managerial decisions evidence-based and timely. In this study, we analyze big data in the form of user reports on adverse events related to medical devices (i.e., high tech innovations-in-use) to investigate both the existence and sources of judgment bias

in detecting and responding to signals of potential medical device recalls (i.e., failures of high tech innovations-in-use).

3.2 Theory and Hypotheses Development

We draw on three theoretical perspectives to develop the study hypotheses. First, we use the theoretical perspectives of signal detection (Salkind, 2007) and system neglect (Massey and Wu, 2005) to introduce the concept of judgment bias and the two types of judgment bias, namely, under-reaction and over-reaction. Second, we use the theoretical perspective of attention based view of firm (Ocasio, 1997) to understand conditions that can lead to under-reaction and over-reaction bias to user feedback.

3.2.1 Signal Detection Theory

The fundamental idea of signal detection is to recognize from the input data stream (e.g., user reported adverse events data stream related to a high tech innovation-in-use) the presence of a specific state of the system (e.g., the failure of a high tech innovation-in-use). A detection method processes the signal and produces a judgment of whether a system level issue is present or not (Harvey, 1992; Wagner et. al., 2001). When the operating condition of a system is perfectly normal, the distribution of the data stream generated will be white noise. However, if a system level issue is present, then the distribution of the data stream will shift and will no longer be a white noise distribution. From the standpoint of signal detection theory (SDT), this distribution is the combined noise and signal distribution. The core idea of SDT is to be able to detect the presence of such a signal in the data stream. The detectability of the signal depends on the strength of the signal as well as the detection process operationalized by the choice of an appropriate decision threshold, as is depicted in Figure [2.1] in Chapter 2.

Let C_s be the sunk cost of a false alarm and C_o be the opportunity cost of missing out on a credible failure signal, a rational decision maker would like to minimize the expected costs associated with the two types of errors in judgment, i.e., *false alarm* and *miss*. The decision criterion of the rational decision maker can be stated as:

$$\min_{X_c}\{C_o * P(Miss) + C_s * P(False\ Alarm)\} \qquad \ldots [3.1]$$

If the noise distribution is $N(\mu_o, \sigma_o^2)$ and the noise plus signal distribution is given by $N(\mu_s, \sigma_s^2)$, then the optimal decision threshold X_c is given by equation [3.2] Normalizing the mean of the

noise distribution to zero (WLOG), we further simplify the optimal value for the decision criterion as follows; see Appendix for proof.

$$X_c = \frac{\mu_s}{2} + \frac{\lambda\sigma^2}{\mu_s};\ \lambda = \log(\beta) = \log\left(\frac{C_s}{C_o}\right) = \log\left(\frac{Sunk\ Cost\ of\ False\ Alarm}{Opportunity\ Cost\ of\ Miss}\right) \quad \ldots [3.2]$$

3.2.2 Judgment Bias: Under-reaction and Over-reaction

In this study we investigate how decision makers react to a stream of data in the form of user feedback on adverse events related to high tech innovations-in-use. The empirical context of this study is the medical device industry. The data-set that is central to conducting this study is the big and unstructured database on user-reported adverse events on medical devices collected by the Food and Drug Administration (FDA). Adverse event reports contain information related to the nature of device, device type, and nature of adverse event from the standpoint of a patient (e.g., hospitalization or death). FDA acknowledges the potential of user reported adverse events to be used as signals for failure detection. The failure detection process depends on a decision threshold that decision makers choose to classify a data stream as either a signal for a potential device failure or as pure white noise. In making this judgment related to potential failure of devices, decision makers can make two possible errors originating from their choice of a decision threshold (equation [3.2]). First, decision makers can wrongly detect a failure of a device when in reality there is none. The rate at which decision makers make this kind of error is called the False Alarm Rate (*FAR*) (Salkind, 2007). Second, decision makers can miss out on detecting a credible failure signal and the related error rate is called the Miss Rate (*MR*) (Salkind, 2007). Both the error rates are dependent on the choice of decision threshold by decision makers. A decision threshold that is higher than the optimal decision threshold (equation [3.2]) leads to a high miss rate, whereas a decision threshold that is lower than the optimal decision threshold would lead to a high false alarm rate. Decision makers who systematically exhibit a propensity towards one type of error rate over the other is said to exhibit a systematic judgment bias in detecting failures of devices in use from market and user feedback data (Massey and Wu, 2005). Thus, *judgment bias* is measured using equation [3.3] (Macmillan and Creelman, 2005; Wickens, 2002).

$$Judgment\ Bias:\ \ c = \frac{False\ Alarm\ Rate\ (FAR) - Miss\ Rate\ (MR)}{2} \quad \ldots [3.3]$$

A positive measure of judgment bias means a systematic propensity of decision makers generate make false alarms over misses. Thus, decision makers exhibit an *over-reaction bias* when

$c > 0$ (equation [3.3]). Similarly, decision makers exhibit *under-reaction bias* when $c < 0$ (equation [3.3]). Having defined judgment bias relevant to the context of detection of device failure while in use from market feedback, we look at possible sources that can be associated with a higher likelihood of one type of judgment bias over the other. In the context of using user feedback related to adverse events to detect potential failures of high tech innovations-in-use, there are two primary sources that can lead to differential likelihood of the two types of judgment bias, i.e., under-reaction bias and over-reaction bias. The first source of differential likelihood of the two types of judgment bias is the characteristics of user-generated data stream that acts as the signal for potential failure of a high tech innovation-in-use that is interpreted by decision makers in making a judgment (e.g., to recall or not to recall). We analyze this first source by using the theoretical perspective of system neglect. The second source of the differential likelihood of the two types of judgment bias is the environmental condition in which the decision makers are situated, e.g., firm conditions and product-market conditions. We analyze this second source using the theoretical perspective of attention based view of the firm (Occasio, 1997).

3.2.3 System Neglect Hypothesis: Characteristics of User Reported Adverse Event Data-stream.

In the published literature, the theoretical perspective of *system neglect* has been used to understand the judgement bias of decision makers in the context of change detection (Massey and Wu, 2005) and in the context of supply chain forecasting (Kremer, Moritz and Siemsen, 2011). Both Massey and Wu (2005) and Kremer, Moritz and Siemsen (2011) show that decision makers systematically deviate from optimal change detection from data streams generated by relevant environment based on certain characteristics of the data that act as the signal for change. Both these papers show that for a given signal strength, the variance of the noise-plus-signal data is a significant predictor of the type and extent of judgment bias. In this study, we use the theoretical perspective of system neglect in the context of user feedback on adverse events in detecting failures of high tech innovations-in-use to theoretically analyze the sources of judgment bias.

The parameter X_c in equation [3.2] provides the optimal decision threshold for decision makers in the context of detection of failures of high-tech innovation-in-use. However, if decision makers adopt a decision threshold that is greater than the optimal decision threshold X_c, the decisions makers will have a high miss rate. Thus, decision makers will under-react to market signals. On the other hand, if decision makers adopt a decision threshold that is lower than the

optimal decision threshold X_c, the decision makers will generate a high rate of false alarms. Thus, the decision makers will over-react to market signals. For a given signal strength, the two factors related to system neglect that are associated with judgment bias are the noise-to-signal ratio $\frac{\sigma^2}{\mu_s}$ and the natural logarithm of the ratio of the perceived cost of errors in detection of failure signals, λ. As the noise-to-signal ratio increases, the likelihood that decision makers adopt a high decision threshold increases. Hence, high noise-to-signal ratio is likely to be associated with under-reaction bias. The severity of adverse events is indicative of the cost associated with a miss rate. Given the inherent risk averseness of decision makers, a high mean severity of the adverse events is likely to increase the perceived cost of a *miss* which is the opportunity cost, thus decreasing the cost ratio λ. Decrease in the perceived cost ratio λ is likely to be associated with a low decision threshold X_c. Hence, mean severity of adverse event is likely to be associated with over-reaction bias. Based on the above discussion, we posit the following set of hypotheses:

> **Hypothesis 1a (H1a):** Noise-to-signal ratio (variance) in user reports on adverse events related to high tech innovations-in-use is positively associated with under-reaction bias.

> **Hypothesis 1b (H1b):** Noise-to-signal ratio (variance) in user reports on adverse events related to high tech innovations-in-use is negatively associated with over-reaction bias.

> **Hypothesis 2a (H2a):** Severity of adverse events is positively associated with over-reaction bias.

> **Hypothesis2b (H2b):** Severity of adverse events is negatively associated with under-reaction bias.

3.2.4 Attention Based View of Firms: Characteristics of the Environment in which a Decision Maker is Situated.

Apart from the characteristics of the data stream that acts as the signal, it is conceivable that the specific firm and market conditions in which decision makers are situated are critical in explaining

the extent and type of judgment bias. The role of decision makers in exhibiting either under-reaction bias or over-reaction bias is evidenced by the choice of decision threshold. While the choice of decision thresholds and corresponding decisions are characteristics of decision makers, decision makers are situated in the context of specific organizations, and markets. Hence, in spite of the fact that decisions are made by decision makers, firm characteristics and specific market characteristics are likely to systematically influence the choice of decision threshold and the resulting type and extent of judgment bias. To understand the impact of the context in which the decision makers are situated on the judgment bias of decision makers, we draw on the theoretical perspective of the attention based view of firms (Ocasio, 1997). The primary tenet of the attention based view of firms is that a firm is a "*system of structurally distributed attention"* (Ocasio, 1997, p. 189) in which the decisions and actions of individuals cannot be determined by knowing the characteristics of the decision makers. Rather, the nature and pattern of decisions can be determined by the organizational contexts in which decision makers find themselves situated. This viewpoint has also been supported by Simon (1947) who contended that organizations influence decision makers by shaping the distribution of organizational stimuli that form the attention process of the decision makers. Ocasio (1997, p. 189) defines *attention* as "noticing, encoding, interpreting, and focusing of time and effort by organizational decision-makers on both (a) *issues:* the available repertoire of categories for making sense of the environment; and (b) *answers:* the available repertoire of action alternatives." Haas, Criscuolo and George (2015), while analyzing managerial attention process in an online community, have found that problem characteristics such as problem length, problem breadth and problem novelty, and problem crowding are significantly associated with differential attentional process. The attention based view of firms has been used in several organizational contexts. Hoffman and Ocasio (2001) use the organizational attention process to analyze how firms and industries react to external events. While external events are critical triggers that can lead to significant changes in an industry and motivate institutional transformation, evolution and improvement, Hoffman and Ocasio (2001) find that firms allocate differential importance and attention to different external events. The *saliency* of external events are important predictors the attention process of a firm. Higher saliency created by the extent of immediate media attention trigger higher and faster reaction towards those events as compared to events that have low saliency to a firm. Li et. al. (2013) have used the attention based view of firm to analyze innovation performance of firms in high tech industry. Joseph and Ocasio (2012) studied a large multi-business firm and found that firm structure and architecture significantly influences managerial attention

process and decision making. The attention based view of the firm is characterized by the following three tenets:

***Focus of attention**:* At any given point in time, decision makers focus their attention on a limited set of issues and external triggers. Given limited managerial resources, firms are able to attend to specific issues at a time. This leads to a sequential nature of the attention process within a firm.

***Situated Attention**:* This concept is closely related to the focus of attention. Given that firms and decision makers can focus their attention to limited sets of issues, at any given time, firms tend to attend to issues that decision makers find themselves situated in depending on a firm's context at that time. Hence, situational attention process leads decision makers to attend to issues that are more salient and immediate in nature depending on a firm's immediate contextual realities and internal and external stimuli.

***Structural Distribution of Attention**:* The resources, structural distribution of resources, and social and industry position of a firm generate a distributed focus of attention among decision-makers participating in the firm's procedural and communication channels. Hence, a larger firm with wider structural distribution of resources is likely to focus less on specific issues unless the saliency and immediacy of the impact of those issues are high.

The key tenets of the managerial attentional process is adequately supported by several scientific studies in neural psychology literature (Walter and Koch, 2006) where the attentional process is considered to be a selective gating mechanism where managers attend to stimuli that are above a saliency threshold which, in turn, is determined by the number of and relative saliency of all available stimuli. The effect of number of issues that a firm needs to deal with, on the relative saliency of a specific issue and the resulting likelihood of addressing the specific issue can be concisely explained by considering the following simple model of firm attention process. Following Walter and Koch (2006), let us consider that at any time a firm and its key decision makers are exposed to N signals related to their internal or external environment. Let the saliency of each of those N issues be represented by the random variable $S_i: i \in \{1, \dots, N\}$ distributed according to the distribution function $S \sim f(s)$ with CDF $F(S)$. Without loss of generality (WLOG), if we assume that a decision maker is able to attend to one issue at a time and that the decision maker attends to the issue with the highest situated and contextual saliency, then the probability that the decision maker reacts to a specific signal S_1 out of those N signals is given by $P[S_1 \geq S_j; j \neq$

$1\big] = [F(S_1)]^{N-1}$. Therefore, we have $\frac{\partial P[S_1]}{\partial N} = -|\log F(S_1)|[F(S_1)]^{N-1}$; see Appendix B for derivation. Since $0 < F(S1) \leq 1$, for $N > 0$, $\frac{\partial P[S_1]}{\partial N} < 0; \forall N > 0$. Hence, the probability that decision makers react to a specific environmental signal is a decreasing function in the number of signals that the decision makers is faced with simultaneously. This is supported by an empirical study of online knowledge sharing community by Haas, Criscuolo and George (2015, p. 686), where the authors have observed that "because attention is a finite resource, beyond some point a large number of concurrently posted problems may reduce the likelihood that the potential knowledge provider decides to allocate attention to a focal problem."

Based on the above discussion, we consider the following *firm, product,* and *market* characteristics and their likely impact on judgment bias related to detection of failures of innovations-in-use from user feedback of adverse events in the market:

Firm Characteristics

Firm Size*:* The size of a firm determines the extent and span of issues that key decision makers in the firm are faced with at any given time. A large firm is likely to have a much wider span of issues to deal with. The structural distribution of different types of issues that decision makers are required to pay attention to, are generally much higher in larger firms as compared to smaller firms. Joseph and Ocasio (2012) argue that large firms are able to put less specific attention on a particular issue due to a wider variety and larger number of issues that they are needed to attend to. The perceived saliency of market feedback related to adverse events pertaining to an individual high tech innovation-in-use is likely to be much less in larger firms as compared to smaller firms. Hence, we hypothesize that the likelihood of under-reaction bias would be high for a large sized firm.

Hypothesis 3a (H3a): Ceteris paribus, firm size is positively associated with under-reaction bias.

Hypothesis 3b (H3b): Ceteris paribus, firm size is negatively associated with over-reaction bias.

Firm Growth*:* Firms grow their business either through expansion of their market share or

extension of their business lines. Extensions of business lines can be achieved either organically through internal resources or inorganically through mergers and acquisitions. Either form of growth strategy pursued by a firm influences the structure of situated attention within the firm which influences decision makers to focus more towards growth related activities and less towards market signals of failures of products already in use (Greve, 2008). Below we analyze both organic and inorganic growth strategies of firms in the context of judgment bias related to interpretation of user generated adverse event reports.

***Firm's Organic Growth**:* Organic growth activity through internal resources like market expansion, product line expansion or both often require substantial investment in fixed assets and manufacturing growth (Ollinger, 1994). Firms with high focus on growth through internal resources need to focus their attention on growth activities such as new market exploration, manufacturing base expansion, supply base expansion and product expansion. The situated attentional process towards growth activities in such firms lead to lower relative saliency of market feedback related to adverse events of individual devices already in use in relation to other issues related to the growth process. The immediacy of impact of issues related to the growth process often require decision makers to focus more on those activities and decisions than existing devices already in use. We hypothesize that firms with higher organic growth rate measured by cumulative average growth of fixed assets within the firm (excluding assets gained through mergers or acquisitions) would be significantly associated with higher levels of under-reaction bias.

Hypothesis 4a (H4a): Ceteris paribus, investments in organic growth is positively associated with under-reaction bias.

Hypothesis 4b (H4b): Ceteris paribus, investments in organic growth is negatively associated with over-reaction bias.

***Firm's Inorganic Growth**:* Apart from organic growth, firms often undertake inorganic growth activities like mergers and acquisitions for rapid expansion into new business lines or new geographies. Both product line extensions and geographic expansions through mergers or acquisition activities significantly impact firm structure and often change the structural boundaries of firms (Hayward, 2002; King et. al., 2004). Structural distribution of situated attention in firms undergoing structural changes through mergers and acquisitions is more likely to be on such

activities which impact the inorganic growth process directly than on issues that do not. Hence, the relative level of attention to market feedback on high tech innovations-in-use is likely to be low leading to higher likelihood of under-reaction bias.

Hypothesis 5a (H5a): Ceteris paribus, investment in inorganic growth activities is positively associated with under-reaction bias.

Hypothesis 5b (H5b): Ceteris paribus, investment in inorganic growth activities is negatively associated with over-reaction bias.

Product Characteristics

Firm's Product Portfolio*:* A firm's product line depth and breadth is a significant predictor of individual product failure (Thirumalai and Sinha, 2011). Both the number of product lines that a firm has and the product portfolio index, measured by the entropy index of a firm's product line depth and breadth also determines the relative saliency of each individual products.

Number of Product Lines: Number of product lines of a firm determines the business diversity of the firm. Multidivisional or multi-business firms have higher structural distribution of resources. Hence, the relative saliency of market feedback on individual products is much less as compared to firms with less number of product or business lines. Hence, diversified firms are more likely to exhibit higher under-reaction bias than focused firms with limited business lines. In case of focused firms, such as single product firms, the relative importance and saliency assigned to market feedback on individual product lines are likely to be high, and, hence, the likelihood of over-reaction bias may be higher compared to multi-business firms.

Hypothesis 6a (H6a): Ceteris paribus, the number of product lines of a firm is positively associated with under-reaction bias.

Hypothesis 6b (H6b): Ceteris paribus, the number of product lines of a firm is negatively associated with over-reaction bias.

Product Portfolio Index*:* Higher levels of product depth and breadth reduces the relative saliency of each product. Hence, the structural attention that firms are able to allocate to user

feedback on individual products is relatively lower than for firms with lower levels of product portfolio index. Hence, we hypothesize that higher levels of product line index is associated with higher likelihood of under-reaction bias.

Hypothesis 7a (H7a): Ceteris paribus, product portfolio index is positively associated with under-reaction bias.

Hypothesis 7b (H7b): Ceteris paribus, product portfolio index is negatively associated with over-reaction bias.

Market Characteristics

Market Competition*:* Product market competition impacts how the firms react to external events and external feedback (Hoffman and Ocasio, 2001). Higher levels of competition often necessitates firms to focus on market and user related issues than internal issues. Hence, higher levels of product market competition increases the attentional focus of firms to market feedback and user reports of adverse events. Hence, higher competition is likely to reduce under-reaction bias and increase over-reaction bias.

Hypothesis 8a (H8a): Ceteris paribus, market competition is positively associated with over-reaction bias.

Hypothesis 8b (H8b): Ceteris paribus, market competition is negatively associated with under-reaction bias.

3.2.5 An Integrative Framework Judgment Bias in Detecting Failures of Innovations-in-Use

We integrate the hypotheses posited above into a framework of judgment bias for detecting failures of high tech innovations-in-use. We propose the integrated framework at two levels of abstractions: (i) at a conceptual level and (ii) at a measurement level.

At the conceptual level, we highlight the underlying similarities in the three theoretical perspectives – signal detection, system neglect, and attention based view of the firm – we have used to identify and analyze judgment bias, and, in turn, attempt to synthesize the three theoretical perspectives to shed light on the evaluation of judgment bias in detecting failures of high tech

innovations-in-use. As depicted in Figure [3.1], we contend that there are two latent factors related to the differential attention process. The *first* factor corresponds to the sources that lead to high noise either in the data-stream, i.e., the noise-to-signal ratio or in the situated environment of decision makers such as firm size, product portfolio index, product line diversity and growth related efforts cause decision makers to pay differential attention to user reported adverse events and exhibit either under-reaction or over-reaction bias. The *second* factor corresponds to the perceived cost of judgment error either in the form of severity of adverse events reports or in the form of market competitiveness influences the type judgment bias. Figure [3.2] depicts the measurement level integrative framework that translates the conceptual level framework (Figure [3.1]) into operational level of constructs and hypotheses.

Figure 3.1 An Integrative Conceptual Level Framework of Judgment Bias Pertaining to Detection of High Tech Innovation Failure from User Reported Adverse Events

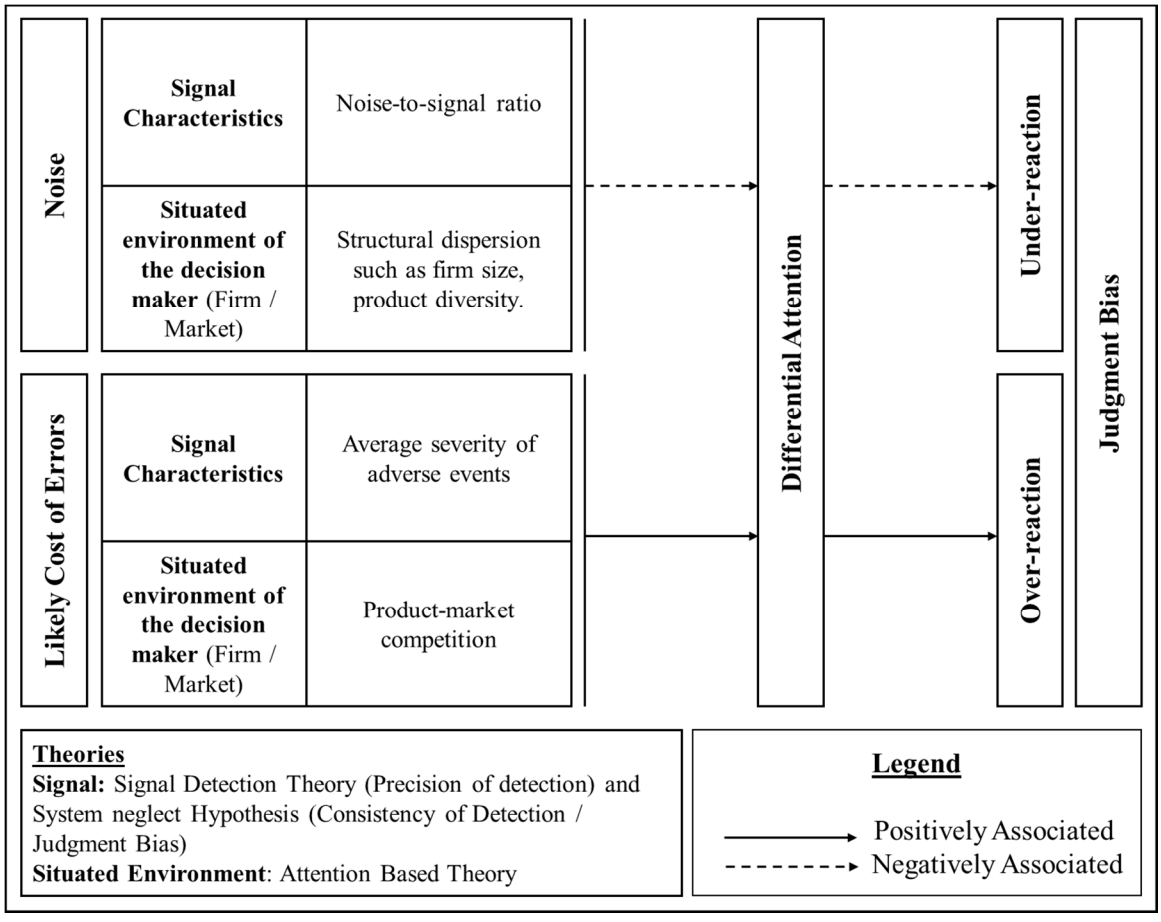
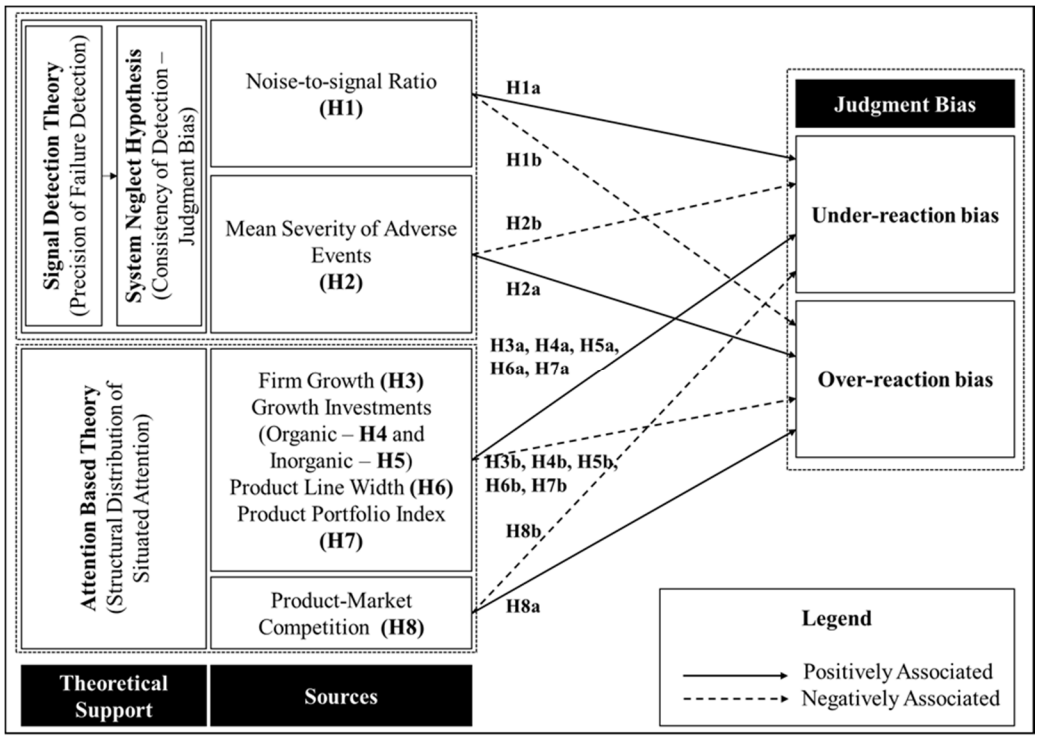


Figure 3.2 An Integrative Measurement Level Framework of Judgment Bias Pertaining to Detection of High Tech Innovation Failure from User Reported Adverse Events



3.3 Empirical Setting and Research Design

The empirical setting for the study is the medical device industry. As mentioned earlier, the data-set that is central to conducting this study is the Manufacturer and User Facility Device Experience (MAUDE) database collected by Food and Drug Administration (FDA). Our sample contains MAUDE reports generated between the year 2002 and 2012 and has upwards of *three million usable data points*. While MAUDE database contains detailed point-of-use information pertaining to individual medical devices, the information is text-based, unstructured, with significant missing information. The MAUDE database is a "big" data-set. Along with the MAUDE database, we tapped into several other databases to assemble the data-set for this study. The other databases are: the recall database, the device classification and registration database, and the Compustat database. The research sample contains 108 firms and 1348 devices identified by unique firm code and device code.

3.3.1 Unit of Observation and Analysis

The primary unit of observation is a product identified by a three letter product code assigned by FDA for each product. Individual products are nested within the level of 'firms'. Each product is associated with several firms and each firm is associated with several products. The primary unit of analysis is a product model that is identified by the combination of a firm code and a product code. User generated adverse events and firm's recall decisions are observed at the level of a product model. The firm related variables are observed at the level of individual firms.

3.3.2 Dependent Variable: Judgment Bias and Measurement of Judgment Bias

The concept of judgment bias stems from the fact that firms either fail to detect failures of high-tech innovations-in-use such as several medical devices from market signals of user reported adverse events, thus missing out on credible failure signals or react too fast to market signals when there is no underlying source of failure, thereby generating false alarms. The ability of decision makers to detect failure signals from user-reported adverse events is critical to making a rational failure judgment of high-tech innovations-in-use in the market. Hence, the first step to estimating judgment bias is to be able to predict the likelihood of failure of high-tech innovations-in-use with precision and consistency from user feedback on adverse events. In the ensuing sub-section, we briefly trace the steps in developing a consistent model of prediction and detection of failure of high-tech innovation from user-reported adverse events in the medical device industry.

3.3.2.1 Consistent Estimation of Judgment Bias from Data.

The critical factor that determines the optimal decision threshold is λ the natural logarithm of the ratio of the costs associated with false alarms and misses (equation [3.2]). The relative marginal costs associated with the false alarms and misses depend on the technological complexity and usage risks associated with a specific device. The expected false alarm rate (*FAR*) for any device i is given by $\mathbb{E}[FAR] = 1 - \Phi(X_c|0, \sigma^2)$ and the expected miss rate (*MR*) is: $\mathbb{E}[MR] = \Phi(X_c|\mu_s, \sigma^2)$. For any device where the technology associated factor λ_i is zero, i.e., the costs associated with a false alarm and a miss are equivalent, the expected number of false alarms and misses would be equivalent, i.e., $\mathbb{E}[FAR] = \mathbb{E}[MR]$. However, in reality, if we observe that decision makers are making more of one or the other, we can compute the judgement bias as a mean

difference between the two error rates, i.e., $Bias_i = \frac{Z(FAR)-Z(MR)}{2}$. For devices where the associated technology factor is not zero, we have $\mathbb{E}[MR] = \mathbb{E}[FAR] + \beta\lambda_i$, where $\beta = \frac{\sigma^2}{\mu_s}$ a characteristics of the signal is. In such a case the judgment bias that decision makers exhibit is estimable as a mean difference between the two error rates as above after controlling for technology fixed effect pertaining to a specific device, i.e.,

$$Bias_i = \frac{Z(FAR) - Z(MR)}{2} + \beta\lambda_i \qquad \ldots [3.4].$$

The MAUDE database along with other related databases such as the recall database provides us with the ability to predict medical device recalls (i.e., failures of high tech innovations-in-use) with precision and consistency, and estimate judgment bias exhibited by the decision makers.

3.3.2.2 Data Organization and Variable Generation

As mentioned earlier, the primary data-set we are using for this study is the Manufacturer and User Facility Device Experience (MAUDE) data-set of the Food and Drug Administration (FDA). Each year, FDA receives several hundred thousand adverse event reports of suspected device-associated deaths, serious injuries and malfunctions. The MAUDE database houses the adverse event reports to FDA by mandatory reporters (manufacturers, importers and device user facilities) and voluntary reporters such as health care professionals, patients and consumers. MAUDE is a "big" database and a rich source of information with millions of data points generated through reports of adverse events involving the usage of medical devices. MAUDE is also an unstructured database, since the adverse event reports are uncodified text; and a noisy database, since the reports submitted can be incomplete, inaccurate, untimely, unverified, or biased data.

Apart from the MAUDE and patient databases, we assembled data from several other sources. The main databases are: (i) the recall database; (ii) the 510(K) and PMA databases – where 510(K) is premarket notification, i.e., the approval process of new models of existing devices, and PMA is pre-market approval, i.e., the approval process of new devices; (iii) the facility registration database; and (iv) COMPUSTAT database from Wharton Research Data Services (WRDS). A complete list of the databases used for the empirical analysis is provided in Table [2.1] in Chapter 2.

Data organization represents an important step towards building a predictive model. As mentioned, the MAUDE database is primarily text base and also incomplete in its codification. We used text analytics to organize the databases and link different databases such as the MAUDE and

the recall database. Through an extensive data organization and data classification exercise we generated the following variables for the purpose of the predictive model building.

3.3.2.3 Response Variable

The unit of analysis for the empirical investigation in this study is a failure of high tech innovation-in-use – i.e., a medical device recall. Every medical device model in the database is identified by a unique identifier which is a combination of the device code and the firm code. By way of an illustrative example, a Medtronic defibrillator model is identified by the combination of the general device code for defibrillators and the firm code for Medtronic. We did not distinguish between versions of the same device since that information is already captured in the number of FDA approvals for the device. The response variable is a binary recall variable with the unit of time being a quarter (three months). If a device has been recalled within a quarter, then the value of the response variable is 1 for the quarter. Otherwise, the value of the response variable is zero.

3.3.2.4 Predictor Variables

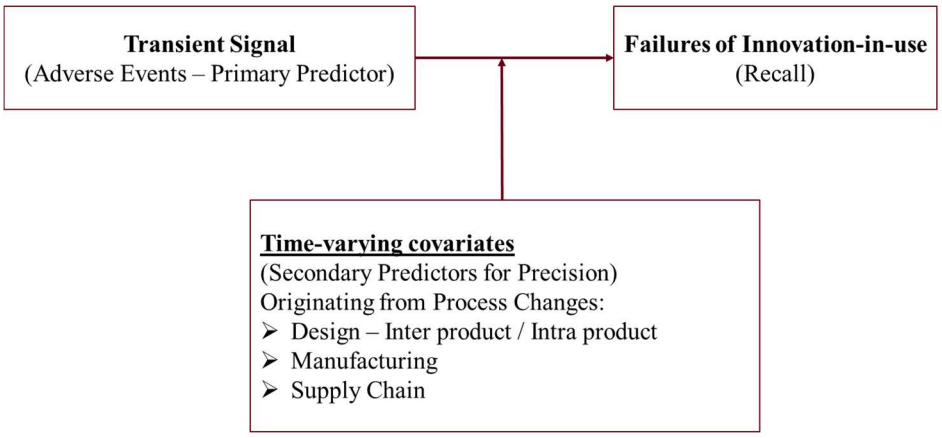
The primary predictor variable is the kernel density of adverse events given by $f_t = \left(\frac{1}{Nh}\right)\sum_{i=1}^{N} I_i * K\left(\frac{t-T_i}{h}\right)$. Taking the kernel density instead of a discretized frequency count has several advantages. Firstly, it allows a continuous time filtering of the signal. Secondly, kernel densities automatically normalize the data pertaining to different products and product classes. We also created severity weighted kernel densities to account for severity of each adverse event. Severity rating was taken from FDA classification of adverse events into one of five severity classes starting from death to no-harm to patients. We also created the 100 days and 300 days lagged variables for both kernel densities and severity weighted kernel densities.

Apart from the primary predictors we also generated variables pertaining to covariates related to design, supply chain and manufacturing. While the kernel densities of the adverse events are the primary predictors, information related to design, supply chain and manufacturing captured through the respective covariates is likely to improve the precision of such prediction. Apart from the covariates we included several control variables for the purpose of building the predictive models.

3.3.2.5 Prediction Modeling Framework

Figure [3.3] provides an overview of the prediction modeling framework. The conceptual framework provides an outline of the predictive model. In spite of the best efforts of the firms and regulators in the form of approval testing, performance testing and quality inspections, many products encounter failures often several times during their effective life-cycle while in use in the market. Failure of innovations-in-use is a result of some problems that unforeseeably creep in one or more of the components in their value chain, namely, design, manufacturing and supply-chain. Any perturbation in the value chain can in turn cause potential failure of the products which manifest in the form of user generated signals of failure, i.e., a shift in the adverse effect distribution from the pure stationary white noise distribution of adverse effects which is expected under normal working conditions even when there is no system level source of failure to a noise-plus-signal distribution. Hence, the primary predictor of a failure is the density distribution of the adverse event reports from users. When the adverse event reports are supplemented by system level covariates like design covariates, supply chain covariates and manufacturing covariates the precision of prediction is likely to improve further. The prediction framework conceptually captures this system dynamics and interaction of the primary signal of failure prediction and the underlying signal generating process. This framework guides the choice of the predictive models which can capture the interaction of the primary signal as well as the important covariates related to the underlying signal generating process. The choice of the theoretical framework of system neglect and managerial attention process has also been guided by this system dynamics that is captured in the conceptual framework of Figure [3.3].

Figure 3.3 Framework for Predictive Model Building for Prediction of Failure of High-tech Innovation in the Medical Device Industry from User Reported Adverse Events.



Analytically, the framework incorporates the following structural dynamics of risk of failures of innovation-in-use. Let us assume that X_{it} (where $i \in \{1, \dots, N\}$ is a subscript for a product and t is a subscript for time) denotes the time-varying covariates originating from the under-lying processes such as design, supply-chain and manufacturing that generate an innovation. Also, let $Y_{it}, i \in \{1, \dots, N\}$ denote the transient signal, i.e., the adverse event reports corresponding to product i at time t. For purposes of normalization of the signals and covariates corresponding to all products let us consider the kernel transformations of the signals, i.e., $K_h(Y_{it})$ and covariates $K_h(X_{it})$, where $K_h(\cdot)$ is an appropriate kernel function such as the Gaussian kernel with h being the bandwidth or the resolution of the kernel function. A change or perturbation of the covariates would be associated with the risk of a failure, say R_{it} given by $R_{it} = \phi_i\left(\frac{\partial K_h(X_{it})}{\partial t}\right)$, where $\phi_i(\cdot)$ is a probabilistic risk function associated with product $i \in \{1, \dots, N\}$. However, the failure is manifested in the transient signals of adverse events. Hence, transient signals are indicative the presence of a failure in the underlying system that generates a product and the risk is measurable through a risk function $R_{it} = \delta_i(\Delta[K_h(Y_{it})])$. Hence, the interaction of the two indicators of failure risk is more predictive of failure of a innovation-in-use. Hence, the risk of failure of an innovation is $R_{it} = F_i\left[\frac{\partial K_h(X_{it})}{\partial t} \times \Delta[K_h(Y_{it})]\right]$. The risk function associated with a product i can be estimated from data using various machine learning models such as support vector machines, random forest or neural network. The choice of the appropriate model would be guided by data and predictive ability of models on hold-out test samples.

3.3.2.6 Variable Selection

Variable selection is a key step in predictive model building (Shmueli, 2010). The criterion for selecting the right set of variables is predictive association of a variable with the response variable (Shmueli, 2010). From a number of likely predictors generated from data, a subset of variables are chosen to be included in the actual prediction stage. This is important from the point of view of predictive model parsimony – i.e., to avoid over-fitting and reduce prediction error.

One of the most commonly used approaches to variable selection is a shrinkage method in fitting a linear model such as Least Absolute Shrinkage and Selection Operator (LASSO) which uses a L_1 penalization on a maximum likelihood estimation and Elastic Net which uses a combination of L_1 and L_2 penalization on a maximum likelihood estimation (Zou, 2006). Other common methods are stepwise regression estimates using information criteria such as Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) for variable selection at each step of model estimation process. We considered a triangulation of LASSO, stepwise generalized Linear Mixed Model (GLMM) and stepwise Generalized Additive Model (GAM). The triangulations helped in removing unnecessary variables while selecting the variables that are predictors of the response variable. Table [2.2] in Chapter 2 provides the list of the selected predictive variables after the variable selection process.

3.3.2.7 Predictive Model Building

We used a random forest model for building the predictive model based on the selected variables from the variable selection step. A random forest model is a decision tree based model that is built on the data by recursive random partitioning of the data while minimizing the out-of-sample prediction error of the model. Random forests work by generating ensembles of regression trees built on independent random subsamples of the training data (Breiman, 2001). The classification ensemble is generated by the modal prediction class. It has been shown that the classification accuracy of random forests depend on the number of classification trees, i.e., the size of the ensemble. The prediction error asymptotically converges almost surely to a limit as number of trees in the forest becomes large. Random forests have been widely used in machine learning, bioinformatics, climate sciences and other natural sciences. Using several data-sets from economics and environment science, Breiman (2001) has demonstrated that random forests achieve significantly superior prediction accuracy over single classifiers or even over other ensemble class

classifiers. The random forest model inherently takes into account the important non-linear interaction effects that exist within the selected variables and hence is suitable for modeling in light of the conceptual model illustrated in Figure [3.3] earlier. The model equation is stated in equation [3.5].

$$P(Recall) \sim F\begin{pmatrix} Maude.Kernel, Control\ Variables. \\ Design\ Covariates, \\ Supply\ Chain\ Covariates. Manufacturing\ Covariates \end{pmatrix} \quad \ldots [3.5]$$

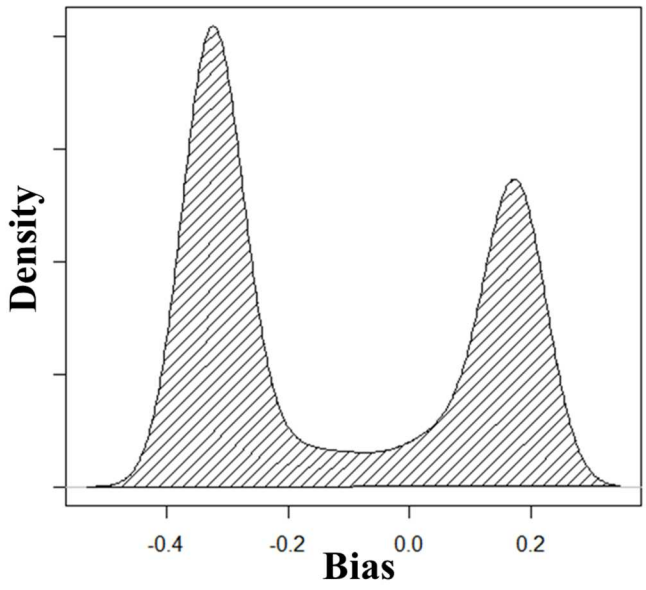
To build the model we used a 80:20 split of the sample into train and test set. The train set is used to estimate the model parameters and the test set is used to test the model accuracy. To account for over-fitting and consistency issues we used a 10 fold cross validation on the training sample (Hastie and Tibsirani, 2009). We used a Receiver Operating Characteristics (ROC) curve to test the accuracy of prediction. The ROC curve is a plot between the false positive rate and the true positive rate on the test data-set. A good predictive model would maximize the true positive rate while keeping the false positive rate to be as low as possible. Hence, higher is the area under the curve (AUC) of the ROC, higher is the predictive accuracy of a model. Figure [2.11] in Chapter 2 provides the ROC curve corresponding to the random forest model. Figure [2.11] in Chapter 2 provides the median ROC curve on the test set and the 95% confidence band generated through 1000 times bootstrapping of the prediction process. The median an AUC is 0.88 with a 95% bootstrapped (1000 times) confidence band of (0.82-0.90).

3.3.2.8 Judgment Bias Estimation

Judgement bias is measured as a continuous measure following equation [3.4]. The random forest model was used to generate failure likelihood for individual products in the study sample and generate the false alarm rate and miss rate for each product at the minimum bias point in the receiver operating characteristics curve. The minimum bias is the limiting judgment bias measure for each product that has been achieved by firms in the study sample pertaining to each product. The prediction model has been estimated with product level fixed effects. The product level fixed effects measure the mean propensity of all firms to commit one type of judgment bias within that product group identified by the FDA assigned three lettered product code. The fixed effects for each product code (which contains multiple technologically similar products intended for similar usages) represents the base level technology factor λ as in equation [3.4] and thus accounts for the differential cost of errors pertaining to a product-firm tuple which is the unit of observation. The difference between the individual products specific to individual firms from the base level judgment

bias pertaining to the specific product code would provide us with a consistent measure of the judgment bias of each observation unit, i.e., individual products nested within individual firms in the study sample. We created two different measures of judgment bias for the purpose of statistical analysis, namely, *(i) Bias,* which is a continuous measure of the bias and *(ii) Bias_OR,* which is a categorical measure of bias codes as *one* for over-reaction and *zero* for under-reaction. Figure [3.4] shows the distribution of the continuous judgment bias measure, *Bias.* Figure [3.4] shows that in general firms have a much higher likelihood of under-reaction bias as compared to over-reaction bias. This observation is valid both in terms of the frequency (measured by kernel density of the bias) with which firms exhibit under-reaction bias as against over-reaction bias and in terms of the extent of under-reaction bias (measured in absolute numerical scale of the bias measure).

Figure 3.4 Distribution of Bias Measure



3.3.3 Independent Variables[6]

The following independent variables were generated for the purpose of statistical analysis.

Noise-to-signal ratio*:* This is measured as the ratio variance of the noise-plus-signal distribution

[6] The terms "product" and "medical device" are used interchangeably

and the mean of the noise-plus-signal distribution. This variable pertains to hypothesis 1 (H1a & H1b).

***Severity**:* This is the weighted average of the severity of adverse events pertaining to each product firm tuple. This pertains to hypothesis 2 (H2a & H2b).

***Firm Revenue**:* This is a logarithm of the mean firm revenue for each firm over the study time-frame. This revenue measure accounts for the changes in firm structures due to spin-offs, mergers and acquisitions over the years. We use logarithm of the mean firm revenue as a measure of firm size pertaining to hypothesis 3 (H3a & H3b).

***CAGR Fixed Asset**:* This is a measure of the cumulative average growth rate of the gross fixed assets of a firm excluding assets acquired through mergers or acquisitions but including assets diluted. This measures primarily the increase or decrease of manufacturing assets of a firm. We use the CAGR of fixed assets as a proxy measure of a firm's internal growth in manufacturing activities, thus measuring organic growth rate of a firm. This pertains to hypothesis 4 (H4a & H4b).

***Mergers Acquisitions**:* This measures a firm's gross investment in mergers and acquisition activities over the study time period. We use the natural logarithm of the gross investment in mergers and acquisitions for the purpose of the analysis. This pertains to hypothesis 5 (H5a & H5b).

***Product Line**:* This is the count of the mean number of product lines each firm has during the study period. Product lines are identified by the specific usage class that a firm is present in. This variable pertains to hypothesis 6 (H6a & H6b).

***Product Index**:* This measures a firm's mean product index measured by the entropy index of the product portfolio following Thirumalai and Sinha (2011). This measures the depth and breadth of product portfolio of a firm and pertains to hypothesis 7 (H7a & H7b).

***Product Competition**:* This is a numeric count of the average number of competing product models within each product code. This measures the product market competition and pertains to hypothesis 8 (H8a & H8b).

3.3.4 Control Variables

We created the following control variables that are likely to influence the variation in the response *Bias* and *Bias_OR:* (i) *Regulation Type PMA,* a categorical variable coded as *one* for pre-market approval (PMA) corresponding to approval routed for new products and *zero* corresponding to approval route for new versions of existing products; (ii) *log product age,* natural logarithm of the number of years a product is in the market; (iii) *product class,* a categorical variable corresponding

to FDA classification of complexity class of products with three possible values, namely, 1 corresponding to least complex products, 2 corresponding to medium complexity products and 3 corresponding to high complexity products; (iv) *implant,* a categorical control coded as *one* for implantable devices and *zero* otherwise; (v) *firm,* firm fixed effects; (vi) *CAGR growth,* a continuous measure of the cumulative average growth of gross operating revenue of a firm; (vii) *change operating margin,* mean year on year change operating margin of a firm over the study period; (viii) *foreign,* a categorical variable coded as *one* for domestic firms and *zero* for foreign firms; (viii) *inventory turnover,* a continuous measure of the mean inventory turnover of a firm measured in number of days of inventory; and (ix) *Usage Class:* a categorical variable pertaining to the 19 usage classes of the products. These usage classes are assigned by FDA at the time of product registration.

3.3.5 Model Specification

We estimated three different models to test the hypotheses. First we estimated a base model with firm fixed effect specified in equation [3.6].

$$\begin{aligned} Bias = \beta_0 + \beta_1(Noise - to - Signal\ Ratio) + \beta_2(Severity) \\ + \beta_{12}(Noise - to - Signal\ Ratio \times \text{Severity}) + \beta_3\big(Log\ Product\ Age\big) \\ + \beta_4\big(Product\ Competition\big) + \beta_5\big(Product\ Class\ II\big) \\ + \beta_6\big(Product\ Class\ III\big) + \beta_7\ \big(Regulation\ Type\ PMA\big) \\ + \sum_{i=8}^{27} \gamma_i(Usage\ Class_i) + \sum_{i} \delta_i(Firm_i) + \epsilon \quad \ldots [3.6] \end{aligned}$$

To test for the firm related sources of judgment bias we used a hierarchical linear modeling (HLM) set-up with firm random effects considering the multi-level effects that we are interested in estimating and the nested structure of the variables. The HLM model specification is shown in equation [3.7].

$$Bias = \beta_0 + \beta_1(Noise - to - Signal\ Ratio) + \beta_2(Severity) + \beta_{12}(Noise - to - Signal\ Ratio \times Severity) + \beta_3(Log\ Product\ Age) + \beta_4(Product\ Competition) + \beta_5(Product\ Class\ II) + \beta_6(Product\ Class\ III) + \beta_7\ (Regulation\ Type\ PMA) + \sum_{i=8}^{27} \gamma_i(Usage\ Class_i) + \sum_i \delta_i(Firm_i) + \epsilon$$

$$\delta_j = \gamma_0 + \gamma_1(Revenue_j) + \gamma_2(CAGR\ Fixed\ Assets_j) + \gamma_3(Mergers\ Acquisitions_j) + \gamma_4(Number\ of\ Product\ Lines_j) + \gamma_5(Product\ Index_j) + \gamma_6(Foreign_j) + \gamma_7(CAGR\ Growth_j) + \gamma_8(Change\ in\ Operating\ Margin_j) + \gamma_9(Inventory\ Turnover_j) + \gamma_{10}(Inventory\ Turnover_j) + \eta_j \quad \text{... [3.7]}$$

To check for the robustness of model estimations we estimated a generalized linear mixed effects model with logit link function for the categorical measure of bias, *Bias_OR* as in equation [3.8].

$$log\left(\frac{P(Bias\ OR = 1)}{1 - P(Bias\ OR = 1)}\right) = \beta_0 + \beta_1(Noise - to - Signal\ Ratio) + \beta_2(Severity) + \beta_{12}(Noise - to - Signal\ Ratio \times Severity) + \beta_3(Log\ Product\ Age) + \beta_4(Product\ Competition) + \beta_5(Product\ Class\ II) + \beta_6(Product\ Class\ III) + \beta_7\ (Regulation\ Type\ PMA) + \sum_{i=8}^{27} \gamma_i(Usage\ Class_i) + \gamma_1(Revenue_j|Firm_j) + \gamma_2(CAGR\ Fixed\ Assets_j|Firm_j) + \gamma_3(Mergers\ Acquisitions_j|Firm_j) + \gamma_4(Divisions_j|Firm_j) + \gamma_5(Product\ Index_j|Firm_j) + \gamma_6(CAGR\ Growth_j|Firm_j) + \gamma_7(Foreign_j|Firm_j) + \gamma_8(Change\ in\ Operating\ Margin_j|Firm_j) + \gamma_9(Inventory\ Turnover_j|Firm_j) + \eta_j + \epsilon \quad \text{... [3.8]}$$

The model estimation results are shown in Table [3.1].

Table 3.1 Model estimation results for judgment bias

	Continuous Bias Measure (Higher value indicates higher likelihood of over-reaction)		**Categorical Bias Measure (1: Over-reaction, 0: Under-reaction)**
	Firm Fixed Effect	**Firm Random Effects (Hierarchical Linear Model, Bootstrap robust error distribution)**	**Firm Random Effects (MCMC[2] GLMM Model)[1]**
	Equation [3.6]	Equation [3.7]	Equation [3.8]
	Random Effects Variance Components (Chi-square p-value)		
(Intercept)		0.08444 (0.036)*	
Firm:Revenue		0.09178 (0.0209)*	
Firm:Cagr Growth		0.0144 (0.3721)	
Firm:Foreign USA		0.1652 (0.0001)***	
Firm:CAGR Fixed Asset		0.1163 (0.001)***	
Firm:Change Operating Margin		0.07361 (0.0501)*	
Firm:Inventory Turnover		0.0023 (0.4531)	
Firm:Mergers Acquisitions		0.1843 (0.0001)***	
Firm:Divisions		0.0934 (0.018)**	
Firm:Product Index		0.0895 (0.025)**	
	Fixed Effects Estimates (t-statistic based p-Values)		
(Intercept)	-0.0527 (0.0000)***	-0.0588 (0.000)***	-0.0352 (0.013)*
Severity	-0.0041 (0.1634)	-0.0047 (0.206)	-0.0079 (0.089)+
Noise-to-Signal Ratio (Variance)	-0.0379 (0.051)*	-0.0401 (0.082)+	-0.0017 (0.045)*
I(Severity:Noise-to-Signal Ratio)	0.1189 (0.003)**	0.1011 (0.0278)*	0.0994 (0.005)**
Log Product Age	-0.0049 (0.0151)*	-0.0058 (0.0034)**	-0.0832 (0.0042)**
Product Competition	0.2002 (0.0000)***	0.2627 (0.0000)***	0.0027 (0.0000)***
Product Class 2	0.0217 (0.0016)**	0.0743 (0.0006)***	0.0127 (0.0000)***
Product Class_3	0.1189 (0.0000)***	0.1816 (0.0000)***	0.0866 (0.0000)***
Implant 1	0.0179 (0.0317)*	0.0151 (0.0533)*	0.0309 (0.0833)+
Regulation Type PMA	0.0016 (0.081)+	0.0009 (0.074)+	0.0012 (0.1739)
Usage Class (Fixed Effects)	S	S	S
Usage Class_(…)	-	-	-
Usage Class SURGICAL (SU)	0.0367 (0.0012)**	0.0427 (0.002)**	0.05215 (0.0021)**
Usage Class CARDIO_VASCULAR (CV)	0.0634 (0.0001)***	0.0720 (0.0091)**	0.0945 (0.0003)***
Usage Class LAB CHEMICAL (CH)	-0.0032 (0.042)*	-0.0015 (0.036)*	-0.0757 (0.0623)+
Usage Class DENTAL (DE)	-0.0073 (0.164)	-0.0050 (0.1845)	0.0002 (0.345)
Usage Class_(…)	-	-	-
Firm (Fixed Effects)	S	-	-
Revenue		-0.0435 (0.003)**	-0.0073 (0.0455)*
CAGR Growth		-0.0024 (0.132)	-0.0001 (0.224)
Foreign USA		0.0981 (0.0032)**	0.0289 (0.072)+
CAGR Fixed Asset		-0.0001 (0.042)*	-0.0015 (0.072)+
Change Operating Margin		0.0014 (0.175)	-0.0061 (0.3224)
Inventory Turnover		0.0001 (0.245)	0.0044 (0.146)
Mergers Acquisitions		-0.0164 (0.0245)*	-0.0143 (0.061)+
Divisions		-0.0164 (0.0035)**	-0.099 (0.0513)*
Product Index		-0.0068 (0.002)**	-0.0193 (0.0224)*
Number of Products	1348	1348	1348
Number of Usage Class	19	19	19
Number of Firms	108	108	108

Note:

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1

[1] Only the posterior parameter estimate is shown.

[2] Max. 60,000 iterations with initial 10,000 burnout iterations

3.4 Results and Discussions

3.4.1 Model Estimation Results

Table [3.1] shows the results of the model estimation. Like the corresponding results in Chapter 2 (Table [2.5]) we find that the interaction of *Severity* and *Noise-to-Signal Ratio* is significant in explaining judgment bias. Also, the main effect of *Noise-to-Signal Ratio* is significant in explaining judgment bias (H1). The sign of the main effect of *Noise-to-Signal Ratio* has a negative sign indicating that as a main effect, high *Noise-to-Signal Ratio* leads to higher likelihood of over-reaction to adverse event reporting from users (H1b). However, when *Noise-to-Signal Ratio* interacts with *Severity* of the adverse events, the over-reaction likelihood increases due to the positive sign of the interaction term. High severity of adverse event reports from users leads to higher likelihood for firms to become risk averse and incline towards higher rate of false alarm than misses (H2a). This is supported by all the three model estimations using both the continuous and categorical measure of bias. From the density of estimation of bias measure (Figure [3.4]), we see that generally there is a higher likelihood of firms to under-react to market signals than over-react.

A firm's size, measured by the gross operating revenue of a firm, is significant in explaining judgment bias exhibited by the firm. We see that larger size firms tend to significantly increase under-reaction likelihood of firms. Firm's and decision makers react to different signals coming from their internal or external environment. However, firms do not react equally to all signals coming from their environment, rather they react more to certain signals and ignore certain others. The differential nature of firms' reaction is determined by saliency and immediacy of impact of different issues. Larger firms are encountered with a high number of signals from the environment and hence the likelihood that a firm under-reacts to a specific signal with moderate or low saliency to the firm is low. This supports hypothesis 3 (H3a) that large firms are likely to under-react more than over-react to market generated signals of potential product failures.

Firms undertaking high rate of organic or inorganic growth activities are also significantly more likely to under-react to market signals of product failures. Given limited managerial capacities, the relative saliency and immediacy of impact on the firm for growth related investments causes firms to assign higher levels of situated contextual attention to such activities. The slope of the model estimation parameters for *CAGR Fixed Assets* and *Mergers Acquisitions* are significant and negative in all three model estimations shown in Table [3.1]. Mergers and acquisitions often lead

to structural discontinuities in firms. Anand, Gray and Simesen (2012) analyses inorganic growth activities like mergers and acquisitions in the pharmaceutical industry and show that such investments lead to structural discontinuities which lead to decay of operational routines specifically related to maintaining and improving product quality. Our study supports and complements the study by Anand, Gray and Siemsen (2012) in extending the relationship between structural breaks in firms and their ability to react optimally to market generated signals related to product performance and failure. This supports both hypothesis 4 (H4a) and hypothesis 5 (H5a).

We find that firms with a higher number of product lines are more likely to under-react to user reported adverse events compared to more focused firm (H6a). Thirumalai and Sinha (2011) finds that higher value of entropy index of product portfolio of a firm leads to higher likelihood of product failure. Our analysis shows that higher value of product portfolio index is also associated with firms' ability to react adequately to user generated signals, i.e. adverse events, of product failures. The parameter estimate for *Product Portfolio Index* is negative and significant in both model estimate (2) for continuous measure *Bias* and (3) for categorical response measure *Bias_OR*. This supports hypothesis 7 (H7a). Also, higher levels of market competition is associated with higher likelihood of over-reaction to market signals. Competition necessitates firms to constantly be aware of device performance while they are in use with users. This increases the saliency of user generated adverse event signals and hence, the attention that firms and managers allocate for such signals is also relatively higher compared to firms in low product market competition. The slope of *Market Competition* is negative and significant. This supports hypothesis 8 (H8a).

Apart from the above, usage class of products is a statistically significant explanatory for differential likelihood of under-reaction bias and over-reaction bias. The intended usage of a device significantly determines how firms react to adverse event reports related to the device. Certain usage class increases the over-reaction likelihood of devices and certain other usage class leads to higher under-reaction likelihood. Usage classes such as cardio-vascular (CV), orthopedic (OR) and surgical (SU) significantly increases over-reaction likelihood and usage classes such as clinical chemistry (CH) and dental (DN) increases the under-reaction likelihood significantly. The overall risk to patients associated with a device while in use is an important consideration for firms as well as regulators. Hence, the reaction towards market feedback of devices is different for different usage classes. This partly explains why majority of the device recalls made by firms are concentrated within a few classes, namely, cardio-vascular, orthopedic, radiology and surgical, among the 19 device usage class. We also see that firms tend to over-react significantly to market

signals related to new products (PMA approved) than when the signals are related to new versions of existing products (510K). Firms' likelihood to over-react is high in case of implants as compared to non-implants. Product class which is a categorical measure of product complexity and criticality is a significant predictor of type and extent of judgment bias exhibited by firms. Also domestic firms are less likely to under-react to market feedback than foreign firms. We did not find any significant effect of firm profitability, inventory turnover or profitability growth of firms.

3.5 Conclusions

This study was motivated by incidences of failures of high tech innovations-in-use, suggesting that firms exhibit judgment bias (under-reaction or over-reaction) in reacting to user feedback on adverse events related to the innovations. Drawing on, and synthesizing the theoretical perspectives of signal detection, system neglect, and attention based view of the firm, we developed an integrative theoretical framework for identifying the sources of judgment bias in detecting failures of high tech innovations-in-use. The empirical setting of this study was the medical device industry. We investigated the failures of high tech innovations-in-use by way of medical device recalls. Specifically, we investigated recalls of new medical devices and new versions of existing medical devices. We analyzed user reports (big and unstructured data) on adverse events related to medical devices using a combination of econometric and predictive analytic (machine learning) methods to test the constituent relationships (hypotheses) of the integrative framework. The primary contribution of this study is the integrative framework which identifies the sources of judgment bias (under-reaction and over-reaction) exhibited by firms in reacting to user reports on adverse events related to high tech innovations-in-use.

The key insights from the study results are as follows. First, we showed that decision makers in firms exhibit judgment bias in reacting to market signals of potential failures of high tech innovations-in-use. Either firms over-react or under-react to user feedback on adverse events related to high tech innovations-in-use. Next, we identified specific firm, product, and market conditions in which firms are likely to under-react or over-react to the market signals.

We believe that the results of this study, based on theoretically-grounded rigorous empirical analyses, will broaden the acknowledgement of the existence of judgment bias in firms as a major barrier to the timely detection of failures of high technology innovations-in-use. A significant implication of this study will be in informing firms and regulators (e.g., FDA and GAO) about the sources of judgment bias in detecting failures of high tech innovations-in-use. The need for making

the detection of failures of medical devices more proactive, consistent and predictive has been expressed by the regulatory agencies such as FDA and GAO. In a report to the United States Senate, GAO has concluded that FDA should use data on device usage available to them for better analysis and proactive management of device failures to minimize public health risks from failures of medical devices (GAO-11-468; June, 2011; p. 35). In a subsequent study, GAO has indicated that user feedback on adverse events related to usage of medical devices, the big data that we mine in the study, can serve as effective signals for detection of risk of potential failures of medical devices (GAO-12-816 , p. 28, 33, 45). Towards that end, the findings of this study will inform firms and regulators (e.g., FDA and GAO) about the sources of judgment bias and improve the post-launch market surveillance of high-tech innovations-in-use (medical devices) by making it more evidence-based and predictive, thereby contribute towards addressing the grand challenge of safe and cost-effective health care delivery that is, increasingly, becoming enabled by high tech innovations.

Chapter 4:

Enabling Health Care Delivery with High Tech Innovation: A Longitudinal Field Study of Robot-Assisted Surgery

4.1 Introduction

> *"I'm very impressed," said Dr. David Shepherd, a Fort Worth, TX urologic surgeon. "I watched a gentleman before me use it (da Vinci surgical robot) and asked him if it was his first time. And he was throwing some sutures in, and he said it was, and he was doing a fine job."*[7] *(Source: NBC News, Nov 5, 2012)*

Variation in the outcomes of medical procedures has been identified as a key concern in health care delivery and management (e.g., Birkmeyer and Dimik, 2009). The effect of outcome variation is of particular concern in the context of critical and complex surgical procedures (Hannan et al., 1990). The variation in health care outcomes of complex surgical procedures has been related to several factors like surgeon experience (Diwas and Staats, 2012), procedure complexity (Hannan et al., 1990), organization (Huckman and Pisano, 2006), and team composition based on familiarity and task specificity (Huckman, Staats, and Upton, 2009). One of the most important sources of variation in surgical outcomes is the skill heterogeneity of surgeons (Ramdas et al., 2010; Waldman et. al, 2003). Skill-based outcome variation not only leads to inconvenience but also has severe economic implications in increasing the overall cost of health care delivery (Porter, 2010). Outcome variation in critical surgical procedures introduces several constraints to the overall health care delivery process. First, from the point-of-view of a patient, variation in outcome can cause lack of confidence in the health care delivery system. For complex procedures, this may result in considerable gap between demand and supply of specific surgical skills due to patient and surgeon self-selection effects. From the point-of-view of a health care delivery organization, this introduces

[7] http://www.nbcnews.com/id/3403901/ns/business-cnbc_tv/t/robots-invade-operating-room/#.VA4S3_ldV8E

several scheduling constraints based on resource availability. One of the key challenges in scheduling complex surgical procedures is to match a specific surgeon's experience-based performance likelihood with patient criticality. The same is also true for team composition and nurse scheduling. From the perspective of a health care delivery system, outcome variation of complex medical procedures not only leads to poor quality of outcome but also impacts the overall availability and affordability of these procedures to the general patient population. Hence, it is important to understand the causes and sources of some of these variations in health care procedures. Moreover, beyond understanding the sources of outcome variation in health care delivery, mitigation of some of these sources of variation is critical to improving the overall availability, cost, and quality of health care delivery services.

Extant literature in health care has looked at ways and means of reducing input heterogeneity in the form of surgeons' experience, staff experience, and team familiarity to reduce outcome variation of critical surgical procedures. However, in this paper, we take a different approach to managing the issue of outcome variation in health care. Given input heterogeneity, which is often not under the full control of health care delivery organizations, we investigate how a health care delivery organization can develop technological capability to reduce the effect of input heterogeneity on outcome variation? To that end, the questions we attempt to address in this paper through a field study are the following:

(i) *Can technology help mitigate the effect of input heterogeneity (surgeon and staff experience and skill heterogeneity) on outcome variation in critical and complex health care delivery procedures?*

(ii) *If technology can reduce the effect of input heterogeneity on outcome variation in critical and complex health care delivery procedures, how does surgeon and staff learning help in developing this technology capability?*

(iii) *What are the dimensions of surgeon learning that are central to developing technological capability for critical and complex health care delivery procedures?*

Specifically, we investigate the effect of a high tech innovation on the outcome variation of critical and complex health care delivery procedures, given input heterogeneity in the form of surgeon and staff experience and skill heterogeneity and heterogeneity in surgical team composition. The empirical setting for the study is a large multi-specialty hospital that adopted a high-tech

innovation – namely, a da Vinci surgical robot – to assist in performing a set of surgical procedures. Because surgeries are primarily surgeon-led procedures, skill and experience heterogeneity of the surgeons performing surgeries play a consequential role in determining the overall outcome variation of surgical procedures. We investigate the effect of robotic technology mediation on the outcome variation of surgical procedures, given input heterogeneity.

The rest of the chapter is organized as follows. In section two, we propose a conceptual framework for the study; in section three, we introduce the key features of the surgical robot; in section four, we develop the study hypotheses; in section five, we describe the data collection and research design; in section six, we present the empirical analysis results corresponding to the stated hypotheses; and in section seven, we perform some robustness checks for the results and in finally section eight we present our concluding remarks.

4.2 Conceptual Framework

4.2.1 Sources of Input Heterogeneity in Health Care

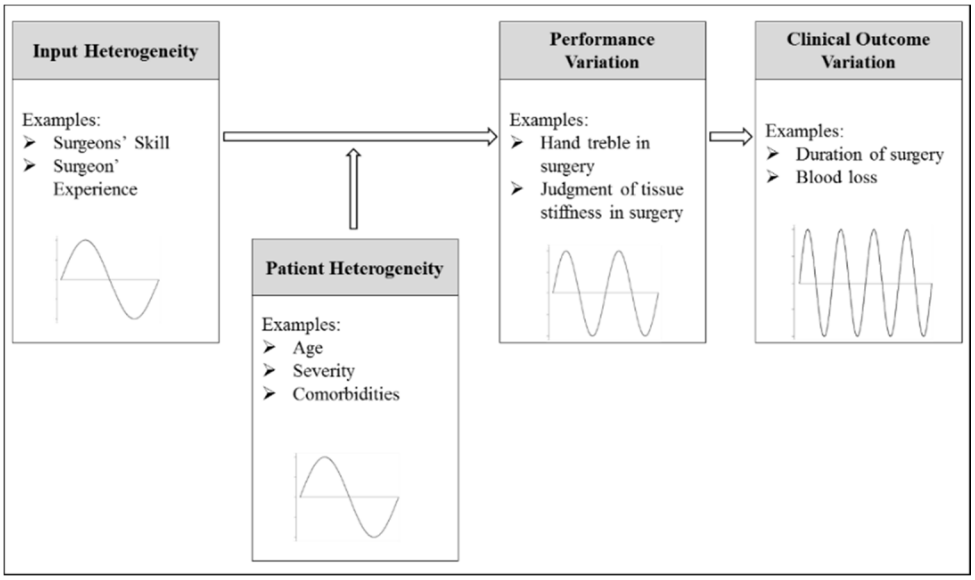
Sources of input heterogeneity in health care can be categorized into two broad classes: controllable sources (sources that can be partially or fully controlled by a health care delivery organization) and non-controllable sources. Controllable sources are related to surgeons, staff, and the surgical team such as surgeon and staff experience and team familiarity. The non-controllable sources are related to patients such as patient profiles in terms of age, severity or comorbidities. While some amount of control can be exercised on patient composition, it is still largely exogenous to the health care delivery organization.

Among the controllable sources of variation, surgeon and staff heterogeneity related to training, skill levels, and experience is a key factor. Atella, Boletti, and Depalo (2011) found through econometric analysis of a large panel dataset related to variation in drug therapy effectiveness that heterogeneity in the skill and experience of surgeons is a significant determinant of drug therapy effectiveness. They found that apart from patients' adherence to a drug regime, the surgeons' experience-based judgment heterogeneity is a critical determinant of the effect of the drug therapy on patients. Investments in increasing surgeons' knowledge, experience, and skill can reduce outcome variation significantly. Experience-based heterogeneity as a source of variation in health care outcome has been pointed out in many other studies (Diwas and Staats, 2012). Apart from individual sources of heterogeneity, another key source of input heterogeneity has been

identified as the team composition and team familiarity (Huckman and Staats, 2011). Apart from the health care management literature, in the clinical literature, the positive influence of team familiarity on surgical performance has been acknowledged. Xu et. al., (2013) found that in addition to surgeons' experience, team familiarity based on prior collaborations among the members of a surgical team significantly contributed towards reducing surgical duration and improving surgical outcome.

The effect of input heterogeneity combined with the largely exogenous heterogeneity in patient profiles lead to substantial variation in surgeon, staff, or team performance like hand treble or judgment of tissue stiffness and the resulting incision pressure during surgeries. The variation in surgeon or staff performance manifests in the form of variation in health care outcome measures like surgical procedure duration, patient outcomes like speed of recovery, or other patient-level quality measures such as blood loss during surgery. This can be schematically shown through the conceptual framework in Figure [4.1].

Figure 4.1 Conceptual Framework for Health Care Outcome Variation



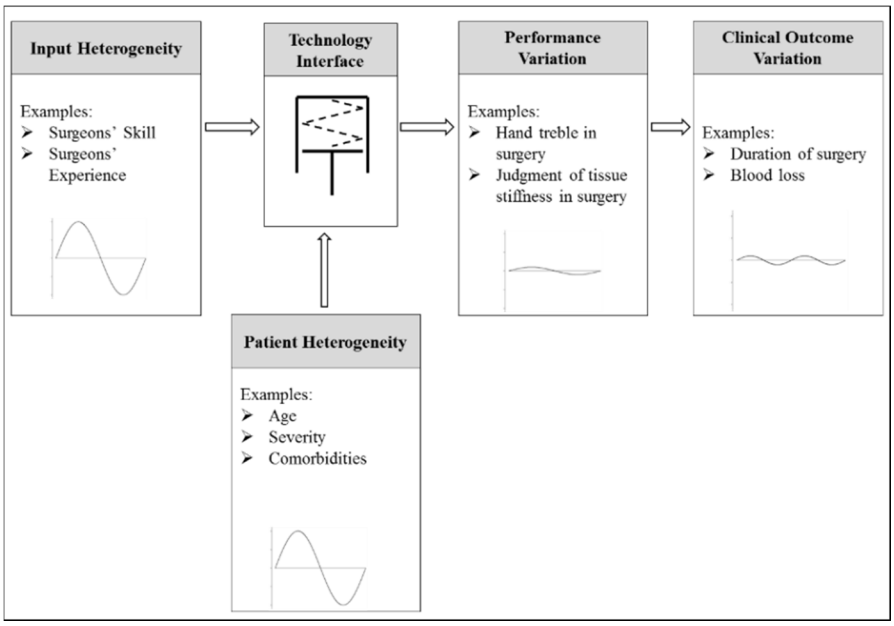
While the input heterogeneity, patient heterogeneity, and clinical outcome variation are measurable, performance is a latent class that is not easily observable or directly measurable except under controlled experimental conditions. Performance is observable through surgical outcome. Outcome variation is a major cause of poor quality and high cost of overall health care delivery in

any economy (Hendryx et al., 2002).

4.2.2 Impact of Technology Mediation

As mentioned earlier, high tech innovations have the potential to play a key role in improving overall performance and reducing outcome variation in health care delivery. This argument can be better understood by acknowledging the fact that health care delivery is a complex adaptive system (Institute of Medicine, 2001; Plasek and Greenhalgh, 2001; Sweeney and Griffiths, 2002). In the context of a health care delivery system, Plasek and Greenhalgh (2001) defined health care as a complex adaptive system, since it is ". . . collection of individual agents with freedom to act in ways that are not always totally predictable, and whose actions are interconnected so that one agent's actions changes the context for other agents. . ." As a complex system, process and outcome variations are natural, where agents act upon localized and internalized sets of rules and heuristics based on their own experiences and contexts in the absence of standardization that can be replicable across agents and across contexts. Variation in outcome also emerges from interactions among agents, like surgeons, staff, and teams. However, technology in health care can help in reducing complexity in health care delivery systems in several ways. First, it can provide a standardized interface between agents in the system to remove many of the interdependencies. Second, technology helps in providing standardized and replicable structural interfaces in health care delivery processes (Plsek, 2003). Third, technology aids in removing skill-based and judgment-based performance variation, either by replacing human skill variation, like hand tremor in surgeries, with standardized interfaces or by making judgment-based decision more objective information-based, as in advanced digital imaging systems. Hence, we hypothesize that high tech innovation in the context of health care delivery mitigates much of the input heterogeneity and helps reduce performance and outcome variation. This is pictorially depicted as the modified conceptual framework in Figure [4.2].

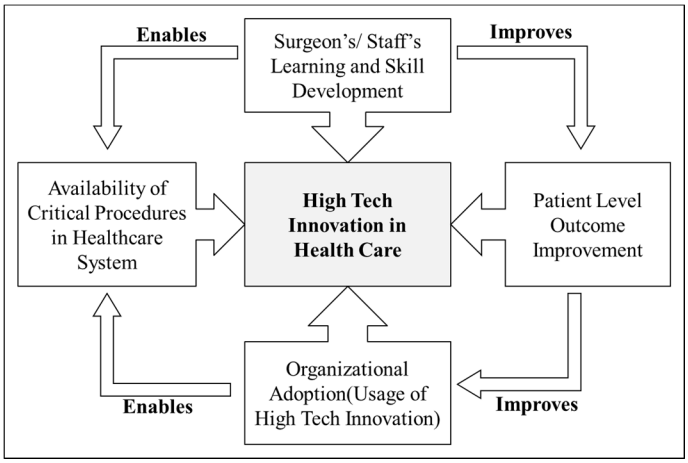
Figure 4.2 Modified Conceptual Framework with Technology Interface as a Dampening Factor for Outcome Variation in Health Care Delivery



4.2.3 Organizational Issues Related to Technology Capability Building

High tech innovation is only one step, albeit a very important one, toward improving performance in the context of health care delivery. Adoption and usage of high tech innovation are among the other key pieces of the puzzle, and without which high tech innovation is likely to be of little significance. In health care delivery systems, adoption and usage of new technology is a complex process. In a conceptual paper, Christensen, Bohmer, and Kenagy (2000) argued that health care organizations benefit the most from disruptive technologies when the key stakeholders of health care delivery, i.e., the surgeons, the clinicians, and the patients, see the benefit in a high tech innovation and embrace the innovation. The authors call for creation of a new organizational culture of learning and feedback to adopt high tech innovations. Hwang and Christensen (2008) also observed that the key to deriving positive results from high tech innovation in health care is proper integration of new technology with organizational learning and change. We integrate these perspectives into a conceptual framework for high tech innovation adoption and usage to enable the conduct of critical and complex health care delivery procedures, as shown in Figure [4.3].

Figure 4.3 Conceptual Model for High Tech Innovation Adoption and Usage



4.3 Empirical Setting

Robot-assisted surgery is the empirical context of this study. To appreciate this empirical context, it is essential to understand the evolution of the modalities of surgical procedures – i.e., from minimally invasive laparoscopic procedures to robot-assisted surgical procedures. Robotic surgical procedure evolved as an alternate to manual laparoscopic surgeries. Some of the constraints that surgeons faced in manual laparoscopic procedures are less degrees of freedom of movement, two-dimensional vision (two-dimensional display), impaired eye-hand coordination (mis-orientation between real and visible movement), and reduced haptic senses (limited tactile feedback mechanism) (Rassweiler, 2006).

The initial motivation for adopting robotic surgical systems was to address many of the constraints present in manual laparoscopic procedures that are major sources of skill and performance variation in normal laparoscopic procedures (Hemal, 2002; 2007). In 1991, the first surgical robot was developed that included a separate surgeon's console and remotely operated telemanipulators (robotic endowrists) was developed in 1991. This first system was called the Stanford Research Institute (SRI) Green Telepresence Surgery System (Ficarra, 2006). The first commercially available robotic surgical system was developed by Intuitive Surgical in 2000 (the first system received FDA approval in July 2000). The first da Vinci surgical robot was an enhanced version of the original SRI surgical robot. The current study focuses on the second generation of the da Vinci robot, which is currently most widely in use in performing robot-assisted surgeries.

Some of the key features of this da Vinci robot include a true three-dimensional immersion

view with up to twelve times magnification, seven degrees of freedom of movement (which is much more than in open surgery as well as normal laparoscopic surgery— normal laparoscopic surgeries provide surgeons only four degrees-of-freedom), alignment of visual and motor axes, auto dampening of surgeon's hand tremors, and tactile pressure sensors of robotic endo-wrists. Much of the performance variation in surgical procedures arises out of fatigue or skill-related hand tremors of surgeons and misjudgment of tactile pressure on tissues while performing open or normal laparoscopic surgeries.

Starting in the latter half of the past decade, robotic surgery has gained substantial popularity and acceptance in several critical application areas: urology, gynecology, gastroenterology, and recently, cardiology. So far approximately 1.5 million surgeries have been done using the da Vinci robotic system in the United States alone. Going forward, it is likely that this high tech innovation will gain higher levels of acceptance in hospitals, and patient communities. Health care delivery organizations, academics, and regulators have stressed the need to conduct further research on the adoption and usage of surgical robots. This study is a step in that direction.

4.4 Literature Review and Hypothesis Development

For developing our study hypotheses, we reviewed and integrated perspectives from two streams of literature. The first stream is related to individual performance and learning in the health care setting. The second stream is related to the clinical studies on robotic surgical procedures.

4.4.1 Robotic Technology and Variation in Surgical Procedure Outcomes

Diwas and Staats (2012) studied performance and learning of cardiothoracic surgeons and found that there is substantial variation in surgeons' performance based on cumulative volume-based experience. They studied a specific cardiothoracic procedure from the introduction of the procedure and found that surgeons learn at different rates, which leads to substantial variance in performance and outcome. They found that specialization and diversity in experience both lead to variation in performance and learning rate. The authors argued that the significant association between specialization, diversity, and volume of experience lead to greater "regionalization" of health care services, resulting in concentration of critical services in few facilities and few regions. Concentration of critical health care services leads to issues related to access and availability of those services to the broader patient population across all strata of the economy and across diverse

geographical areas in the nation (Birkmeyer and Dimick, 2009). Apart from health care settings, in the broader operations literature, task variety and specialization have been found to be positively related to performance variation (Boh et al., 2007; Staats and Gino, 2012). The context of our study is similar to that of these previous studies in many ways. Like Diwas and Staats (2012), we too study the adoption of a new clinical procedure starting from the time of first adoption of the procedure. However, the major point of deviation of our study from the previous studies is that in robotic-assisted surgery, the learning of the surgeon and the surgical team is mediated by high tech innovation that is intended to shorten learning cycles and address many of the sources of individual performance and learning variation.

Diwas and Staat (2012) also found that specialization and task variety is strongly associated with performance variation. In a literature meta-study of major published studies on surgical outcomes, Chowdhury, Dagashand, and Pierro (2007) found that specialization of surgeons is strongly positively associated with improved performance. They also found that the effect of specialization in determining performance variation is more significant for more complex clinical procedures. This has led to progressive specialization in the medical field, which adds significant complexity to workforce planning and availability of trained surgeons within subspecialties (Stitzenberg and Sheldon, 2005). The specialization-based performance variation is related to specific experience and knowledge gained through repeated performance of similar procedures. Hence, both specialization and number of years of individual surgeons' experience are expected to be significant factors associated with performance variation of surgeons under traditional health care delivery practices. However, the da Vinci robotic technology interfaces with many of the sources of variation in individual surgeons, like magnified three-dimensional view of the surgery field, tactile feedback, and auto tremor elimination. Several prior studies found the learning curve on robotic surgical procedures to be significantly shortened as compared to traditional procedures.

Apart from individual sources of variation, prior literature has identified the effect of the surgical team to be associated with procedural heterogeneity and outcome variation. Huckman, Staats, and Upton (2009), and Huckman and Staats (2011) found team composition and team familiarity to significantly influence performance outcome in cardiac surgeries. Also, Diwas, Staats, and Gino (2013) found significant learning variation among cardiac surgeons performing minimally invasive cardiac surgeries based on their own successes and others' failures. This is in a way related to peer effect on performance variation in a surgical setting. However, the nature of the robotic surgical technology removes much of the within-team interactions and interdependencies

within surgical teams. The team size in robotic surgeries goes down marginally as compared to that of traditional surgical procedures. Also, the robotic interface standardizes the interactions of the team with the primary surgeon, to a large extent.

In the medical literature related to robotic surgical procedures, the clinical benefits of the da Vinci procedure are clearly established. Ahlering et al. (2004) studied radical prostatectomy procedures using the da Vinci surgical robot and found significant improvement in clinical outcomes as compared to open as well as normal laparoscopic procedures. The study showed that robotic surgical procedures led to significant reduction in blood loss, postoperative hemoglobin change, and length of hospital stay. Also, the study found that the usage of the da Vinci robot significantly reduced the learning curve of surgeons as compared to normal laparoscopic procedures. Accordingly, we hypothesize that the adoption of the robotic surgical procedure leads to mitigation of much of the individual performance and learning variation in surgical settings that is otherwise expected in normal manual procedures.

HYPOTHESIS 1: Robotic technology decreases the variation in surgical procedure outcomes between surgeons.

4.4.2 Robotic Technology and the Learning of Surgeons and Teams

Individuals learn through cumulative volume of experience in health care delivery (Chowdhury et al., 2007; Edmondson, Winslow, Bohmer, and Pisano, 2003). In the operations literature, both individuals and teams have been found to learn significantly through cumulative volume of experience in any specific task (McCarter et al., 2000; Edmondson, Dillon, and Roloff, 2007). Consequently, we also expect individual surgeons as well as teams to learn with cumulative volume of experience on the surgical robot. However, as mentioned earlier, since the robotic surgery reduces the team size marginally and augments the role of the principal surgeon in the procedure, we expect that the learning curve on the robot will be steeper for the principal surgeon as compared to the overall team learning.

HYPOTHESIS 2a: With robotic technology, a surgeon learns by doing a surgical procedure.

HYPOTHESIS 2b: With robotic technology, a team learns by doing a

surgical procedure.

HYPOTHESIS 2c: With robotic technology, surgeon-learning-by-doing a surgical procedure is greater than team-learning-by-doing the surgical procedure.

Does regularity with which a surgeon performs surgical procedures on the robot matter after accounting for cumulative volume? Regularity of experience is likely to induce a positive reinforcement mechanism in the learning process, while irregularity may result in some forgetting and necessitate relearning to some extent, thereby delaying the overall learning process. We hypothesize that the rate of learning on the robot is positively associated with the regularity with which a surgeon performs surgical procedures on the da Vinci robot. While we investigate the effect of regularity in the context of the robot, we feel that this effect is probably applicable to a much wider set of scenarios, where individuals or groups learn based on cumulative volume of experience. To our knowledge, while the effect of cumulative volume on individual learning is abundant and well developed, the study of individual learning on the frequency and regularity of experience is limited so far.

HYPOTHESIS 3a: The rate of surgical procedure performance improvement for a surgeon is positively associated with the local (time) frequency with which the surgeon performs a surgical procedure.

HYPOTHESIS 3b: The rate of surgical procedure performance improvement for a surgeon is positively associated with the complexity of a surgical procedure.

4.4.3 Learning of Surgeons and Surgical Teams and Usage of Robotic Technology

Organizations improve their efficiency and effectiveness through individual or group learning within the organizations. In the specific case of surgical robots, it is of key significance to analyze issues related to usage of the robot, due to the higher cost of surgical robotic technology. Several studies on surgical robots have found that the overall cost of a procedure goes up in case of robotic

surgeries (Steinberg et al., 2008). Hence, usage is of critical significance in improving the cost effectiveness of surgical robots. Hollingsworth (2008) identified productivity and efficiency improvement in high tech health care services to be factors of considerable impact in health economics. Effective adoption and usage of high tech innovation in health care services have significant impact on cost, availability, and affordability. We expect cumulative learning of both surgeons as well as surgical teams to result in significant improvement in usage of the robot. However, we expect surgeon learning to be more significantly associated with improvement in usage of the robot within the specific hospital. This is because robot-assisted surgery is more surgeon-led than are traditional surgical procedures. The usage of robotic technology mitigates many of the team-based variation in performance.

HYPOTHESIS 4a: The usage of robotic technology for a surgical procedure is positively associated with surgeon learning.

HYPOTHESIS 4b: The usage of robotic technology for a surgical procedure is positively associated with team learning.

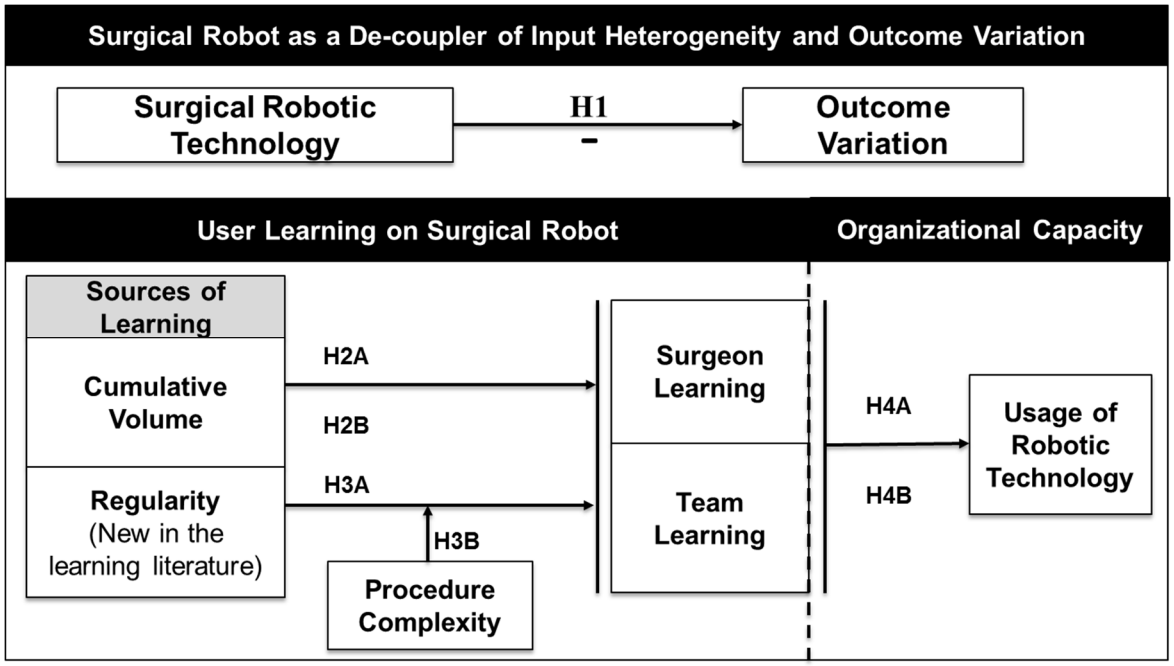
HYPOTHESIS 4c: The impact of surgeon learning on the usage of robotic technology for a surgical procedure is greater than the impact of team learning on the usage of robotic technology for a surgical procedure.

4.4.4 Integrative Framework for Technological Capacity Building within Organizations Delivering Surgical Healthcare

As discussed in previous sections, technology mediation in healthcare delivery can result in several benefits. Technologies in healthcare add functionalities that are not available otherwise. Also, technology mediation improves efficiency and effectiveness of healthcare delivery. Finally, technology mediation in healthcare helps in standardizing interfaces between inputs and outcomes, thus buffering the impact of input heterogeneity on outcome variation. In the context of the robotic surgical procedure, while all the benefits exist, the most significant benefit is that robotic surgical procedures dampen the impact of input heterogeneity (namely, surgeons' and teams' experience heterogeneity, heterogeneity in team familiarity), and patient heterogeneity on surgical outcome variation (namely, surgical duration and quality measured in terms of patient blood loss during

surgeries). Technology mediation, by way of robotic technology interface, between a surgeon and a patient opens up the possibility of using several technological features such as auto-dampening of the surgeon's hand trembling, three dimensional magnified immersion vision leading to better understanding of the surgical field and higher degrees of freedom of movement leading to better accessibility and precision which leads to significant reduction of skill and experience performance variation between surgeons and teams. This allows hospitals to mitigate several organizational constraints such as availability of highly experienced surgeons, scheduling constraints emerging from team formation with high levels of experience and familiarity and constraints emerging from long learning cycle of surgeons and teams otherwise. Thus, technology mediation in surgical care has the potential to lead to improved organizational capacity and improved usage of organizational resources. However, the benefits of technology mediation in healthcare are not realized automatically. We contend that the benefits of technology mediation in healthcare delivery leading to improved capacity and usage evolve over time through surgeons' and teams' learning to use a technology. Learning to use a technology depends on several factors, namely, learning-by-doing through cumulative experience and through regularity of usage. Not only is cumulative experience important in learning with respect to the use of a technology, but how the cumulative experience has been achieved is also critical in determining the level of learning with respect to the use of the technology. Users' learning with respect to a technology leads to improved usage of the technology, which, in turn, leads to organizational capacity building. The set of hypotheses that we have developed embodies the above insights on technology mediation in healthcare delivery and organizational capacity building relevant to performing complex surgical procedures. We integrate the hypotheses into a framework for organizational capacity building for performing robot-assisted (technology mediated) surgical procedures. Figure [4.4] depicts the framework.

Figure 4.4 Integrated Framework for Technology Mediated Surgical Healthcare Delivery



4.5 Data and Methodology

4.5.1 Data

The data for the study were collected from a large multi-specialty hospital in the United States. The hospital had commissioned a second-generation da Vinci robot in early 2008. The robot has primarily been used for several gynecological and urological surgical procedures—specifically, hysterectomy and sacrocolpopexy (gynecological procedures); pelviscopy and prostatectomy (urological procedures). According to the manufacturer of the surgical robot, Intuitive Surgical Inc., the monthly usage of the robot at the multi-specialty hospital that served as the empirical setting for this study has been among the highest in the United States. The study period over which the data were collected is 2008 to 2013, i.e., the entire usage period of the robot since its initial purchase. Our field study spanned a period of two years (2011-13). One member of the research team was stationed in the hospital for six months during this two-year period. This team member worked closely with the surgical and biomedical teams, understanding critical issues pertaining to robot-assisted surgeries from the practical viewpoint of a robot user, and collecting data for the entire

usage cycle of the robot. We also had monthly meetings with the hospital's biomedical department which is responsible for the management and maintenance of the da Vinci robot. These meetings informed the data collection effort and also helped us to understand the data.

4.5.1.1 Response Variables

Beyond clinical outcomes, efficiency is a major consideration for robot-assisted surgery. Hence, the primary response variable we have selected for our study is the duration of a surgical procedure. We collected two measures of surgical procedure time. First is the operating room procedure time, which measures the duration from when a patient is wheeled into the operating room to when the patient is wheeled out (also referred to as "in-to-out" time). The second measure is the surgical procedure time, which is the duration from when the robot arm makes the first incision on a patient to the final stitch (also referred to as "cut-to-close" time). In essence, operating room procedure time includes the surgical procedure time.

4.5.1.2 Explanatory Variables

Factors influencing the surgical procedure time include procedure complexity, the condition of the patient (procedure severity), and surgeon experience. Factors influencing the operating room procedure time include the experience of the staff on a surgical team in addition to the factors influencing the surgical procedure time. Thus, the trend of surgical procedure time provides an estimate of surgeon learning after controlling for all other contributing factors. Similarly, the operating room procedure time provides an estimate of the learning of the surgical team inclusive of the surgeon.

The study data-set includes data on 18 surgeons performing four robotic surgical procedures over a five-year period (2008 to 2013), totaling to a sample size (n) = 1380 surgeries. There is considerable heterogeneity across the four procedures, as is evident from their mean surgical procedure time (in minutes) that are as follows: (i) Prostatectomy, mean = 149.07, std. dev. = 29.17; (ii) Hysterectomy, mean = 141.62, std. dev. = 58.30; (iii) Pelviscopy, mean = 133.47, std. dev. = 53.01; and (iv) Sacrocolpopexy, mean = 207.46, std. dev. = 54.90. (Figure [4.4]). Also Figure [4.5] provides a glimpse of the usage history of the surgical robot in the hospital.

Figure 4.5 Comparison of Robot-Assisted Surgical Procedures

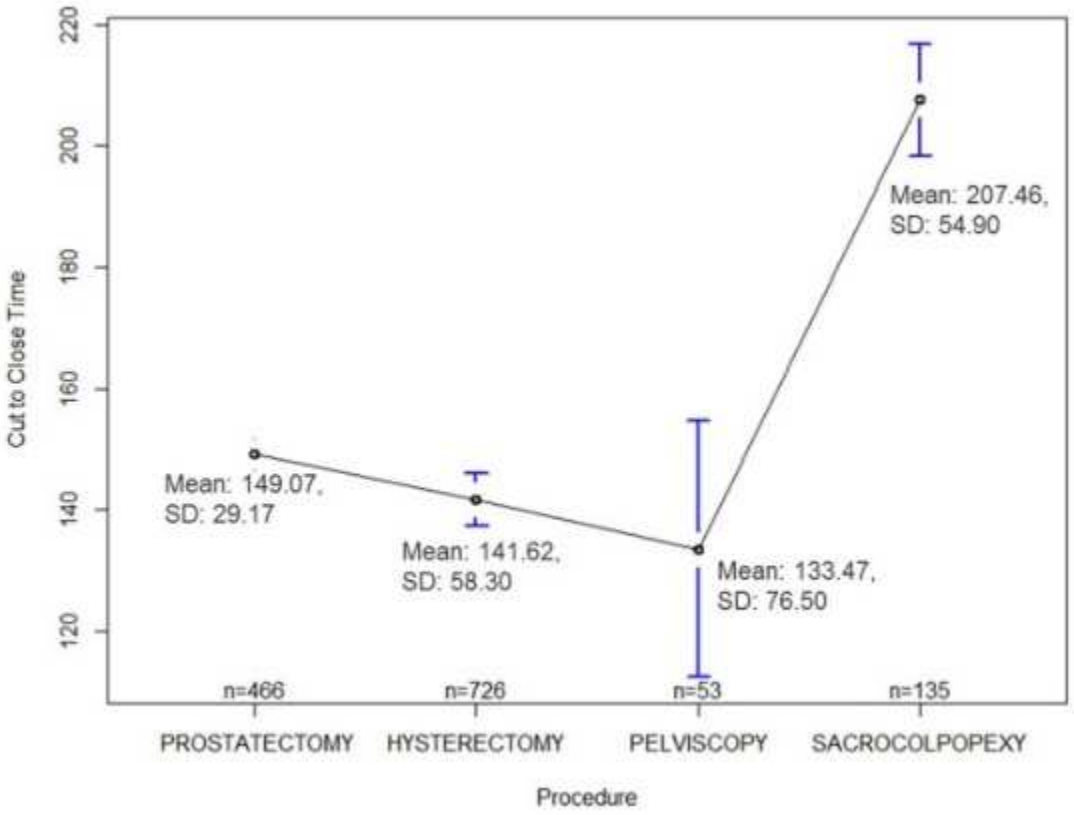
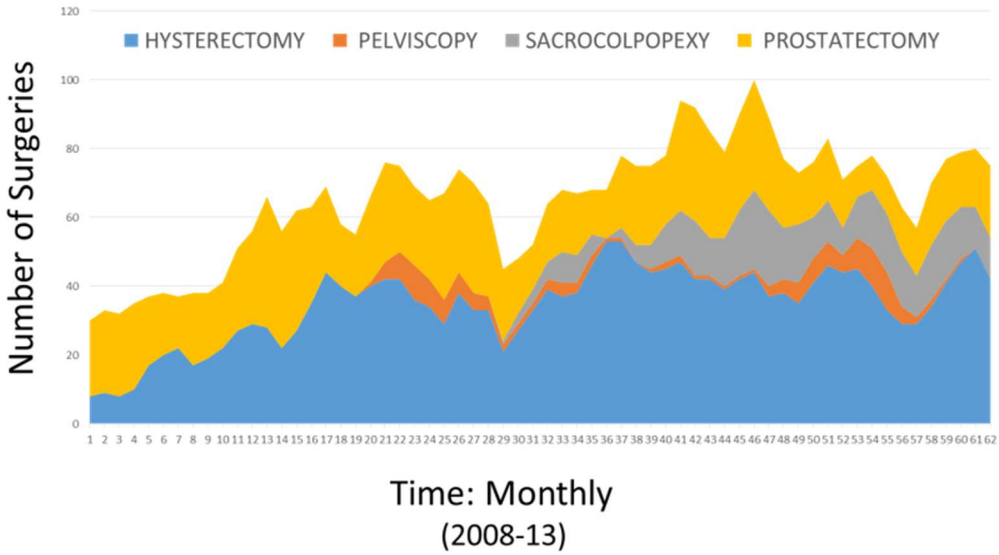


Figure 4.6 History of Usage of the Surgical Robot



The following variables were generated from the data for the purpose of the empirical analysis: (i) cumulative experience, the number of a specific surgical procedure performed by a surgeon; (ii) prior experience, the number of years of experience for a surgeon; (iii) related cumulative experience, the cumulative number of surgeries of other procedure types performed by a surgeon; (iv) diversity, the Herfindahl-Hirschman index *(*HHI, defined as the sum of the squares of the proportion of each procedure type performed during a certain time frame) calculated for a surgeon from the number of different types of surgical procedures performed; (v) team experience, the average experience of the team including the primary surgeon; and (vi) team familiarity, cumulative average number of times any member of a surgical team has worked with any other member of the same team prior to a surgical procedure. For a detailed list of variables and their meaning in the context of the study, refer to Table [4.1].

Table 4.1 List of Variables

Sl	Variable	Description
1	*Surgical Procedure Time (Cut.to.Close)*	Time taken by the surgeon from initial incision to the close of incision, i.e., the main surgery time. This is primarily doctor's time in robotic surgery.
2	*Operating Room Procedure Time (In.to.Out)*	Time for which the patient was in the operations theatre. This is surgery time plus the pre- and post- surgery time. This is the time taken by the team to complete the entire procedure.
3	*Pre- and Post-surgery Time (In.to.Cut +Close.to.Out)*	Pre- and post surgery preparation time. This is purely the team time without the primary surgeon.
4	*Surgeon Cumulative Experience*	Number of prior robotic surgeries done of the same type.
5	*Surgeon Related Experience*	Number of prior robotic surgeries done but not of the same type.
6	*Surgeon Total Experience*	Categorical variable (High, medium and low) representing a doctor's total experience in medical profession. The median number of years of total experience is 11 years in the sample.
7	*Surgeon Diversity*	Herfindahl index of surgical procedures done by a doctor. Skill diversity on robotic surgery.
8	*Team Experience*	Mean experience of the surgical team other than the primary surgeon
9	*Team Familiarity*	The average number of times each member has wrked with every other member of the team. (Following Huckman, Straat and Upton, 2009)
10	*Proc*	Procedure code
11	*Doc*	Surgeon code

4.5.2 Empirical Research Methods

4.5.2.1 Alternate Designs: Choosing the Right Estimation Model

We considered several alternate designs for the study. Since this is panel data over five years and there are two natural units of observations, the surgeons and the procedures, we considered both

fixed effect as well as random effect design for the surgeons and the procedures individually. Specifically, we considered the following alternate designs.

$$(\boldsymbol{Doc\ FE\ and\ Proc\ FE})\ \log\left(procedure.duration_{ijt}\right) = \beta_0 + \beta_1(Cumm.Expr_{it}) + \beta_3(Surgeon.Experience_{it}) + \alpha_{Doc_i} + \gamma_{Proc_j} + \delta\ (Doc_i : Cumm.Expr_{it}) + \epsilon_{ijt} \quad \ldots . [4.1]$$

$$(\boldsymbol{Doc\ RE\ and\ Proc\ FE})\ \log\left(procedure.duration_{ijt}\right) = \beta_0 + \beta_1(Cumm.Expr_{it}) + \beta_2(Surgeon.Diverity_{it}) + \beta_3(Surgeon.Experience_{it}) + \sigma\ U_{Doc_i} + \gamma_{Proc_j} + \sigma\ V_{Doc_i}\ (Doc_i : Cumm.Expr_{it}) + \epsilon_{ijt}; U \sim N(0,1), V \sim N(0,1) \quad \ldots . [4.2]$$

$$(\boldsymbol{Doc\ FE\ and\ Proc\ RE})\ \log\left(procedure.duration_{ijt}\right) = \beta_0 + \beta_1(Cumm.Expr_{it}) + \beta_2(Surgeon.Diverity_{it}) + \beta_3(Surgeon.Experience_{it}) + \alpha_{Doc_i} + \sigma\ W_{Proc_j} + \delta\ (Doc_i : Cumm.Expr_{it}) + \epsilon_{ijt}; W \sim N(0,1) \quad \ldots . [4.3]$$

To choose between the three alternate designs, we performed a series of Hausman tests for random effects of observation units. The Hausman Chi-square test statistic for a test between surgeons' fixed effect versus random effect is 2.729 with a *p*-value of 0.4353 (between equation [4.1] and [4.2]). Since, the Hausman test for surgeons' fixed effect versus random effect is insignificant, a random effect for surgeon learning is admissible, and we choose to use a surgeons' random effect design. If admissible, a random effect design has several key advantages over a fixed effect design. First, random effect model estimation is more efficient than fixed effect estimation because of the reduced number of parameters that need to be estimated in random effect designs as compared to fixed effect designs. Secondly, random effect designs allow estimation of time invariant variables for the observation unit. This is not possible in a fixed effect design, since in a fixed effect design, all time-invariant variables are partialed out (by mean differencing) in the estimation process. So factors like total number of years of experience and other relatively less time-invariant factors related to surgeon or team characteristics can be estimated if random effect design is consistent. In the case of surgeons, the random effect design is both efficient and consistent. In case of procedure, fixed effect versus random effect Hausman test (between equation [4.1] and [4.3]), the value of the Chi-square test statistic is 16.5772 with a *p*-value of 0.0001. Hence, in the case of procedures, a random effect design would be inconsistent and hence not admissible.

Therefore, we chose to go with a fixed effect design for procedure type. The results of the tests between various alternate designs are shown in Table [4.2].

Table 4.2 Model Estimation Results for Tests of Model Choice

Model Choices	Test	Statistic	P-Value	Conclusion
Pooled OLS vs. Surgeon FE	F Test	3.2889	0.0016**	Doctor's fixed effect design is preferred over pooled OLS.
Surgeon FE vs. Surgeon RE	Hausman Test (Chi sq)	2.729	0.4353	Doctor's random effect design is admissible and more efficient.
Procedure FE vs. Procedure RE	Hausman Test	16.5772	0.0001***	Proc fixed effect is preferred.
Surgeon RE	Serial Correlation Chi-sq Test (Breush-Godfrey-Wooldridge)	5.667 (Chi df: 3)	0.8314	No serial correlation for the doctor random effect model. So differencing would not be required.

4.6 Model Estimation, Results and Discussion

4.6.1 Testing Hypotheses 1 and 2 (H1 and H2a, 2b, 2c)

We investigated the variation in the performance outcomes of surgical procedures (H1) and surgeon and surgical team learning (H2) associated with robot-assisted surgeries by estimating a generalized linear mixed model (GLMM) with surgical procedure type fixed effects and surgeon random effects. As shown in Table [4.3], the model estimation was done with respect to two different response variables, i.e., natural logarithm of surgical procedure time (*cut.to.close*) and natural logarithm of operating room time (*in.to.out*). The estimation model for *cut.to.close* captures a surgeon's learning and performance in conducting a surgical procedure with the assistance of a robot. The estimation model for *in.to.out* time captures surgical team learning and performance. Each of the models were estimated twice: first, with data from the complete sample (n=1380) that we refer to as M1; and second, with data from the sub-sample (n=830) that we refer to as M2. The subsample excludes data for the surgical procedure prostatectomy but contains data related to team composition. Note, the explanatory variables related to team composition (*team.experience* and *team.familiarity*) were included in model estimation with data from the smaller sub-sample (M2) since data related to team composition were not available for prostatectomy. Other than explanatory variables related to team composition, the explanatory variables in the two models estimated are the same. Also, for completeness we have estimated a small procedure fixed effect generalized linear model (GLM) for pre- and post- surgery time, which is the time when the surgical team excluding the surgeon is performing a procedure. The results of the estimation are in Table [4.3].

Table 4.3 Generalized Linear Mixed Model (GLMM) Estimation Results for Surgeon and Surgical Team Learning in Conducting Robot-Assisted Surgical Procedures

	Surgical Procedure Time (M1)	Surgical Procedure Time (M2)	Operating Room Procedure Time (M1)	Operating Room Procedure Time (M1)	Pre and Post Surgery Time (M2)
Random Effects Variance (p-Value)					
Surgeon Intercept	0.0093 (0.9997)	0.0103 (0.9994)	0.0107 (0.9997)	0.0093 (0.9996)	-
Surgeon Slope (Cum.Expr)	3.455 e-09 (0.9999)	2.56 e-08 (0.9999)	2.317 e-07 (0.9999)	2.217 e-07 (0.9999)	-
Surgeon: Procedure Intercept	0.0156 (0.9996)	0.0173 (0.9993)	0.0276 (0.9995)	0.0301 (0.9994)	-
Surgeon: Procedure Slope (Cum.Expr)	3.326 e-07 (0.9999)	2.734 e-07 (0.9999)	1.519 e-06 (0.9999)	2.014 e-06 (0.9999)	-
Fixed Effect Estimates (p-Value)					
Intercept	**5.124 (0.0000)*** **	**5.083 (0.0000)*** **	**5.354 (0.0000)*** **	**5.335 (0.0000)*** **	**3.7227 (0.0000)*** **
Cumulative Experience	**-0.0051 (3.714 e-09)*** **	**-0.0071 (1.47 e-05)*** **	**-0.0038 (6.453 e-06)*** **	**-0.0052 (6.09 e-06)*** **	-
Squared Cumulative Experience	**9.910 e-06 (0.0002)*** **	**1.952 (0.003)** **	**9.439 e-06 (1.631 e-05)*** **	**1.573 e-05 (0.0015)** **	-
Surgeon Diversity	-0.1810 (0.7785)	-0.0923 (0.6345)	-0.1373 (0.7856)	-0.1242 (0.6753)	-
Prior Related Experience	**-0.2619 (0.0051)*** **	**-0.1437 (0.0043)** **	**-0.1421 (0.0105)** **	**-0.1326 (0.02)***	-
Total Surgeon Experience (Low)	0.1617 (0.1338)	0.1832 (0.1009)	**0.1876 (0.0245)***	0.1797 (0.1183)	-
Total Surgeon Experience (Medium)	0.1350 (0.2207)	0.1273 (0.1632)	0.1645 (0.0610)+	0.1745 (0.073)+	-
Team Experience		0.0011 (0.2194)		0.0008 (0.196)	**-0.0011 (0.010)***
Team Familiarity Index		-0.0078 (0.2464)		-0.0044 (0.3507)	-0.0021 (0.263)
Procedure Code	S	S	S	S	NS
Number of Procedures	4	3	4	3	3
n	1380	830	1380	830	830

Notes:

*1. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Test	Chi-sq value (p-value)
Surgeon random effect before accounting for procedure fixed effect	29.744 (1.56e-06***)
Durgeon random effect after accounting for procedure fixed effect (H_0:δ^2=0)	3.8392 (0.2794)

The results in Table [4.3] corresponding to the response variable *cut.to.close* indicate that H1 is supported. That is, the variation in surgical procedure time between surgeons is not statistically significant in robot-assisted surgeries. The random intercepts and random slopes for surgeons' performance on the robot are not statistically significant. One of the key reasons why hospitals adopt robotic surgical technology and provide a high tech interface between a surgeon and a patient in an otherwise manual skill-driven procedure is to mitigate skill-based performance variation across individual surgeons, and, thereby, address the downside risks related to the heterogeneity in the skills and experiences of surgeons.

Also, as shown in Table [4.3], there is empirical support for H2a and H2b. We find that surgeons (response variable *cut.to.close* in M1 and M2*)* as well as surgical teams (response variable *in.to.out* in M1 and M2; and response variable *pre-and-post-surgery-time* in M2) learn by doing robot-assisted surgical procedures. Cumulative number of surgeries performed with the robot is associated with reduction in the duration of surgical procedures. However, the learning curve of surgeons has a steeper slope than that of surgical teams (slope of cumulative experience for

cut.to.close in M1: -0.0038, p<0.001; and slope of cumulative experience for *in.to.out* in M1: -0.0051, p<0.001). This is also evident from the respective slopes of surgeon experience and team experience from model estimations corresponding to the sub-sample M2 for *in.to.out* (surgeon experience: -0.0071, p<0.001; team experience: 0.0011, p>0.2) and *cut.to.close* (surgeon experience: -0.0052, p<0.001; team experience: 0.0008, p>0.2). Team experience becomes significant in the model estimated for the response variable *pre-and-post-surgery-time* (team experience: -0.0011, p=0.01). The significance level for surgeon experience is higher than the significance level for team experience. This results lends support to H2c. The explanation for the result supporting H2c is that a robot-assisted surgical procedure is largely a surgeon-led procedure, and the role of other members on the surgical team diminishes considerably during the actual surgical procedure. Many of the activities of the team, like change of cutter and tips; stitching; and handling of gauges, catheters, and other disposables are done through the robotic interface. Also, team familiarity is not a significant predictor of performance on the robot, as is evidenced from the results of the models estimated and presented in Table [4.3]. In none of the model estimations, team familiarity is significant. Even in the model estimated for the response variable *pre-and-post-surgery-time*, which is primarily the surgical team's operation time excluding the surgeon, team familiarity is not significant. This is a new finding that is contrary to the findings in the existing published literature (cf. Xu et. al, 2013).

4.6.2 Testing Hypothesis 3 (H3a and H3b)

In this sub-section, we discuss the estimation of models for testing H3a and H3b related to the nature of surgeon learning in the context of robot-assisted surgeries. Typically, linear models aggregate over the entire support space (domain) in calculating the slope parameters. In doing so, local dynamics and temporal evolution of the data related to the underlying phenomenon is lost. While cumulative volume has, typically, been used in the extant literature as a surrogate for learning-by-doing, some studies suggest that learning can depreciate over time (Argote et al., 1990), and, hence, recency or regularity of learning-by-doing can be of consequence. For estimating how regularity (defined as the local frequency on temporal dimension) affects the learning rate of surgeons, we use the following model formulation. If $D = \log(procedure\ duration)$ represents the response variable for individual surgeons, we can model this using a general function $D = f(X_t)$ where X_t is the cumulative volume of procedures done at time t. By a Taylor series expansion, we can write,

$$D = f(X_t) + (X - X_t)^T \frac{\partial f(X)}{\partial X} + \frac{1}{2}(X - X_t)^T \frac{\partial^2 f(X)}{\partial X^2}(X - X_t) + o_p(\cdot) \quad \ldots.[4.4]$$

We estimate the function using a kernel $(K_h, h: Bandwidth)$ weighted loss function,

$$\min \sum_{i=1}^{n} \left\{ D_i - f(X_t) - (X - X_t)^T \frac{\partial f(X)}{\partial X} - \frac{1}{2}(X - X_t)^T \frac{\partial^2 f(X)}{\partial X^2}(X - X_t) \right\}^2 K_h(X_i - X_t) \quad \ldots.[4.5]$$

The third term $\frac{\partial^2 f(X)}{\partial X^2}$ of the local function estimation gives us the rate of learning. The local frequency of estimation or regularity is estimated using a kernel density function for individual surgeons in the sample:

$$R_t(X) = \frac{1}{nh} \sum_{i=1}^{k} K\left(\frac{X_i - X}{h}\right) \quad \ldots.[4.6]$$

Finally, the local learning rate is estimated as a function of local frequency of procedures done using a simple generalized linear model,

$$\left[\frac{\partial^2 f(X)}{\partial X^2}\right]_i = \gamma_0 + \gamma_1 R_t(X_i) + \gamma_2 X_i + \sum_{j=1}^{3} \alpha_{j,Proc} + \eta_i \quad \ldots.[4.7]$$

The results of the model estimation are shown in Table [4.4]. We estimated the model using the complete sample (n=1380). Also, we estimated models pertaining to individual years from 2008 to 2013 to see if the effect of regularity changes over time as surgeons become more skilled using the robot.

Table 4.4 Model Estimation Results for the Local Frequency Effect of Robot-Assisted Surgeries on Surgeon Learning

	Complete	Year 1	Year 2	Year 3	Year 4	Year 5
(Intercept)	0.0756 (0.04)*	0.2278 (0.0023)**	0.3027 (0.0005)***	0.072 (0.01)**	0.0807 (0.051)*	0.019 (0.062)+
Cumulative Volume	-0.0003 (0.1301)	**-0.0129** **(0.0021)****	**-0.0059** **(0.0027)****	**-0.0014** **(0.05)***	-0.0012 (0.1413)	-0.0005 (0.463)
Immediate Frequency	**-0.0155** **(0.01)****	**-0.0473** **(0.0010)*****	**-0.0599** **(0.0005)*****	**-0.0153** **(0.043)***	-0.0211 (0.173)	-0.0073 (0.251)
Cumulative Volume * Immediate Frequency	0.0002 (0.2121)	0.0002 (0.091)	0.0001 (0.02)*	-0.0003 (0.216)	0.0002 (0.192)	0.0014 (0.465)
Proc.Hysterectomy	0.0038 (0.2617)	0.0035 (0.08)+	0.0081 (0.201)	0.0072 (0.181)	0.0002 (0.2079)	0.0002 (0.258)
Proc.Sacrocolpopexy	**-0.0034** **(0.0611)+**	**-0.0266** **(0.001)*****	**-0.012** **(0.031)***	**-0.0051** **(0.091)+**	-	-

Notes:

1. *Significance code 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '+' 1*
2. *For fair comparison year-wise data has been calculated based on individual doctor's usage timeline from*
3. *Base procedure is taken as Prostatectomy for procedure control. Pelviscopy numbers for individual doctors*
4. *Sacrocolpopexy as a procedure was started after almost two years of the robotic surgical procedure was*
5. *Data for year 5 is for 11 months and not complete year, due to the sample characteristics.*

From the results presented in Table [4.4], we infer that the learning process is more dependent on the frequency with which a surgeon performs a surgery than just the cumulative volume. The local slope of surgical time is significantly dependent on the rate at which surgeons performs a surgical procedure (slope of regularity in the full model: -0.0155, $p<0.01$; slope of cumulative experience: -0.0003, $p>0.1$). This is also observable from the year-wise subsample model estimates. Specifically, for 2008 (slope of regularity: -0.0473, $p<0.001$), 2009 (slope of regularity: -0.0599, $p<0.0005$) and 2010 (slope of regularity: -0.0153, $p<0.043$). However, for subsequent years neither regularity nor cumulative volume lead to further learning. Hence, the effect of regularity appears to be significant in the initial stages of usage of the robot. This result confirms H3a according to which learning is a self-reinforcing mechanism where repetitions at high frequency lead to a positive feedback cycle, and on the other hand, low rate of repetitions may lead to some attrition or forgetting in the learning process.

In the hospital that serves as our field setting, the above finding has two practical implications. First, the hospital is currently considerably constrained in using the da Vinci robot for gynecological and urological procedures. The hospital has also tried out some of the gastroenterology and general surgical procedures on the robot. However, since the robot is always scheduled to almost full capacity with gynecological procedures (hysterectomy and sacrocolpopexy) and urological procedures (pelviscopy and prostatectomy), the regularity with which the other

disciplines can get surgeries scheduled on the robot is severely constrained. Hence, in spite of performing quite a few gastroenterological surgeries with the robot, the hospital is not able to develop a reliable and sustainable capability for performing robot-assisted surgeries in other disciplines. Second, from a point of view of an individual surgeon, we find that though a number of surgeons in the hospital have undergone training and certification for performing robot-assisted surgical procedures, only a few are performing the surgical procedures on a regular basis. Several of the surgeons, even in gynecology and urology disciplines, who have performed a number of robot-assisted surgeries but not on a regular basis, have eventually dropped out of the list of surgeons capable of performing surgeries with the da Vinci robot. Understanding the nuances of the learning process of a surgeon is critical to improving the usage of the robot.

Out of all the four different surgical procedures that act as our context, sacrocolpopexy is clinically the most complex procedure, as is also evidenced by the mean duration and variation of the *cut.to.close* time depicted in Figure [4.4]. We observe from the results in Table [4.6] that the learning rate for sacrocolpopexy is significantly higher than for the other procedures (coefficient estimate for the fixed effect of sacrocolpopexy in the full model: -0.0034, $p<0.06$; in sub-model for 2008: -0.0266, $p<0.001$; in sub-model for 2009:-0.012, $p<0.013$; in sub-model for 2010: -0.0051, $p<0.091$). This leads us to conclude that surgeon learning is relatively higher in more complex procedures. Thus, we find support for H3b that learning on more complex procedures is higher than relatively less complex procedures after accounting for cumulative volume and regularity of experience.

4.6.3 Testing Hypothesis 4 (H4a and H4b)

For investigating the effect of surgeon and surgical team learning on the usage of the robot (H4a and H4b), we estimated a generalized additive model (GAM). GAM is appropriate because we expect that the effect of surgeon and surgical team learning on the usage of the robot would be nonlinear in nature, as can be seen in Figure [4.6]. GAM uses an additive spline regression to estimate the effect of the predictor variables on the response. It is a distribution-free robust estimation model where the significance of the predictors can be estimated through a robust nonparametric estimation method. The results of the GAM estimation are presented in Table [4.5].

Figure 4.7 Monthly Usage of the da Vinci Robot at the Multi-Specialty Hospital

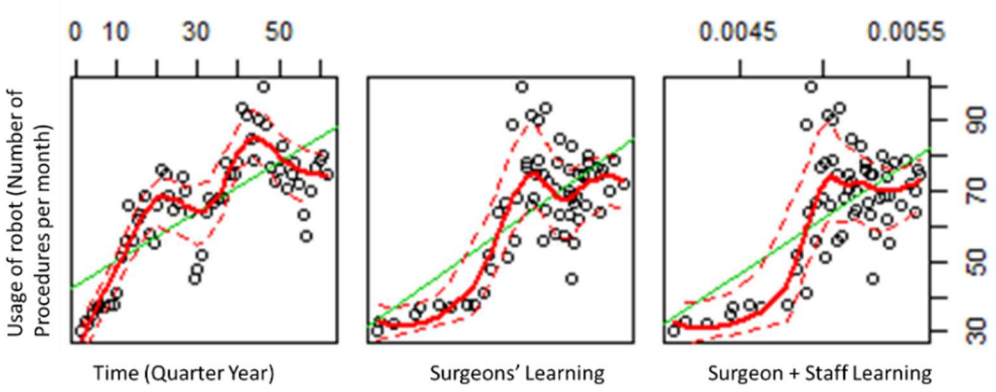


Table 4.5 Generalized Additive Model (GAM) Estimation Results for the Monthly Usage of the Surgical Robot

	M1	M2	M3	M4
Time	7.257***	6.379***	7.029***	6.394***
Surgeon Learning		6.417**		6.431*
Team Learning			1.932*	1
Surgical Procedure Mix	NS	NS	NS	S
Deviance Explained		92.20%	89.10%	92.40%
n	62	62	62	62

The results of GAM estimation in Table [4.5] provides support for H4a and H4b. We estimated four separate models to investigate the effect of surgeon as well as team learning separately and jointly. As expected, we find that both surgeon learning and surgical team learning are associated with improvement in the usage of the surgical robot. This relationship is evidenced at a more aggregate level in that the hospital is now able to schedule three surgeries per day on the surgical robot compared to two surgeries two years ago, in 2010-11. The impact of surgeon learning is more significant than surgical team learning on improvement in the usage of the robot (H4c). Again, this is because the robot-assisted surgery is largely a surgeon-led procedure.

4.7 Robustness Checks for the Model Estimation Results

To establish robustness of the results we performed several checks. Specifically, we performed robustness checks with respect to model specification for data heteroscedasticity and surgical procedure outcomes, specifically, procedure duration, clinical quality outcome and length of stay. Finally, we accounted for the endogenous relationship between clinical quality outcome and

surgical procedure duration.

4.7.1 Robustness Check Related to Data Heteroscedasticity

To check for robustness of the GLMM estimation results, we need to account for the likely heteroskedastic nature of the data. Hence, we performed a Lagrange Multiplier test on the data use for GLMM estimation. The test confirmed, (p<10e-06), that there is substantial heteroscedasticity in the response variable. Since the variance structure of the response variable is influenced by the underlying phenomenon of learning by doing, we attempted to understand the dependence structure of the variance on cumulative experience and not just control for it. We adopted a Bayesian formulation of the GLMM to account for simultaneous influence of learning by doing on both the mean and variance of the surgical procedure duration. Like the GLMM, we assumed that the natural logarithm of the surgical procedure duration has a Gaussian distribution with model parameters $\beta \; and \; \sigma^2$. If y denotes the response, natural logarithm of the surgical procedure duration, and X denotes the predictor variables as in the GLMM, we can write the structure of the problem as follows:

$$p(y|X,\beta,\sigma_X^2) \sim N(y|X'\beta,\sigma_X^2 I_n) \qquad \ldots . [4.8]$$

$$(Random\ Effect)\ p(\beta|\sigma_X^2) \sim N(\beta|\mu_\beta,\sigma_X^2 I_n) \qquad \ldots . [4.9]$$

$$p(\sigma_X^2|X,a,b) \sim IG(a,bX) \sim \frac{(bX)^a}{\Gamma(a)}\left(\frac{1}{\sigma_X^2}\right)^{a+1} \exp\left(-\frac{bX}{\sigma_X^2}\right), \sigma_X^2 > 0 \qquad \ldots . [4.10]$$

where, a, b, μ_β are the hyper-parameters of the Bayesian structure and IG is an Inverse Gamma distribution, which is a natural conjugate prior for the variance parameter of Gaussian distribution of the data. The details of the derivation of the posterior estimation model is shown in Appendix [C.1].

We estimate the Bayesian model parameters using a Markov Chain Monte Carlo (MCMC) approach in R-BUGS (Bayesian inference Using Gibbs Sampling, a package in R). This formulation is also useful in simulating scenarios based on the fitted model for checking robustness and generalizations of the GLMM model. The results of model estimation are presented in Table [4.6]. The Bayesian model parameters were estimated on the full data sample (n=1380). As described in the formulation of the model, we estimated the models for both the mean and variance of natural logarithm of *cut to close*.

Table 4.6 Robustness Check Related to Data Heteroscedasticity – Estimation Results for Bayesian Normal Inverse Gamma (NIG) Model

	Response : Log(Cut.to.Close) *Posterior Marginal: Multivariate Students*				*Response: Variance (Log_Duration)* *Posterior Marginal: Inverse Gamma*			
Random Effects (Variance)								
	Posterior mean	**l-95% CI**	**u-95% CI**	**pMCMC**	**Posterior mean**	**l-95% CI**	**u-95% CI**	**pMCMC**
Surgeon	0.02737	0.002649	0.0724	0.5839	0.05828	0.006132	0.1555	0.2689
Surgeon:Experience	0.001759	0.0001779	0.004242	0.8875	0.0009802	0.0001188	0.002774	0.8859
Residual (Units)	0.08638	0.07972	0.09358	-	0.04765	0.009563	0.02396	-
Fixed Effects								
	Posterior mean	**l-95% CI**	**u-95% CI**	**pMCMC**	**Posterior mean**	**l-95% CI**	**u-95% CI**	**pMCMC**
Intercept	5.63E+00	3.13E+00	8.38E+00	<0.001***	8.8303669	6.1654597	10.9693	<0.001***
Cumulative Experience	-3.69E-03	-4.78E-03	-2.66E-03	<0.001***	-0.9795635	-1.3955867	-0.5018113	<0.001***
Squared Cumulative Experience	9.58E-06	5.04E-06	1.40E-05	<0.001***	-	-	-	-
ProcHYSTERECTOMY	-1.10E-01	-5.98E-01	3.68E-01	0.564	-1.1790245	-1.6226238	-0.791816	0.002**
ProcPELVISCOPY	-2.09E-01	-6.56E-01	3.12E-01	0.324	-0.2770877	-0.9227214	0.3195546	0.296
ProcSACROCOLPOPEXY	1.02E-02	-4.67E-01	5.47E-01	0.978	1.8352	0.58253	2.19263	0.001***
Doctor Diversity	-6.36E-01	-3.35E+00	1.83E+00	0.558	-	-	-	-
Prior Related Experience	-2.63E-03	-6.57E-02	5.46E-02	0.94	-	-	-	-
Total Surgeon Experience (Low)	2.53E-01	-5.70E-02	5.70E-01	0.102	-	-	-	-
Total Surgeon Experience (Medium)	2.44E-01	-4.28E-02	5.14E-01	0.098+	-	-	-	-

Notes:
*1. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
2. Iterations = 3001:12991
3. Thinning interval = 10
4. Sample size = 1000

From the model estimation results in Table [4.6], we see that the performance variance between surgeons (variance estimate for surgeon experience: 0.001759, p=0.8875) is not statistically significant indicating that the variance between surgeons is not significant. This lends support to H1. In addition, we observe that the variance of surgical procedure duration for surgeons has reduced significantly (slope of cumulative experience: -0.9795, p<0.001) with experience. This lends additional support to H1. We also observe that surgeons learn significantly in conducting robot-assisted surgeries with cumulative volume of experience (slope of cumulative volume: -0.0037, p<0.001). This lends additional support to H2. Also, we observe that a surgeon's performance with the robot is not significantly dependent on the surgeon's prior experience in the

medical profession or the surgeon's degree of specialization. Similarly, we checked for the robustness of surgical team learning with respect to heteroskedastic error variance through a similar GLMM estimation. The results are not reported here because of limitations of space.

4.7.2 Effect of Robotic Technology Adoption on Clinical Quality Outcome and Length of Stay

From the standpoint of surgical practice, robot-assisted surgical procedure is considered to be an advancement over manual laparoscopic surgical procedure. Hence, it is important to check if robot-assisted surgery is superior to manual laparoscopy in terms of clinical quality outcomes. To that end, we were able obtain a small sample of data (n=200) on the gynecological procedure of hysterectomy that included both robot-assisted surgeries and manual laparoscopic surgeries. The data-set for this sample included data on patient characteristics (*BMI [Body Mass Index], uterine weight* and *comorbidities*) and data on quality outcomes for a subset of the small sample, namely, *blood loss* in ml and *length of stay* in the hospital in hours. We conducted a subset analysis to compare the outcomes related to robot-assisted surgeries and manual laparoscopic surgeries. The response variables for quality outcomes are *blood loss* in ml and *length of stay* in the hospital in hours. Simultaneously, we also estimated models with *cut to close* duration as response variable for both the robot-assisted surgery and manual laparoscopy. The model estimation results are presented in Table [4.7].

Table 4.7 Generalized Linear Model (GLM) Estimation Results for the Impact of Robotic Assisted Surgeries on Clinical Quality Outcome and Surgical Procedure Duration

Response:	Blood_Loss_ML		Cut_to_Close		Length_of_Stay	
	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
(Intercept)	-119.1578	0.0008***	50.9952	0.0332*	21.3277	0.0217*
BMI	5.2440	4.80e-05***	1.2899	0.1242	-0.0499	0.8767
Uterine_Weight	0.1534	0.0002***	0.0664	0.0183*	0.0015	0.8855
Comorbiditiesy	-25.7551	0.0470*	3.0476	0.7273	1.9465	0.5654
Procedure_Type_Manual.Laparoscopy	281.2054	1.58e-06***	123.0918	0.0013**	19.6343	0.1723
Uterine_Weight:Procedure_Type_Manual .Laparoscopy	0.3393	0.0156*	0.3830	0.0205*	0.0628	0.3184
BMI:Procedure_Type_Manual.Laparosco py	6.9538	0.0002***	1.3571	0.2687	0.189499	0.6889
n	200		200		200	

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

The results in Table [4.7] show that the adoption of the surgical robot technology has led to significant improvement in patient level outcome, i.e., blood loss during surgeries (fixed effect

coefficient estimate for manual laparoscopy: 281.21, p<0.0001). Mean blood loss for manual laparoscopic surgery is significantly higher than mean blood loss for robot-assisted surgery. We also found that higher the criticality of a patient's condition, higher is the clinical benefit of the robotic technology mediation in surgical procedures. This can be observed from the interaction of procedure type and patient characteristics. The slope of the interaction of *uterine weight* and *manual laparoscopy* procedure is 0.3393 (p<0.0156) and the slope of the interaction of *Body Mass Index (BMI)* and *manual laparoscopy* procedure is 6.9538 (p<0.0002). Similar effect is observed with the response variable, surgical procedure duration *cut.to.close* (interaction estimates of *uterine weight* and *manual laparoscopy:* 0.3830, p<0.0205). That is, patients with higher uterine weight for hysterectomies, an indicator of higher severity of patient condition, benefit more from robot-assisted surgery. However, we did not find such effect of robot-assisted surgery on the *length of stay.* This may be because a patient's length of stay in a hospital is also dependent on several other exogenous variables which are beyond the scope a surgeon's robot-assisted surgical procedure performance. Therefore, all in all, we can conclude that the results pertaining to the testing of the study hypothesis presented in section 6 are robust after accounting for patient condition and clinical quality outcome.

4.7.3 Accounting for the Endogenous Relationship between Clinical Quality Outcome and Surgical Procedure Duration

Above, we checked for the robustness of robot-assisted surgical procedure performance variation across surgeons with the surgical procedure outcome being a clinical quality measure: *blood loss* in ml. However, the clinical quality outcome and surgical procedure duration are likely to be *endogenous*. Hence, we estimated the effect of surgeon experience on clinical quality outcome measure as *blood loss* and surgical procedure duration measured as natural logarithm of *cut to close* time using a system of simultaneous equations. The estimation was done using the subset of data on hysterectomies (n=200), as described in the previous sub-section. Since this subsample has information on a smaller number of surgeons (6 surgeons), we used a surgeon level fixed effect for model estimation. Table [4.8] presents model estimation results corresponding to the simultaneous estimates of *blood loss* and *cut to close* duration. Details of the simultaneous equation model for the estimation are provided in Appendix [C.2].

Table 4.8 Simultaneous Equations Estimation Results for the Impact of Robotic Assisted Surgeries on Clinical Quality Outcome and Surgical Procedure Duration

Response: log(Cut_to_Close)				
	Estimate	**Std. Error**	**t-Statistic**	**p-Value**
Intercept	4.0473	1.0899	3.7132	0.0004***
log(Blood_Loss_ML)	0.2190	0.0114	19.0565	0.87e-10***
Doc_GYNMD03	0.1469	1.3291	0.1105	0.9122
Doc_GYNMD05	-0.0477	0.5461	-0.0875	0.9306
Doc_GYNMD06	-0.0461	0.8912	-0.0725	0.9424
Doc_GYNMD10	0.1345	2.3609	0.0567	0.9547
Doc_GYNMD11	-0.0035	0.0175	-0.2009	0.8412
Cumm_Expr	-0.1191	0.0815	1.4617	0.1475
Comorbidities_Y	0.0947	0.0551	1.7184	0.0893+
log(Uterine_Weight)	0.0315	0.0151	2.0959	0.0391*
Response: log(Blood_Loss_ML)				
	Estimate	**Std. Error**	**t-Statistic**	**p-Value**
Intercept	-13.6491	5.3302	-2.5607	0.0122*
log(Cut_to_Close)	3.4049	1.2091	2.8161	0.0061**
Doc_GYNMD03	-0.2699	0.7792	-0.3464	0.7299
Doc_GYNMD05	0.0917	0.2669	0.3436	0.7215
Doc_GYNMD06	0.2669	0.3944	0.6767	0.5005
Doc_GYNMD10	0.7468	0.5851	1.2764	0.2054
Doc_GYNMD11	-0.0124	0.0209	-0.5933	0.5528
Cumm_Expr	-0.5438	0.2683	-2.0268	0.0458*
Comorbidities_Y	0.4123	0.1396	2.9534	0.0041**
BMI	0.0031	0.0015	2.0667	0.0489*
n	153			
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*				

The results in Table [4.8] suggest that surgeon level fixed effects are not significant for both *cut to close* duration as well as *blood loss*. These results confirm the robustness of the relationship posited in H1. Robotic technology mitigates individual surgeon level skill and experience heterogeneity in determining the variation in surgical procedure outcomes. We find that longer surgical procedure duration is associated with higher blood loss. Hence, surgeon learning and reduction in *cut to close* duration are likely to be positively associated with the clinical quality measure (i.e., *blood loss* – less blood loss is a better quality outcome). Also, we find that surgeon experience is not associated with the duration of a surgical procedure, however, surgeon experience is positively associated with clinical quality outcome (i.e., reduction in *blood loss*). An explanation for the results is that the sample data was from the year 2012 when all the surgeons in the sample had substantial experience performing robot-assisted hysterectomy. Also, we find that measures of patient condition (*Uterine_Weight, Comorbidities and BMI)* are significant predictors of both

clinical quality outcome (*blood loss*) and surgical procedure duration (*cut to close*).

We also performed robustness checks with respect to varying patient population through a simulation estimate. The details of the simulation study are provided in Appendix [C.3].

4.8 Conclusion

The fundamental question that motivated this study was: Does the adoption of a high tech innovation like a surgical robot mitigate the effect of experience and skill heterogeneity among surgeons and surgical teams on surgical procedure outcome variation? While investigating this question, we also examined other consequential issues like the nature and mechanisms of learning that are related to the adoption of a high tech innovation in health care delivery settings. Specifically, we conducted a longitudinal field study at a large multi-specialty hospital that had adopted robotic surgical technology (da Vinci robot). One member of the research team was stationed in the premises of the hospital for a period of six months, working closely with the surgical and biomedical teams, understanding critical issues pertaining to robot-assisted surgeries from a robot user's perspective, and collecting data for the entire usage cycle of the robot.

The results of this study provided strong support for the claim that the robotic technology mediation reduces the effect of many sources of input heterogeneity on outcome variation in robot-assisted surgical procedures. A key insight obtained from the study is that a surgeon's specialization or a team's familiarity level did not affect the variation in surgical procedure duration. We found nuanced mechanisms of surgeon and surgical team learning that could lead to better and more efficient usage of robotic technology in the hospital. While we found that both surgeons and surgical teams learn by doing robot-assisted surgeries, surgeon learning-by-doing is greater than surgical team-learning-by-doing. A surgeon's learning is associated not only with cumulative volume of surgeries performed but also the regularity with which a surgeon performs robot-assisted surgical procedures. The effect of regularity on learning is more significant in the initial stages of a surgeon performing robot-assisted surgeries. And, surgeons learn more while performing complex surgical procedures.

We found that surgeon as well as surgical team learning lead to better usage of the surgical robot, but the effect of surgeon learning is more significant than the surgical team learning in improving the usage of the robot. Finally, we found that the adoption of robotic technology is associated with improvement in clinical quality and reduction in surgical procedure duration.

We believe that findings of this study are significant from the standpoint of a health care

delivery system. One of the key issues that constrain the accessibility of critical and complex health care delivery procedures is the availability of surgeons and surgical staff with appropriate and adequate training and skills. In this study, we showed that technology mediation by way of a surgical robot can mitigate much of the negative effects of skill requirements to effectively perform surgical procedures. Moreover, the findings of this study provide health care delivery organizations insights into the mechanisms of learning related to the effective adoption of a robotic surgical technology.

In conclusion, we believe that this study will expand the focus of the extant literature in health care operations management. To date, the academic and practitioner literature on health care operations management has focused on controlling input heterogeneity to reduce variation in health care delivery outcome. However, this study demonstrated that a high tech innovation such as a surgical robot can reduce variation in health care delivery outcome in spite of input heterogeneity, thereby contributing towards the broadening of perspective of academic researchers and practitioners, and providing a springboard for new lines of future inquiries.

Chapter 5:

Concluding Remarks

5.1 Conclusions and Key Contributions

The dissertation research is motivated by real incidences of failures of high tech innovations-in-use in several industries with severe consequences to the users and manufacturers alike. I tried to answer two broad questions: (i) how can firms proactively manage the downside risk of failure of high tech innovations-in-use? and (ii) how can users realize the full potential of high tech innovation? Consequently, the dissertation is divided into two broad parts. Part one consists of two studies using predictive analytics and econometrics on a 'big data-set' related to used feedback of adverse events of medical devices, along with several related data-sets. The summary contributions of the two studies are as follows.

First, the studies demonstrate that the prediction of failures of high tech innovations-in-use is possible with sufficient lead time by analyzing big and unstructured data related to user feedback on adverse events in the marketplace; and that the precision of such predictions can be significantly improved by accounting for time-varying covariates related to design, supply chain and manufacturing. Second, the study yields novel and nuanced insights into why firms often fail to correctly detect market signals of failures of high tech innovations-in-use. In particular, the study (i) shows that the existence of judgment bias results in firms to under-react or over-react to market signals in the form of user feedback on adverse events, and (ii) identifies factors – such as the severity of the adverse events and noise-to-signal ratio in the user feedback data stream on adverse events – that influence firms to under-react or over-react. The second study details out product, firm and industry related conditions for systematic variation of the extent and type of judgment bias, namely, under-reaction and over-reaction bias. Finally, this study being the very first to have applied predictive analytics to develop models with a big and unstructured database to predict failures of high tech innovations-in-use and evaluate judgment bias, opens up new empirical research possibilities of conducting technology and innovation management research. Such possibilities include: (i) going beyond an orientation of innovation success to innovation failure, especially

when the innovations are in use in the marketplace; (ii) going beyond an orientation of explanation to prediction while accounting for judgment bias; and (iii) going beyond the analysis of structured data-sets containing hundreds or thousands of observations to analysis of unstructured data-sets containing millions of observations or more. I believe that apart from innovation management, the studies have significant practice and policy implications in terms of regulations and regulatory surveillance related to high tech innovations.

In the second part of the dissertation I have undertaken a field study to: (i) understand the impact of surgical robot on the variation in the outcomes of surgical procedures, given input heterogeneity – namely, heterogeneity in the skills and experience of surgeons and surgical team members; (ii) analyze the nuanced mechanisms of surgeon and surgical team learning that lead to effective usage of robots in performing surgeries. The major findings of the study are the following. First, we find that the robotic surgical interface does indeed mitigate outcome variations that are otherwise expected across surgical procedures. Second, we find that the learning mechanism in the context of a robot-assisted surgery is more nuanced than cumulative volume-based learning. Third, we find that given specific levels of surgical volume, individual learning of a surgeon is significantly dependent on the regularity with which the surgeons perform robot-assisted surgeries. Finally, this study sheds light on the interdependency of duration and quality outcome measures of robot-assisted surgical procedures, thereby providing new insights into the speed versus quality debate in managing health care operations.

5.2 Future Research Direction

This dissertation research opens up a few new research directions which can be explored further. Firstly, this is one of the earlier research that focuses on prediction of failures of innovations-in-use. Most of the past research in this area are explanatory in nature. I believe that this research will open up new avenues of prediction of innovation performance, both success and failures, in many different industries such as automobiles, pharmaceuticals, electronics and telecommunication and consumer durables. Also, this research is likely to encourage usage of different types of data and variables such as firm specific supply chain, manufacturing and design related variables to set-up studies which has higher resolution of prediction both in terms of the level of analysis as well as in terms of the time frame of prediction. Methodologically, this study is one of the earlier studies in using predictive analytics on ‘big data’ such as social media data, text data, voice data, mobile data, real time health monitoring data, real time user feedback, and transaction data. This study lays out

a schema for using big unstructured data to set-up and conduct scientific enquiry in management sciences and is likely to be used as a guide for designing similar studies in the future. All in all, this study is likely to encourage fairly novel and new directions of research in designing research studies using predictive analytics on big unstructured data from diverse sources.

Bibliography

Adler, W., Brenning, A., Potapov, S., Schmid, M., & Lausen, B. (2011). “Ensemble classification of paired data”. *Computational Statistics & Data Analysis*, *55*(5), 1933-1941.

Ahlering, Thomas E., et al. (2004). "Robot-assisted versus open radical prostatectomy: a comparison of one surgeon's outcomes." *Urology* 63(5), 819-822.

Anand, G., Gray, J., & Siemsen, E. (2012). "Decay, shock, and renewal: operational routines and process entropy in the pharmaceutical industry." *Organization Science,* 23(6), 1700-1716.

Argote, Linda, Sara L. Beckman, and Dennis Epple. (1990). "The persistence and transfer of learning in industrial settings." *Management Science.* 36(2), 140-154.

Atella, Vincenzo, Federico Belotti, and Domenico Depalo. (2011). *Disentangling true determinants of drug therapy effectiveness: A double fixed effects approach*. No. 186. Tor Vergata University, CEIS.

Banker, R. D., Khosla, I., & Sinha, K. K. (1998). "Quality and competition." *Management Science,* 44(9), 1179-1192.

Barnard, C., and Herbert A. Simon. Administrative behavior. (1947). *A study of decision-making processes in administrative organization*. Macmillan, New York.

Benediktsson, J. O. N. A., Swain, P. H., & Ersoy, O. K. (1990). “Neural network approaches versus statistical methods in classification of multisource remote sensing data”. *IEEE Transactions on geoscience and remote sensing*, *28*(4), 540-552.

Binder, J., and W. Kramer. (2001). "Robotically-assisted laparoscopic radical prostatectomy." *BJU International,* 87(4), 408-410.

Birkmeyer, John D., and Justin B. Dimick. (2009). "Understanding and reducing variation in surgical mortality." *Annual Review of Medicine,* 60, 405-415.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). “Latent dirichlet allocation”. *The Journal of machine learning research*, *3*, 993-1022.

Breiman, L. (2001). “Random forests”. *Machine learning*, *45*(1), 5-32.

Breiman, L. (1999). *Using adaptive bagging to debias regressions*. Technical Report 547, Statistics Dept. UCB.

Breslow, N. E. (1984). “Extra-Poisson variation in log-linear models”. *Applied Statistics*, 38-44.

Breslow, N. E., & Clayton, D. G. (1993). “Approximate inference in generalized linear mixed

models". *Journal of the American Statistical Association*, *88*(421), 9-25.

Carlucci, Daniela, Paolo Renna, and Giovanni Schiuma. (2013). "Evaluating service quality dimensions as antecedents to outpatient satisfaction using back propagation neural network." *Health care management science*, 16(1), 37-44.

Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. (2012). "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS quarterly,* 36(4), 1165-1188.

Chowdhury, M. M., H. Dagash, and A. Pierro. (2007). "A systematic review of the impact of volume of surgery and specialization on patient outcome." *British Journal of Surgery,* 94(2), 145-161.

Christensen, Clayton M., Richard Bohmer, and John Kenagy. (2000). "Will disruptive innovations cure health care?" *Harvard Business Review,* 78(5), 102-112.

Cutler, David M., and Robert S. Huckman.(2003). "Technological development and medical productivity: the diffusion of angioplasty in New York state." *Journal of Health Economics* 22(2), 187-217.

Edmondson, Amy C., et al. (2003). "Learning how and learning what: Effects of tacit and codified knowledge on performance improvement following technology adoption." *Decision Sciences* 34(2), 197-224.

Edmondson, Amy C., James R. Dillon, and Kathryn S. Roloff. (2007). "Three Perspectives on Team Learning: Outcome Improvement, Task Mastery, and Group Process." *The Academy of Management Annals,* 1(1), 269-314.

Edmondson, Amy C., Richard M. Bohmer, and Gary P. Pisano. (2001). "Disrupted routines: Team learning and new technology implementation in hospitals." *Administrative Science Quarterly* 46(4), 685-716.

Fong Boh, Wai, Sandra A. Slaughter, and J. Alberto Espinosa. (2007). "Learning from experience in software development: A multilevel analysis." *Management Science,* 53(8), 1315-1331.

Fuhr, T., George, K., & Pai, J. (2013). "The Business Case for Medical Device Quality". *McKinsey & Company.*

Ghasemi, Jahan B., and Hossein Tavakoli. (2013). "Application of random forest regression to spectral multivariate calibration." *Analytical Methods,* 5(7), 1863-1871.

GAO-11-468. (2011). FDA Should Enhance its Over-sight of Recalls. GAO Report. (http://www.gao.gov/products/GAO-11-468)

GAO (United States Government Accountability Office) (August 2012). Medical Devices: FDA

Should Expand Its Consideration of Information Security for Certain Types of Devices, GAO-12-816. (http://www.gao.gov/products/GAO-12-816)

George, G., Haas, M. R., & Pentland, A. (2014). “Big data and management”. *Academy of Management Journal*, 57(2), 321-326.

Goldstein, H. (1991). “Nonlinear multilevel models, with an application to discrete response data”. *Biometrika*, 45-51.

Gokpinar, Bilal, Wallace J. Hopp, and Seyed MR Iravani. (2010). "The impact of misalignment of organizational structure and product architecture on quality in complex product development." *Management Science,* 56(3), 468-484.

Gokpinar, Bilal, Wallace J. Hopp, and Seyed MR Iravani. (2013). "In-House Globalization: The Role of Globally Distributed Design and Product Architecture on Product Development Performance." *Production and Operations Management.* 22(6), 1509-1523.

Green, P. J. (1987). “Penalized likelihood for general semi-parametric regression models”. *International Statistical Review/Revue Internationale de Statistique*, 245-259.

Greve, H. R. (2008). “A behavioral theory of firm growth: Sequential attention to size and performance goals”. *Academy of Management Journal*, 51(3), 476-494.

Haas, M. R., Criscuolo, P., & George, G. (2015). “Which Problems to Solve? Online Knowledge Sharing and Attention Allocation in Organizations”. *Academy of Management Journal*, 58(3), 680-711

Harvey Jr, L. O., Hammond, K. R., Lusk, C. M., & Mross, E. F. (1992). “The application of signal detection theory to weather forecasting behavior”. *Monthly Weather Review*, *120*(5), 863-883.

Hastie, T., & Tibshirani, R. (1986). “Generalized additive models”. *Statistical science*, 297-310.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning theory*.

Hastie, Trevor, et al. (2009). *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer,

Hannan, Edward L., et al. (1990). "Adult open heart surgery in New York State: an analysis of risk factors and hospital mortality rates." *Journal of American Medical Association,* 264(21), 2768-2774.

Hemal, A. K., and M. Menon. (2002). "Laparoscopy, robot, telesurgery and urology: future perspective." *Journal of Postgraduate Medicine,* 48(1), 39.

Hemal, A. K., et al. "Laparoscopic versus open radical nephrectomy for large renal tumors: a long-term prospective comparison." *The Journal of Urology,* 177(3), 862-866.

Hendryx, Michael S., et al. (2002). "Access to health care and community social capital." *Health*

Services Research, 37(1), 87-104.

Hoffman, A. J., & Ocasio, W. (2001). "Not all events are attended equally: Toward a middle-range theory of industry attention to external events". *Organization Science*, 12(4), 414-434.

Hollingsworth, Bruce. (2008). "The measurement of efficiency and productivity of health care delivery." *Health Economics* 17(10), 1107-1128.

Huber, Peter J. (2011). *Robust statistics*. Springer: Berlin Heidelberg.

Hu, Jim C., et al. (2009). "Comparative effectiveness of minimally invasive vs. open radical prostatectomy." *Journal of American Medical Association,* 302(14), 1557-1564.

Huckman, Robert S., and Bradley R. Staats. (2011). "Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance." *Manufacturing & Service Operations Management,* 13(3), 310-328.

Huckman, Robert S., Bradley R. Staats, and David M. Upton. (2009). "Team familiarity, role experience, and performance: Evidence from Indian software services." *Management Science,* 55(1), 85-100.

Huckman, Robert S., Bradley R. Staats, and David M. Upton. (2009). "Team familiarity, role experience, and performance: Evidence from Indian software services." *Management Science* 55(1), 85-100.

Hwang, Jason, and Clayton M. Christensen. (2008). "Disruptive innovation in health care delivery: a framework for business-model innovation." *Health Affairs,* 27(5), 1329-1335.

Institute of Medicine (US). (2001). Committee on Quality of Health Care in America. *Crossing the quality chasm: A new health system for the 21st century*. *National Academies Press*.

Jacome, Enrique G., April E. Hebert, and Frank Christian. (2013). "Comparative analysis of vaginal versus robotic-assisted hysterectomy for benign indications." *Journal of Robotic Surgery*, 7(1), 39-46.

Joseph, J., & Ocasio, W. (2012). "Architecture, attention, and adaptation in the multibusiness firm: General Electric from 1951 to 2001". *Strategic Management Journal*, 33(6), 633-660.

Kansagara, Devan, et al. (2011). "Risk prediction models for hospital readmission: a systematic review." *Journal of Medical Association,* 306(15), 1688-1698.

Kass-Hout, T., & Zhang, X. (Eds.). (2010). *Biosurveillance: Methods and case studies*. Taylor & Francis US.

KC, Diwas Singh, and Bradley R. Staats. (2012). "Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance." *Manufacturing & Service*

Operations Management, 14(4), 618-633.

Kremer, M., Moritz, B., & Siemsen, E. (2011). “Demand forecasting behavior: System neglect and change detection”. *Management Science*, *57*(10), 1827-1843.

Li, Q., Maggitti, P. G., Smith, K. G., Tesluk, P. E., & Katila, R. (2013). “Top management attention to innovation: The role of search selection and intensity in new product introductions”. *Academy of Management Journal*, 56(3), 893-916.

Lyles, M. A., Flynn, B. B., & Frohlich, M. T. (2008). “All supply chains don't flow through: Understanding supply chain issues in product recalls”. *Management and Organization Review*, *4*(2), 167-182.

Lynch, Clifford. (2008). "Big data: How do your data grow?" *Nature* 455(7209), 28-29.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide.*

Madsen, Rasmus E., David Kauchak, and Charles Elkan. (2005). "Modeling word burstiness using the Dirichlet distribution." *Proceedings of the 22nd International Conference on Machine Learning. ACM*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute.

Martino, Martin A., et al. (2014). "A Comparison of Quality Outcome Measures in Patients Having a Hysterectomy for Benign Disease: Robotic vs. Non-robotic Approaches." *Journal of Minimally Invasive Gynecology,* 21(3), 389-393.

Marucheck, A., Greis, N., Mena, C., & Cai, L. (2011). “Product safety and security in the global supply chain: Issues, challenges and research opportunities”. *Journal of Operations Management*, *29*(7), 707-720.

Massey, C., & Wu, G. (2005). “Detecting regime shifts: The causes of under-and overreaction”. *Management Science*, *51*(6), 932-947.

McCarter, Freda D., et al. (2000). "Institutional and individual learning curves for focused abdominal ultrasound for trauma: cumulative sum analysis." *Annals of Surgery,* 231(5), 689.

McKee, Martin, et al. (1998). "Organisational change and quality of health care: an evolving international agenda." *Quality in Health Care*: QHC, 7(1), 37.

Menon, Mani, et al. (2004). "Vattikuti Institute prostatectomy, a technique of robotic radical prostatectomy for management of localized carcinoma of the prostate: experience of over 1100 cases." *Urologic Clinics of North America,* 31(4), 701-717.

Mitchell, T. M. (1999). “Machine learning and data mining”. *Communications of the ACM*, *42*(11), 30-36.

Nigam, Kamal, et al. (2000). "Text classification from labeled and unlabeled documents using EM." *Machine Learning,* 39(2-3), 103-134.

Ocasio, William. (1997) "Towards an Attention-Based View of the Firm”, *Strategic Management Journal*, 18(Special Issue: Organizational and Competitive Interactions). 187-206.

Ocasio, W. (2011). “Attention to attention”. *Organization Science*, 22(5), 1286-1296.

Piao, Yongjun, et al. (2012). "An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data." *Bioinformatics,* 28(24), 3306-3315.

Plsek, Paul. (2003). "Complexity and the adoption of innovation in health care."*Accelerating Quality Improvement in Health Care: Strategies to Accelerate the Diffusion of Evidence-Based Innovations. Washington, DC: National Institute for Healthcare Management Foundation and National Committee for Quality in Health Care*.

Porter, Michael E. (2010). "What is value in health care?." *New England Journal of Medicine,* 363(26), 2477-2481.

Rassweiler, Jens, et al. (2006). "Laparoscopic and robotic assisted radical prostatectomy–critical analysis of the results." *European Urology,* 49(4), 612-624.

Riviere, Cameron N., R. Scott Rader, and Nitish V. Thakor. (1998). "Adaptive cancelling of physiological tremor for improved precision in microsurgery." *Biomedical Engineering, IEEE Transactions,* 45(7), 839-846.

Riviere, Cameron N., Wei Tech Ang, and Pradeep K. Khosla. (2003). "Toward active tremor canceling in handheld microsurgical instruments." *Robotics and Automation, IEEE Transactions*, 19(5), 793-800.

Rumelhart, D. E. (1986). "McClelland, Back Propagation Training Algorithm Processing."

Savoy, J., & Zubaryeva, O. (2012). “Simple and efficient classification scheme based on specific vocabulary”. *Computational Management Science*, *9*(3), 401-415.

Salkind, N. J. (2007). *Encyclopedia of Measurement and Statistics 3-Volume Set*. California, USA: SAGE Publications.

Schapire, R. E. (1999, July). “A brief introduction to boosting”. In *Ijcai*, 99, 1401-1406.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). “Boosting the margin: A new explanation for the effectiveness of voting methods”. *The annals of statistics*, *26*(5), 1651-1686.

Shah, R., Ball, G. & Netessine, S. (2013). "Plant Operations and Product Recalls in the Automotive

Industry: An Empirical Investigation." Working Paper. (SSRN ID 2356315)

Shmueli, G., & Koppius, O. (2010). Predictive analytics in information systems research. *Robert H. Smith School Research Paper No. RHS*, 06-138.

Shmueli, G. (2010). “To explain or to predict?”. *Statistical Science*, *25*(3), 289-310.

Simpson, A. J., & Fitter, M. J. (1973). “What is the best index of detectability?”. *Psychological Bulletin*, *80*(6), 481.

Soliman, Pamela T., et al. (2011). "Radical hysterectomy: a comparison of surgical approaches after adoption of robotic surgery in gynecologic oncology." *Gynecologic Oncology,* 123(2), 333-336.

Staats, Bradley R. (2012). "Unpacking team familiarity: The effects of geographic location and hierarchical role." *Production and Operations Management,* 21(3), 619-635.

Staats, Bradley R., and Francesca Gino. (2012). "Specialization and variety in repetitive tasks: Evidence from a Japanese bank." *Management Science,* 58(6), 1141-1159.

Steinberg, Peter L., et al. (2008). "The cost of learning robotic-assisted prostatectomy." *Urology* 72(5), 1068-1072.

Stitzenberg, Karyn B., and George F. Sheldon. (2005). "Progressive specialization within general surgery: adding to the complexity of workforce planning." *Journal of the American College of Surgeons,* 201(6), 925-932.

Sweeney, Kieran, and Frances Griffiths, (2002). *Complexity and Healthcare: An Introduction*. Radcliffe Publishing.

Swets, John A. (1986a) "Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance." *Psychological Bulletin,* 99(2), 181.

Swets, John A. (1986). "Indices of discrimination or diagnostic accuracy: their ROCs and implied models." *Psychological Bulletin,* 99(1), 100.

Taylor, Russell, et al. (1999). "A steady-hand robotic system for microsurgical augmentation." *The International Journal of Robotics Research*, 18(12), 1201-1210.

Thirumalai, S., & Sinha, K. K. (2011). “Product recalls in the medical device industry: an empirical exploration of the sources and financial consequences”. *Management Science*, *57*(2), 376-392.

Tibshirani, R. (1996a). *Bias, variance and prediction error for classification rules*. University of Toronto, Department of Statistics.

Tibshirani, R. (1996b). “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Verikas, Antanas, Adas Gelzinis, and Marija Bacauskiene. (2011). "Mining data with random

forests: A survey and results of new tests." *Pattern Recognition,* 44(2), 330-349.

Wagner, M. M., Tsui, F. C., Espino, J. U., Dato, V. M., Sitting, D. F., Caruana, R. A. & Fridsma, D. B. (2001). "The emerging science of very early detection of disease outbreaks". *Journal of Public Health Management and Practice*, *7*(6), 51-59.

Waldman, J. Deane, Steven A. Yourstone, and Howard L. Smith. (2003). "Learning curves in health care." *Health Care Management Review,* 28(1), 41-54.

Walther, D., & Koch, C. (2006). "Modeling attention to salient proto-objects". Neural networks, 19(9), 1395-1407.

Wickens, Christopher D. (2002). "Spatial awareness biases."

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 144-148.

Wright, Jason D., et al. (2013). "Robotically assisted vs. laparoscopic hysterectomy among women with benign gynecologic disease." *Journal of American Medical Association,* 309(7). 689-698.

Xu, Rena, et al. (2013). "The teaming curve: a longitudinal study of the influence of surgical team familiarity on operative time." *Annals of Surgery,* 258(6), 953-957.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, *86*(413), 79-86.

Zhang, H. (2004). The optimality of naive Bayes. *A A*, *1*(2), 3.

Zhu, X., & Davidson, I. (Eds.). (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. Igi Global.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429.

Appendices

A Appendices Related to Chapter 2.

A.1 Device Classification from Text Description

Start

Step 1: *Generate bag of unique words from train data set (510K and PMA classification). Split sample into 80% train and test set.*

Step 2: *Create matrix of size n(number of unique words) X k(number of device classes) Frequency[i,j]=0 (initialize)*

Step 3: *Populate matrix*

For in in 1: N (Total number of words in train set with repetition), do

If(Word[i] → Class[k])

Frequency[i,k]++

End For

Step 4: *Calculate Entropy*

$Entropy[i]=Sum[Frequency[i,k]/Sum_{(k)}\{Frequency[i,k]]$

Step 5: *Calculate Threshold values [Alpha and Beta] using grid search*

For all Alpha [0,1] and all Beta [0,max(Entropy)], do

Calculate Probability[Device[i] → Class[k]] using equation [2.9] on test set.

Repeat for all devices.

Calculate fraction of devices correctly classified.

Retain results for max[correct]

End For

End

The algorithm as well as many of the predictive models were programmed in Python and R environment, and run on a supercomputer using multithreading.

A.2 Descriptions of the Prediction Models Estimated

A.2.1 Generalized Linear Mixed Effects Model (GLMM)

GLMMs are widely used in econometric estimations, particularly for data with over dispersion that is often observed in data-sets with binomial, Bernoulli or Poisson distribution (Williams, 1982; Breslow, 1984; Breslow and Clayton, 1993). GLMMS are also appropriate where the response variable is longitudinally distributed with fixed or random variation within observation units. The data-set for this study is longitudinal in nature with several observation units. The primary observation units are the products. There are 766 products distributed among 105 firms in the final data-set after eliminating some products where there were either no data or very sparse data, or where data on all the variables were not available. We eliminated some firms and some products where the observations were so sparse so as not to be meaningful for model estimation purposes. Also, during the course of organizing the study data we had largely accounted for mergers, take-overs, spin-offs and change of ownership in the industry. We estimated the GLMMs with random effects for the product as well as the firms. Both the number of firms and the number of products were not large enough to estimate any fixed effect models. We did, however, included fixed effects for the product class, the usage class, the approval class and the implant status. There are several methods for estimating GLMMs. The primary among them are marginal quasi likelihood (Goldstein, 1991), penalized quasi likelihood (Green, 1987) Bayesian procedures using Gibbs sampling technique (Zeger and Karim, 1991). We adopted the Gibbs sampling approach to estimate the models using the MCMCglmm package in R statistical computing environment.

A.2.2 Generalized Additive Model (GAM)

GAMs assume a linear function for the covariate effects. However, from a predictive standpoint, it is important to account for the non-linearity in the data. Accommodating for non-linearity in the data is important for predictive accuracy. For explanatory models, where the mean level of directional information is more critical than the local non-linearities, it may be acceptable to confine the model building effort to linear models. However, in predictive models, non-linear dynamics of the predictive variables can be critical for improving the predictive accuracy of the models. GAM replaces the linear functions of the linear models ($\sum_i X_j \beta_j$) by a sum of smooth non-linear non-parametric functions ($\sum_i s_j(X_j)$), where $s_j(.)$'s are unspecified functions estimated using local spline smoothers (Hastie and Tibshirani, 1984). Instead of the standard log-likelihood maximization, GAM minimizes the Kullback-Leibler divergence of the true model from the data.

Both log-likelihood and Kullback-Leibler divergence asymptotically converges to the true model parameters and hence are consistent estimators of the true model.

A.2.3 Artificial Neural Network (ANN) Classifier

One efficient way of analyzing data with complex relationships is to decompose the relationships into simple units of relationships and then recombining the simple units into a complex output. ANN enables doing the above by representing the data analysis problem as a network of interconnected relatively simpler computational units called nodes. The nodes receive inputs in the form of data or output from other nodes. The nodes process the inputs usually using a function like a logistic function and produce outputs. The connection between the nodes determines the pattern of information flow in the entire network. Information flow can be unidirectional or bidirectional. One of the mechanisms to reduce error of classification model is by a mechanism called "back propagation" (Rumelhart, 1986) in multi-layered feed-forward ANNs. This means that the nodes are organized in multiple layers (usually two or more) and the signal travels "forward" from the input layers to the output layers, and the error is propagated backwards from the output layer to the input layer. The basic idea is to use the feedback loop of errors to minimize the error of classification. We estimated the ANN network with three layers (1 hidden layer) with feed-forward back-propagation mechanism to achieve higher classification accuracy. The first step in the model building is to train the network with training data-set which, in this case, is 80% of the data points chosen randomly. In the training step the network estimates the weights of each of the nodes. This step is called the learning step where the network learns the weights. The next step is the classification step where the network predicts classification probabilities on the test data set and calculates the classification error. Neural networks have been used in several areas of scientific inquiry such as natural sciences, social sciences and health care. Carlucci et al (2013) use neural networks with back-propagation to analyze the effect of several dimensions of service quality in hospitals on the final patient satisfaction. The authors demonstrate the suitability of ANN as an effective knowledge discovery technique for identifying the service quality dimensions that are important to outpatients. Given the complexity of relationships and interdependence of the dimensions they show that ANN performs better than a standard statistical model. Situations where the population distribution can be fairly accurately estimated, and where the interdependence between the variables are relatively simple, standard statistical estimation models outperform ANN. However, where the data dimension and distribution complexity is high with complex

interdependencies, as in our case, ANN has shown to be a much better estimation model than any of the standard statistical regression models (Benediktsson, Swain, Ersoy, 1990).

A.2.4 Naïve Bayes Estimator

A Naïve Bayes Classifier (Mitchell, 1997) is a probabilistic classification model for data with high dimension and variable interdependence. Naïve Bayes classification estimation has been shown to be superior to standard statistical models for conditions of high interdependence of predictive data (Zhang, 2004). Naïve Bayes classifiers have been extensively used for text classification (Savoy and Zuberyeva, 2012) and other classification problems. Naïve Bayes classification models have also gained popularity in knowledge discovery in public health and health care delivery. The primary reason for using Naïve Bayes is its simplicity of estimation and efficiency of classification of high dimensional complex data-sets.

A.2.5 Ensemble Classifiers

The idea of ensemble classifiers is to combine the outcome of several classifiers to achieve higher accuracy of classification (Adler et al, 2011). The key step in ensemble classifiers is to form an ensemble of multiple classifiers from a single training set. This is achieved by either subsampling the data using bootstrap methods or by modifying the classification algorithm by exploiting the randomization components of the classification algorithms like randomizing starts of decision trees. The combination of the outputs is done using one of several techniques such as frequency counts, aggregating using respective error weights, or by simple averaging. Ensemble classifiers have been shown to have superior classification and prediction accuracy as compared to single classifiers (Hastie, Tibshirani and Friedman, 2001). In our problem, we consider two popular ensemble class models, namely, boosting and random forests.

A.2.5.1 Boosting Classifier

Boosting classifier is an ensemble classifier that uses subsamples of the train set to build a number of classification models (Schapire, 1990 & 1996; Brieman, 1996). In boosting, the subsampling choice is done based on the performance of the earlier classifiers. Observations that are incorrectly predicted by the earlier classifier are assigned higher probability of inclusion than observations that were correctly predicted for the subsequent subsample in the next iteration of the model. Thus, boosting classification attempts to generate new classifiers at each iteration that are able to predict

better the observations that were not correctly predicted by the ensemble thus far. At each iteration, misclassified observations are accentuated and a better classifier fit is tried. One of the drawbacks of boosting is that it has a tendency to over-fit in the presence of high noise (Brieman, 1996). The boosting efficiency is susceptible to noise and can produce relatively higher classification error in the presence of high noise in the data.

A.2.5.2 Random Forest Classifier

Random forest is an ensemble classifier that can be used for either classification or regression. Random forests work by generating ensembles of regression trees built on independent random subsamples of the training data (Breiman, 2001). The classification ensemble is generated by the modal prediction class. It has been shown that the classification accuracy of random forests depend on the number of classification trees, i.e., the size of the ensemble. The prediction error asymptotically converges almost surely to a limit as number of trees in the forest becomes large. Random forests have been widely used in machine learning, bioinformatics, climate sciences and other natural sciences. Using several data-sets from economics and environment science, Breiman (2001) has demonstrated that random forests achieve significantly superior prediction accuracy over single classifiers or even over other ensemble class classifiers. Unlike boosting classifiers, random forests are relatively more noise tolerant and produce better class probability estimates than boosting in the presence of considerable noise in the data.

A.3 Derivation of Equation [2.7]

$$Firm\ objective: \quad \min_{X_c}\{C_o * P(Miss) + C_s * P(FA)\}$$

$$\min_{X_c}\left\{C_o * \int_{-\infty}^{X_c} N(x|\mu_s, \sigma_s^2)dx + C_s * \int_{X_c}^{\infty} N(x|\,\mu_o, \sigma_o^2)dx\right\}$$

The first order condition is given by:

$$\frac{\partial}{\partial X_c}\left\{C_o * \int_{-\infty}^{X_c} N(x|\mu_s, \sigma_s^2)dx + C_s * \int_{X_c}^{\infty} N(x|\,\mu_o, \sigma_o^2)dx\right\} = 0$$

$$\left\{C_o * \frac{\partial}{\partial X_c}\int_{-\infty}^{X_c} N(x|\mu_s, \sigma_s^2)dx + C_s * \frac{\partial}{\partial X_c}\int_{X_c}^{\infty} N(x|\,\mu_o, \sigma_o^2)dx\right\} = 0$$

$$\left\{C_o * \frac{\partial}{\partial X_c}[\Phi(\mathrm{X_c}|\mu_s, \sigma_s^2) - \Phi(-\infty)] + C_s * \frac{\partial}{\partial X_c}[\Phi(\infty) - \Phi(\mathrm{X_c}|\mu_s, \sigma_s^2)]\right\} = 0$$

$$C_o * N(x|\mu_s, \sigma_s^2) = C_s * N(x|\mu_o, \sigma_o^2)$$

Therefore,

$$Likelihood\ Ratio\ \ \beta = \frac{N(x|\mu_s, \sigma_s^2)}{N(x|\mu_o, \sigma_o^2)} = \frac{C_s}{C_o}$$

Let,

$$\lambda = \log(\beta) = \log\left(\frac{C_s}{C_o}\right) = \log\left(\frac{Sunk\ Cost}{Opportunity\ Cost}\right)$$

Therefore, the optimal decision criterion is obtained from the equation:

$$\lambda = \log\left(\frac{N(x|\mu_s, \sigma_s^2)}{N(x|\mu_o, \sigma_o^2)}\right) = \log\big(N(x|\mu_s, \sigma_s^2)\big) - \log\big(N(x|\mu_o, \sigma_o^2)\big)$$

For, simplification considering $\sigma_s = \sigma_o = \sigma$, we get

$$-\frac{1}{2}\log(2\pi\sigma) - \frac{1}{2}\frac{(X_c - \mu_s)^2}{\sigma^2} \pm \frac{1}{2}\log(2\pi\sigma) + \frac{1}{2}\frac{(X_c - \mu_o)^2}{\sigma^2} = \lambda$$

$$2(\mu_s - \mu_o)X_c - \{(\mu_s^2 - \mu_o^2) - 2\lambda\sigma^2\} = 0$$

Hence,

$$X_c = \frac{\mu_s + \mu_o}{2} + \frac{\lambda\sigma^2}{\mu_s - \mu_o}$$

WLOG letting the white noise distribution mean to be zero, $\mu_o = 0$

$$X_c = \frac{\mu_s}{2} + \frac{\lambda\sigma^2}{\mu_s}$$

A.4 Derivation of Equation [2.9]

Let f_{ij} represent the number of times word w_i appear in class c_j. Naïve estimates of some of the probability values would be:

$$P(w = w_i|c = c_j) = \frac{f_{ij}}{\sum_i f_{ij}}; P(w = w_i) = \frac{\sum_j f_{ij}}{\sum_j \sum_i f_{ij}}; P(c = c_j) = \frac{\sum_i f_{ij}}{\sum_j \sum_i f_{ij}}$$

The classification problem is to find out the probability that a device belongs to a class c_j given that it contains a set of words $W = \{w_1, \dots, w_k\}$. We assume that the probability that a word with frequency f_{ij} belongs to a specific class c_j is distributed according to the dirichlet process $Dir\{f_{ij}|\theta_j\}$. Then

$$P(\{w_1, \dots, w_k\}|c = c_j) = Dir\{f_{ij}|\theta_j\} = \frac{1}{B(\boldsymbol{f})}\Pi_{\{j=1\}}^{k}\theta_j^{\{f_{ij}\}}, where\ \boldsymbol{f} \in \{f_{ij}\}, B(\boldsymbol{f}) = \frac{\{\Pi_{\{j=1\}}^{k}\Gamma(f_{ij})\}}{\{\Gamma(\sum_{\{j=1\}}^{k} f_{ij})\}}$$

The objective of text classification is to find the maximum likelihood estimates for the probability vector θ_j.

$$argmax_{\theta_j}\ \ell(\boldsymbol{f}|\boldsymbol{\theta_j}) = \sum_{\{j=1\}}^{k} \log\Gamma(f_{ij}) - \log\Gamma\left(\sum_{\{j=1\}}^{k} f_{ij}\right) + \sum_{\{j=1\}}^{k} f_{ij}\log\theta_j\,, s.t.\ \sum_{\{j=1\}}^{k}\theta_j = 1$$

This is equivalent to maximizing the Lagrangian of the above problem,

$$argmax_{\theta_j}\ L_\lambda = \sum_{\{j=1\}}^{k} \log\Gamma(f_{ij}) - \log\Gamma\left(\sum_{\{j=1\}}^{k} f_{ij}\right) + \sum_{\{j=1\}}^{k} f_{ij}\log\theta_j + \lambda\left(1 - \sum_{\{j=1\}}^{k}\theta_j\right)$$

The first order condition is given by,

$$\frac{\partial L_\lambda}{\partial\theta_j} = \frac{f_{ij}}{\theta_j} - \lambda = 0, \forall j$$

$$\Rightarrow\ \theta_j = \frac{f_{ij}}{\lambda}$$

$$\Rightarrow \sum_j \theta_j = 1 = \frac{\sum_j f_{ij}}{\lambda}$$

$$\Rightarrow \lambda = \sum_j f_{ij}$$

Therefore we have the maximum likelihood Dirichlet parameter,

$$\theta_j = \frac{f_{ij}}{\sum_j f_{ij}}$$

$$P\big(c = c_j \big| w = \{w_1, \dots, w_k\}\big) = \frac{P\big(\{w_1, \dots, w_k\} \big| c = c_j\big) P\big(c = c_j\big)}{P(\{w_1, \dots, w_k\})}$$

$$= \frac{\left\{\frac{1}{B(\boldsymbol{f})} \prod_{i=1}^{k} \frac{P\big(w_i \big| c_j\big)^{f_{ij}}}{f_{ij}!}\right\} P\big(c = c_j\big)}{\prod_{i=1}^{k} P(w_i)}$$

Therefore,

$$P\big(c = c_j \big| w = \{w_1, \dots, w_k\}\big) = \frac{\left\{\frac{1}{B(\boldsymbol{f})} \prod_{i=1}^{k} \frac{\left(\frac{f_{ij}}{\sum_i f_{ij}}\right)^{f_{ij}}}{f_{ij}!}\right\} \frac{\sum_i f_{ij}}{\sum_j \sum_i f_{ij}}}{\prod_{i=1}^{k} \frac{\sum_j f_{ij}}{\sum_j \sum_i f_{ij}}}$$

B Appendices Related to Chapter 3.

B.1 Derivation of the Attention Saliency model

$$P\left[S_1 \geq S_j; j \neq 1\right] = P[S_1 \geq S_2, S_1 \geq S_3, \ldots, S1 \geq S_N] = \prod_{i=2}^{N} P[S_1 \geq S_i]; i.i.d\ S_i$$

$$\Rightarrow P\left[S_1 \geq S_j; j \neq 1\right] = \prod_{i=2}^{N} P[S_i \leq S_1] = \prod_{i=2}^{N} F(S_1) = [F(S_1)]^{N-1}$$

$$\Rightarrow \log P\left[S_1 \geq S_j; j \neq 1\right] = (N-1) * log[F(S_1)]$$

$$\Rightarrow \frac{\partial \log P\left[S_1 \geq S_j; j \neq 1\right]}{\partial N} = log[F(S_1)]$$

$$\Rightarrow \frac{\partial \log P\left[S_1 \geq S_j; j \neq 1\right]}{\partial P\left[S_1 \geq S_j; j \neq 1\right]} \times \frac{\partial P\left[S_1 \geq S_j; j \neq 1\right]}{\partial N} = \frac{1}{P\left[S_1 \geq S_j; j \neq 1\right]} \times \frac{\partial P\left[S_1 \geq S_j; j \neq 1\right]}{\partial N}$$
$$= -|log[F(S_1)]|; \{since, F(S_1) \leq 1 \Rightarrow \log[F(S_1)] \leq 0\}$$

$$\frac{\partial P\left[S_1 \geq S_j; j \neq 1\right]}{\partial N} = -|log[F(S_1)]| P\left[S_1 \geq S_j; j \neq 1\right] = -|log[F(S_1)]|[F(S_1)]^{N-1}$$

C Appendices Related To Chapter 4.

C.1 Derivations of the Posterior Model for Estimation of the System of Equations [4.8], [4.9] and [4.10].

Given,

$$p(y|X,\beta,\sigma_X^2)\sim N(y|X'\beta,\sigma_X^2 I_n)$$

$$(Random\ Effect)\ p(\beta|\sigma_X^2)\sim N(\beta|\mu_\beta,\sigma_X^2 I_n)$$

$$p(\sigma_X^2|X,a,b)\sim IG(a,bX)\sim \frac{(bX)^a}{\Gamma(a)}\left(\frac{1}{\sigma_X^2}\right)^{a+1}\exp\left(-\frac{bX}{\sigma_X^2}\right),\sigma_X^2>0$$

By Bayes formula,

$$p(\beta,\sigma_X^2|X,a,b)=p(\beta|\sigma_X^2)p(\sigma_X^2|X,a,b)$$

$$=\frac{(bX)^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\left(\frac{1}{\sigma_X^2}\right)^{a+\frac{n}{2}+1}\exp\left[-\frac{1}{\sigma_X^2}\left\{bX+\frac{1}{2}(\beta-\mu_\beta)^T(\beta-\mu_\beta)\right\}\right]$$

$$\propto\left(\frac{1}{\sigma_X^2}\right)^{a+\frac{n}{2}+1}\exp\left[-\frac{1}{\sigma_X^2}\left\{bX+\frac{1}{2}(\beta-\mu_\beta)^T(\beta-\mu_\beta)\right\}\right]$$

This is the Normal-Inverse-Gamma (NIG) prior for the slope and variance parameters.

Therefore the posterior joint density given the response is (by Bayes rule):

$$p(\beta,\sigma_X^2|y,X,a,b)=\frac{Prior\times Conditional}{Marginal}$$

$$=\frac{p(\beta,\sigma_X^2|X,a,b)\times p(y|X,\beta,\sigma_X^2)}{p(y)}$$

Where,

$$p(y)=\int p(\beta,\sigma_X^2|X,a,b)\times p(y|X,\beta,\sigma_X^2)\,d\beta d\sigma_X^2$$

Since the parameters are integrated out over the support space, the resulting marginal distribution of the response is free from the parameters and hence we can write,

$$p(\beta,\sigma_X^2|y,X,a,b)\propto\ p(\beta,\sigma_X^2|X,a,b)\times p(y|X,\beta,\sigma_X^2)$$

$$\propto\left(\frac{1}{\sigma_X^2}\right)^{a+\frac{n}{2}+1}\exp\left[-\frac{1}{\sigma_X^2}\left\{bX+\frac{1}{2}(\beta-\mu_\beta)^T(\beta-\mu_\beta)\right\}\right]\times\left(\frac{1}{\sigma_X}\right)\exp\left\{-\frac{1}{\sigma_X^2}(y-X\beta)^T(y-X\beta)\right\}$$

$$\propto\left(\frac{1}{\sigma_X^2}\right)^{a+\frac{n+1}{2}+1}\exp\left[-\frac{1}{\sigma_X^2}\left\{bX+\frac{1}{2}(\beta-\mu_\beta)^T(\beta-\mu_\beta)+(y-X\beta)^T(y-X\beta)\right\}\right]$$

We use the following *multivariate ellipsoidal rectification* identity to complete the derivation of the joint posterior density:

$$\boldsymbol{u^{TAu} - 2\alpha^T u = \left(u - A^{-1}\alpha\right)^T A \left(u - A^{-1}\alpha\right) - \alpha^T A^{-1}\alpha}$$

Where **A** is a symmetric positive definite (hence invertible) matrix. The application of this identity immediately gives us,

$$\left(\frac{1}{\sigma_X^2}\right)\left[bX + \frac{1}{2}(\beta - \mu_\beta)^T(\beta - \mu_\beta) + (y - X\beta)^T(y - X\beta)\right] = \frac{1}{\sigma_X^2}\left[b^* + \frac{1}{2}(\beta - \mu^*)^T(\beta - \mu^*)\right]$$

Using which we can write the posterior as

$$p(\beta, \sigma_X^2|y) \propto \left(\frac{1}{\sigma_X^2}\right)^{a*+\frac{3}{2}} \exp\left\{-\left(\frac{1}{\sigma_X^2}\right)\left[b^* + \frac{1}{2}(\beta - \mu^*)^T V^{*-1}(\beta - \mu^*)\right]\right\}$$

Where,

$$\mu^* = (I_n + X^T X)^{-1}(\mu_\beta + X^T y),$$

$$V^* = (I_n + X^T X)^{-1},$$

$$a^* = a + \frac{n}{2},$$

$$b* = bX + \frac{1}{2}\left[\mu_\beta^T \mu_\beta + y^T y - \mu^{*T} V^{*-1} \mu^*\right]$$

It can be immediately seen that the marginal distribution of the variance parameter is obtained by integrating out the slope parameter from the joint distribution and rearranging relevant terms,

$$p(\sigma_X^2|y) \propto \left(\frac{1}{\sigma_X^2}\right)^{a*+1} \exp\left\{-\frac{b^*}{\sigma_X^2}\right\} \int_{\mathrm{B}} \left(\frac{1}{\sigma_X^2}\right)^{\frac{1}{2}} \exp\left\{-\left(\frac{1}{\sigma_X^2}\right)\left[\frac{1}{2}(\beta - \mu^*)^T V^{*-1}(\beta - \mu^*)\right]\right\} d\beta$$

$$\propto \left(\frac{1}{\sigma_X^2}\right)^{a*+1} \exp\left\{-\frac{b^*}{\sigma_X^2}\right\} \int_{\mathrm{B}} N(\beta|\mu^*, \sigma_X^2) d\beta$$

$$\propto \left(\frac{1}{\sigma_X^2}\right)^{a*+1} \exp\left\{-\frac{b^*}{\sigma_X^2}\right\} \times \mathbf{1}$$

Clearly, the marginal also belongs to the *Inverse Gamma* family and hence,

$$p(\sigma_X^2|y) \sim IG(a^*, b^*) = \frac{b^{*a^*}}{\Gamma(a^*)}\left(\frac{1}{\sigma_X^2}\right)^{a^*+1} \exp\left(-\frac{b^*}{\sigma_X^*}\right)$$

Similarly,

$$p(\beta|y) = \int p(\beta, \sigma_X^2|y) d\sigma_X^2$$

$$\propto \int_{\Sigma} \left(\frac{1}{\sigma_X^2}\right)^{a*+1} \exp\left\{-\left(\frac{1}{\sigma_X^2}\right)\left[b^* + \frac{1}{2}(\beta-\mu^*)^T(\beta-\mu^*)\right]\right\} d\sigma_X^2$$

$$\propto \left[1 + \frac{(\beta-\mu^*)^T\Sigma^{*-1}(\beta-\mu^*)}{\nu^*}\right]^{-\frac{\nu^*+1}{2}}, \nu^* = 2a^*, \Sigma^* = \left(\frac{b^*}{a^*}\right)V^*$$

Clearly, this is a multivariate *t*-distribution with degrees of freedom ν^*.

Therefore,

$$p(\beta|y) \sim MVSt_{\nu^*} = \frac{\Gamma\left(\frac{\nu^*+1}{2}\right)}{\Gamma\left(\frac{\nu^*}{2}\right)\pi^{\frac{1}{2}}|\nu^*\Sigma^*|^{\frac{1}{2}}}\left[1 + \frac{(\beta-\mu^*)^T\Sigma^{*-1}(\beta-\mu^*)}{\nu^*}\right]^{-\frac{\nu^*+1}{2}}$$

C.2 Details of Simultaneous Equation Model Estimation for Blood Loss and Length of Stay for Hysterectomies

Any model where productivity as well as quality related process variables are included, endogeneity related to simultaneity or reverse causality of the variables is a concern. *'Cut_to_Close'* duration of a surgical procedure may lead to higher blood loss due to longer exposure. On the other hand higher blood loss due to other extraneous clinical conditions like hemorrhage, high blood glucose levels, or use of blood thinning drugs causing hypo-coagulability, may increase the duration of the surgical procedure since extra precaution and care is required for such conditions in patient. Thus, *'Cut_to_Close'* duration and *'Blood_Loss_ML'* are simultaneous variables. A simple linear model would be inappropriate to estimate the model parameters with consistency. We use a system of equations approach for estimating the models.

For ease of modelling, let us assume the following notations for the variables of interest.

$$y: Cut\_to\_Close$$

$$z: Blood\_Loss\_ML$$

$$x: \{Doc, Expr, Comorbidities\}$$

$$T: Uterine\_Weight$$

$$Q: BMI$$

The primary estimation models are:

$$\boldsymbol{y_i = \beta_0 + \beta_1 z_i + \sum_{j=2}^{7} \beta_j x_{ij} + \beta_8 T_i + \epsilon_i} \ldots\ldots [\mathbf{C.1}]$$

$$\boldsymbol{z_i = \gamma_0 + \gamma_1 y_i + \sum_{j=2}^{7} \gamma_j x_{ij} + \gamma_8 Q_i + \eta_i} \ldots\ldots [\mathbf{C.2}]$$

Including both *'Uterine_Weight'* and *'BMI'* in both the estimation equations would lead to identification issues due to unavailability of free parameters in the system. *'Uterine_Weight'* is more correlated with the duration of a surgery than the blood loss. On the other hand *'BMI'* is more correlated with blood loss than with the duration of a surgical procedure.

The reduced form equations of the system proposed is given by:

$$\boldsymbol{y_i = A_0 + \sum_{j=2}^{7} A_j x_{ij} + A_8 Q_i + A_9 T_i + u_i} \quad \ldots\ldots [\mathbf{C.3}]$$

$$\boldsymbol{z_i = B_0 + \sum_{j=2}^{7} B_j x_{ij} + B_8 T_i + B_9 Q_i + v_i} \quad \ldots\ldots [\mathbf{C.4}]$$

Where,

$$\beta_1 = \frac{A_8}{B_9}$$

$$\gamma_1 = \frac{B_8}{A_9}$$

$$\beta_j = A_j - \beta_1 B_j; \forall j \in \{0,2,3,4,5,6,7\}$$

$$\gamma_j = B_j - \gamma_1 A_j; \forall j \in \{0,2,3,4,5,6,7\}$$

$$\beta_8 = (1 - \beta_1\gamma_1)B_8$$

$$\gamma_8 = (1 - \beta_1\gamma_1)A_8$$

$$\epsilon_i = u_i - \beta_1 v_i$$

$$\eta_i = v_i - \gamma_1 u_i$$

$$\sigma(\beta) = \sqrt{\frac{\epsilon^t \epsilon}{n}(\Phi^t \Phi)^{-1}},\ \Phi = \begin{bmatrix} 1\ z\ x\ T \end{bmatrix}$$

$$\sigma(\gamma) = \sqrt{\frac{\eta^t \eta}{n}(\Psi^t \Psi)^{-1}},\ \Psi = \begin{bmatrix} 1\ y\ x\ Q \end{bmatrix}$$

C.2.1 Derivation of the Reduced Forms [C.3] and [C.4]:

By replacing [C.4] in [C.3] we get,

$$y_i = \beta_0 + \beta_1\left(\gamma_0 + \gamma_1 y_i + \sum_{j=2}^{7}\gamma_j x_{ij} + \gamma_8 Q_i + \eta_i\right) + \sum_{j=2}^{7}\beta_j x_{ij} + \beta_8 T_i + \epsilon_i$$

$$= (\beta_0 + \beta_1\gamma_0) + \beta_1\gamma_1 y_i + \sum_{j=2}^{7}\gamma_j\beta_1 x_{ij} + \beta_1\gamma_8 Q_i + \sum_{j=2}^{7}\beta_j x_{ij} + \beta_8 T_i + \epsilon_i + \beta_1\eta_i$$

$$or, (1 - \beta_1\gamma_1)y_i$$

$$= (\beta_0 + \beta_1\gamma_0) + \sum_{j=2}^{7}(\beta_j + \gamma_j\beta_1)x_{ij} + \beta_1\gamma_8 Q_i + \beta_8 T_i + (\beta_1\eta_i + \epsilon_i)$$

$$or, y_i = \left(\frac{\beta_0 + \beta_1\gamma_0}{1 - \beta_1\gamma_1}\right) + \sum_{j=2}^{7}\left(\frac{\beta_j + \gamma_j\beta_1}{1 - \beta_1\gamma_1}\right)x_{ij} + \left(\frac{\beta_1\gamma_8}{1 - \beta_1\gamma_1}\right)Q_i + \left(\frac{\beta_8}{1 - \beta_1\gamma_1}\right)T_i + \left(\frac{\beta_1\eta_i + \epsilon_i}{1 - \beta_1\gamma_1}\right)$$

$$y_i = A_0 + \sum_{j=2}^{7}A_j x_{ij} + A_8 Q_i + A_9 T_i + u_i$$

$$where, A_0 = \frac{\beta_0 + \beta_1\gamma_0}{1 - \beta_1\gamma_1}; A_j = \frac{\beta_j + \gamma_j\beta_1}{1 - \beta_1\gamma_1}, \forall j \in \{2,..,7\}; A_8 = \frac{\beta_1\gamma_8}{1 - \beta_1\gamma_1};$$

$$A_9 = \frac{\beta_8}{1 - \beta_1\gamma_1}; u_i = \frac{\beta_1\eta_i + \epsilon_i}{1 - \beta_1\gamma_1}$$

$$Similarly, z_i = B_0 + \sum_{j=2}^{7}B_j x_{ij} + B_8 T_i + B_9 Q_i + v_i$$

$$where, B_0 = \frac{\gamma_0 + \gamma_1\beta_0}{1 - \gamma_1\beta_1}; B_j = \frac{\gamma_j + \beta_j\gamma_1}{1 - \gamma_1\beta_1}, \forall j \in \{2,..,7\}; B_8 = \frac{\gamma_1\beta_8}{1 - \gamma_1\beta_1};$$

$$B_9 = \frac{\gamma_8}{1 - \gamma_1\beta_1}; v_i = \frac{\gamma_1\epsilon_i + \eta_i}{1 - \gamma_1\beta_1}$$

Therefore,

$$\frac{A_8}{B_9} = \left(\frac{\beta_1\gamma_8}{1 - \beta_1\gamma_1}\right) \div \left(\frac{\gamma_8}{1 - \gamma_1\beta_1}\right) = \beta_1$$

Similarly,

$$\gamma_1 = \frac{B_8}{A_9}.$$

For the other parameters, we get the following equations,

$$\begin{bmatrix} 1 & \beta_1 \\ \gamma_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_j \\ \gamma_j \end{bmatrix} = (1 - \beta_1\gamma_1) \begin{bmatrix} A_j \\ B_j \end{bmatrix}, \forall j \in \{0,2,3, \dots ,7\}$$

$$\begin{bmatrix} \beta_j \\ \gamma_j \end{bmatrix} = (1 - \beta_1\gamma_1) \begin{bmatrix} 1 & \beta_1 \\ \gamma_1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} A_j \\ B_j \end{bmatrix}, \forall j \in \{0,2,3, \dots ,7\}$$

$$= (1 - \beta_1\gamma_1) \times \frac{1}{\mathbf{det}\left(\begin{bmatrix} 1 & \beta_1 \\ \gamma_1 & 1 \end{bmatrix} \right)} \begin{bmatrix} 1 & -\beta_1 \\ -\gamma_1 & 1 \end{bmatrix} \begin{bmatrix} A_j \\ B_j \end{bmatrix}$$

$$= (1 - \beta_1\gamma_1) \times \frac{1}{(1 - \beta_1\gamma_1)} \begin{bmatrix} 1 & -\beta_1 \\ -\gamma_1 & 1 \end{bmatrix} \begin{bmatrix} A_j \\ B_j \end{bmatrix}$$

$$\begin{bmatrix} \beta_j \\ \gamma_j \end{bmatrix} = \begin{bmatrix} A_j - \beta_1 B_j \\ B_j - \gamma_1 A_j \end{bmatrix}, \forall j \in \{0,2,3, \dots ,7\}$$

Therefore,

$$\beta_j = A_j - \beta_1 B_j;\ \forall j \in \{0,2,3, \dots ,7\}$$

$$\gamma_j = B_j - \gamma_1 A_j;\ \forall j \in \{0,2,3, \dots ,7\}$$

Similarly, it can be shown that,

$$\epsilon_i = u_i - \beta_1 v_i$$

$$\eta_i = v_i - \gamma_1 u_i$$

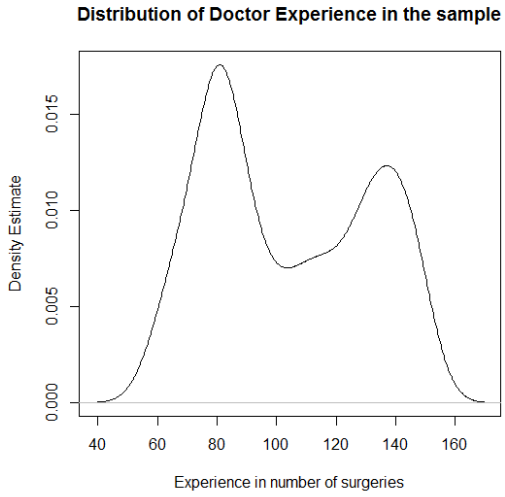
By solving for the parameters in the original system of equations from the estimated parameters of the reduced form, we get a consistent estimate of the parameters in the original system of equation. Since, the error structure of the original system is retrievable from the reduced form equations, we can estimate the standard error and hence the significance levels of the parameters in the original equation. Hence, now we need to have a consistent estimate of the reduced form equation. To get a consistent estimate of reduced form equation we would need to account for any selection bias, if present, in the sample.

C.2.2 Selection Issues: Surgeon – Patient Selection Bias

Selection bias in the sample may arise from the way patients are assigned to surgeons for the purpose of the surgical procedures. If patients with higher levels of criticality are assigned to

surgeons with higher levels of experience on the surgical procedure then that may lead to significant selection bias in the sample which may lead to inconsistent estimates of the reduced form equations unless the selection bias is adequately accounted for. We check for selection bias at the level of experience class of surgeons as well at the level of individual surgeons. Figure [C.1] shows the distribution of experience of surgeons in the subsample.

Figure C.1 Distribution of surgeon's experience on da Vinci robot



Clearly, the distribution of surgeons is bimodal which shows that there are two classes of surgeons, one having high level of experience and another having low level of experience. Let's assume that a patient i can be assigned to a surgeon with high experience $Expr^H \sim N(\mu_H, \sigma_H^2)$ or to a surgeon with low level of experience $Expr^L \sim N(\mu_L, \sigma_L^2)$. Also, let π be the proportion of high experience and low experience surgeons in the sample. Then, the classification of surgeons into high and low experience class is achieved by minimizing the Gaussian mixture likelihood,

$$\boldsymbol{max}_{\pi,\mu_H,\sigma_H,\mu_L,\sigma_L} \prod_{i=1}^{n} \{\boldsymbol{\pi N}(\boldsymbol{Expr}_i|\boldsymbol{\mu_H}, \boldsymbol{\sigma_H^2}) + (\mathbf{1}-\boldsymbol{\pi})\boldsymbol{N}(\boldsymbol{Expr}_i|\boldsymbol{\mu_L}, \boldsymbol{\sigma_L^2})\} \ldots [\mathbf{C.5}]$$

This formulation makes it possible to classify a surgeon as low experience during the beginning of the sampling period and high experience at a later date during the sampling period. However, in our sample this was not the case. The estimated values of means of the component distribution of the mixture Gaussian are $\mu_H = 136$ and $\mu_L = 77$. Two of the surgeons were classified as high experience surgeons and rest four of the surgeons were classified as low

experience surgeons. We estimated a logit model to estimate any selection effect if present:

$$\log\left(\frac{P(Doc\_Expr\_Class = H)}{1 - P(Doc\_Expr\_Class = H)}\right) = \beta_0 + \beta_1 Uterine\_Weight + \beta_2 BMI + \beta_3 (Comorbidities = Y) \dots\dots [C.6]$$

For the individual surgeon level selection bias estimate we estimated a multinomial logit model with each of the surgeon codes as choice classes. The estimation results are shown in Table [C.1].

Table C.1 Surgeon's Selection Effect Estimate for Patient Assignments for Surgeries

Binoimial Logit Model				
Response: Experience_Class (High: 1, Low:0)				
	Estimate	**Std. Error**	**z Value**	**P Value**
Intercept	0.0089	0.9365	0.010	0.992
BMI	0.0247	0.0327	0.755	0.450
Uterine_Weight	-0.0019	0.0012	-1.510	0.131
Comorbidities Y	-0.4082	0.5096	-0.801	0.423
Multinomial Logit Model				
Response: Doctor_Code				
	Estimate	**Std. Error**	**t Value**	**P Value**
GYNMD03:Intercept	-0.2836	1.4846	-0.1910	0.8490
GYNMD05:Intercept	-1.7172	2.2875	-0.7507	0.4552
GYNMD06:Intercept	-3.8464	2.6597	-1.4462	0.1522
GYNMD10:Intercept	-4.1579	2.6707	-1.5569	0.1236
GYNMD11:Intercept	-1.1508	1.4878	-0.7735	0.4416
GYNMD03:BMI	0.0138	0.0514	0.2685	0.7890
GYNMD05:BMI	0.0022	0.0832	0.0264	0.9790
GYNMD06:BMI	0.0884	0.0563	1.5702	0.1205
GYNMD10:BMI	**0.1156**	**0.0539**	**2.1447**	**0.0352***
GYNMD11:BMI	0.0437	0.0523	0.8356	0.4060
GYNMD03:Uterine_Weight	-0.0037	0.0029	-1.2759	0.2059
GYNMD05:Uterine_Weight	0.0017	0.0019	0.8947	0.3738
GYNMD06:Uterine_Weight	0.0024	0.0015	1.6000	0.1137
GYNMD10:Uterine_Weight	-0.0016	0.0023	-0.6957	0.4887
GYNMD11:Uterine_Weight	-0.0015	0.0029	-0.5172	0.6065
GYNMD03:Comorbidities Y	-0.3722	0.7835	-0.4750	0.6361
GYNMD05:Comorbidities Y	-0.7353	1.0081	-0.7294	0.4680
GYNMD06:Comorbidities Y	-0.5298	0.8408	-0.6301	0.5305
GYNMD10:Comorbidities Y	-0.0871	0.9683	-0.0900	0.9285
GYNMD11:Comorbidities Y	-0.8836	0.7362	-1.2002	0.2338
n	153			

From both the estimation models we see that selection effect is negligible. In the multinomial logit model for individual surgeon level selection, *GYNMD04:BMI* slope is significant. From our discussions with the hospital surgical staff and management, we understand that patient to surgeon as well as nurse assignment is done randomly based on availability of surgeon and nurse. Also, hysterectomy being a fairly standard procedure with low levels of mortality or adverse outcome risk, availability based scheduling of patients to surgeon and surgical teams is more

efficient from a organizational point of view. Our selection model results reconfirm this assertion from the hospital surgical stuff. Hence, we believe a standard generalized linear model estimate for the reduced form equations [C.3] and [C.4] would lead to consistent estimates of the reduced form parameters from which we can consistently retrieve the original system.

C.3 Robustness Checks for Varying Patient Population

The study and the sample is from a hospital in a specific region in the United States. We appreciate that this is a critical limitation of the study from the point of view of generalizability of the study. The questions that need to be addressed for asserting some amount of generalizability of the study would be, how robust are the findings of the study to a different set of patient profiles? How robust are the findings to a different patient to surgeon assignment scheme? While actual data would allow us to answer these questions with higher confidence, in the absence of real data from other hospitals, we used a simulation approach. We confine our simulation study to hysterectomy patients and for only the Cut_to_Close surgical procedure duration outcome. The simulation approach has two fundamental assumptions. First, any patient sample would be drawn from the general population of patients. Second, the general distribution of the outcome variable Cut_to_Close with Gaussian perturbation, representing the scenario that the same set of surgeons as in the sample are to operate on varying patient samples randomly drawn from the general population of patients. The second assumption seems reasonable, since in general the surgeons in the sample have substantial experience on the surgical robot for the specific procedure of hysterectomy patients.

To estimate the hysterectomy patients' population level parameters we did a thorough literature survey of several recent clinical studies which report patient characteristics of hysterectomy patients in the United States. We confined our survey to the papers which study hysterectomy patients from the United States only to maintain homogeneity of population level parameters since some of the parameters may be influenced by the healthcare system in specific economies. Wright et. al. (2013) study data from more than 200 hospitals in the United States and report distribution of hysterectomy patients on four dimensions *Age, BMI, Comorbidities* (count of number of comorbidities) and *Uterine_Weight.* Similarly, Soliman et. al (2011) study comparative effectiveness of robotic versus manual laparoscopy of gynecological oncology patients and report several of the patient parameters. Martino et. al (2014) and Jacome, Herbert and Christian (2013) have reported patient parameters on several of these dimensions from separate samples in their

respective study on hysterectomy patients. The patient parameters reported in most of these studies including our sub-sample collected for this study have very similar distributions. A triangulation of these studies and several others lead to a general population level patient parameters as shown in Table [C.2].

Table C.2 Population Level patient Profiles

Pupulation Level patient parameters	
Age Distribution	
<40	26.2%
40-44	22.4%
45-49	23.9%
50-54	12.6%
56-60	5.8%
>60	9.2%
Comorbidity Score	
0	27.0%
1	20.5%
>=2	52.6%
BMI	
Mean	27.6
SD	5.7
Range	18.3-47.4
Uterine Weight	
Mean	169.3
SD	124.6
Range	29.0-1081.0

We used the following simulated data generation schemes for the simulation study based on the patient population level parameters.

$$Age_i \sim Beta(\alpha, \beta), Age\ is\ scaled\ between\ 0\ and\ 1.$$

From Table [B.2], we can see that we are given a few quantile points $\{(x_i,\ p_i): i \in (1, \ldots n)\}$ on the entire age distribution of patients and the corresponding empirical cumulative probability distribution computed as the fraction of the sample falling below the point x_i. To estimate the distributional parameters (α, β) we minimize the quadratic loss function:

$$\min_{\{\alpha,\beta\}} \sum_{i=1}^{n} \left(p_i - \int_0^x \boldsymbol{Beta}(x|\alpha,\beta) \right)^2 \quad \ldots [\mathbf{C.7}]$$

Similarly we sample the BMI using the following distribution:

$$BMI_i \sim Beta(\gamma, \delta), BMI\ is\ scaled\ between\ 0\ and\ 1.$$

In the reported studies, BMI has been shown to be independent of age. However, the other two parameters *Comorbidities* and *Uterine_Weight* have been shown to be jointly dependent on *Age* and *BMI.* Hence, we adopt the following distributional scheme.

$$Comorbidity|Age, BMI \sim Multinomial(p_i: i \in \{0,1,2\})$$

$$Severity|Comorbidity, Age, BMI \sim Multinomial(p_i: i \in \{1,2,3,4\})$$

$$s.t, the\ qunatiles\ of\ the\ distributions\ match\ the\ reported\ quantiles.$$

We used a Monte Carlo Markov Chain based Gibbs Sampling approach to generate the sample of patients at the population level.

To generate the response variable *Cut_to_Close* time we used the following log-normal distributional scheme, (a random variable follows a log-normal distribution when the log of the random variable follows a normal distribution, which is the case with the *Cut_to_Close* time in our original estimate as shown in Table [B.2].

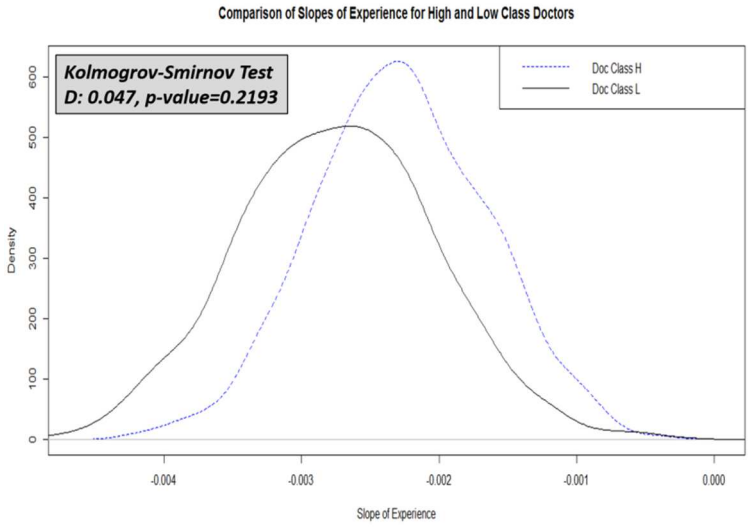
$$\boldsymbol{Cut\ to\ Close} \sim x \sim \frac{1}{x\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}; \quad \ldots\ldots [\mathbf{C.8}]$$

$$\mu = \mu_0 + \mu_1 Expr + \mu_2 Uterine\ Weight + \mu_3 BMI + \mu_4 Comorbidities + \delta\eta; \eta \sim Norm(0,1)$$

The parameters $\{\mu_0, \mu_1, \mu_2, \mu_3, \mu_4, \delta\}$ are estimated from the sample of actual data on hysterectomy patients in the hospital using a GLM estimation δ being the standard error of the GLM estimate and η are sampled randomly from a standard normal distribution independent of the data. To match the simulated response and the simulated patient characteristics we used a probability weighted sampling scheme, where the probability of sampling a patient profile corresponding to a response sample is inversely proportional to the absolute difference in the patient profile used for the response sample as shown in equation [B.8] and the patient profile in the simulated patient sample. This probability weighted matching sampling schemes ensures that the matched sample resembles a somewhat realistic scenario and is not a complete random matching. In Table [C.3] we report the GLMM estimation results of one such matched sample. As we can see from the most of the main results hold out in this simulated example. We ran the simulation for over 1000 matched samples and in Figure [C.2]. As we can see from the Kolmogorov-Smirnov test

of difference in distribution of surgeon's learning between high experience class and low experience class, the difference is not statistically significant thus supporting the primary claim that adoption of robotic surgical procedures leads to standardization of outcome.

Figure C.2 Simulation example distribution of slope of surgeon's learning



Next, we focused on the second issue of selection bias based on experience. For this, instead of matching the simulated response sample and the simulated patient sample based on the patient characteristics in both the samples, we matched the samples using the samples based on surgeon's experience and patient characteristics in the two simulated samples respectively. So for a sample point in the simulated response sample i, we first computed the cumulative probability at $Expr_i$ from the empirical distribution of surgeon's experience. Similarly, we computed the joint cumulative probability of the patient characteristics form the patient sample from using the sampling distributions for simulating the patient profiles discussed earlier. The matching sample probability weights have been programmed to be inversely proportional to the absolute difference between the two cumulative probabilities. This sampling scheme ensures that there is significant selection bias in assigning patients with higher severity to surgeons with higher experience. Table [C.3] shows one sample output of the Heckman selection model we ran on the simulated data-set.

Table C.3 Simulation Results for Surgeon to Patient Selection Strategy

	No Selection Strategy		Moderate Selection Strategy		High Selection Strategy	
Selection Quation: Tobit 5						
Response: Experience Class (High: 1, Low: 0)						
	Estimate	**p_value**	**Estimate**	**p_value**	**Estimate**	**p_value**
(Intercept)	-0.2915	0.552	-21.051	6.38e-11***	-9.7493	<2.00e-16***
xsAge	-0.0358	0.88	5.672	3.28e-05***	2.7138	1.08E-06
xsBMI	0.2525	0.512	6.839	1.39e-06***	2.1138	0.00104
xsComorb	0.2039	0.287	10.963	4.57e-14***	5.6629	<2.00e-16***
xsSeverity	0.2823	0.545	15.945	7.06e-09***	7.4993	3.42E-13
Outcome Equation 1: Experience Class: High						
Response: log(Cut.to.Close)						
	Estimate	**p_value**	**Estimate**	**p_value**	**Estimate**	**p_value**
(Intercept)	9.2191	0.0027**	4.9591	<2e-16***	4.742	<2.00e-16***
x1Expr	-0.0441	<2.0e-16***	-0.0027	8.43e-05***	-0.0027	<2.00e-16***
x1Doc_Diverse	0.0831	0.9827	0.4359	0.263	0.4993	0.203
xmat1Age	-0.6547	0.451	-0.0669	0.504	0.0586	0.571
xmat1BMI	0.9062	0.4111	0.164	0.236	-0.0549	0.602
xmat1Comorb	1.3953	0.0015**	0.0748	0.386	-0.0619	0.629
xmat1Severity	0.5834	0.6593	0.1532	0.275	0.0348	0.813
Outcome Equation 2: Experience Class: Low						
Response: log(Cut.to.Close)						
	Estimate	**p_value**	**Estimate**	**p_value**	**Estimate**	**p_value**
(Intercept)	8.3927	0.0002***	4.5909	<2e-16***	4.4859	<2.00e-16***
x2Expr	-0.0411	0.0219*	-0.0022	1.23e-07***	-0.0022	<2.00e-16***
x2Doc_Diverse	0.0881	0.5639	0.1585	0.7682	-0.0097	0.9825
xmat2Age	0.051	0.7693	-0.0229	0.8089	0.0034	0.9708
xmat2BMI	0.8186	0.0041**	0.2359	0.0645.	0.1687	0.0875
xmat2Comorb	0.4103	0.0098**	0.0336	0.8348	0.2029	0.1602
xmat2Severity	0.4646	0.1791	0.0598	0.7369	0.2134	0.1966

As we can see from the simulation results, the basic hypothesis set hold out even with different patient profiles and with selection effect. This leads to some degree of generalizability of the results within the limitations of the simulation assumptions.

C.3.1 Time varying patient Profiles: Simulation

Another issue that can impact the results is if the patients' profiles change over time due to change in demographic parameters or change in health parameters of general population. Can some of the results related to surgeon learning be explained by a favorable (benign) change in patient profiles? In other words, if the patient profiles had become more favorable over the years during the observation period, then also we would observe improvements in patient level outcomes of surgical duration. Can this effect, if it were to happen, change the results related to surgeon's learning and surgeon's heterogeneity of outcome?

To investigate these questions, we discussed these issues with the hospital team and we

understand that this is not the case. From experience, the hospital team asserts that there has been no substantial change in patient profile over the observation period. However, to check the robustness of our findings we extended the simulation approach described earlier to generate a patient profile which becomes more favorable (less severity and criticality) over time. We do not explain the exact algorithmic details here for sake of space. The results of this simulation study are shown in Table [C.4]. From the results of the table we see that the main results again hold out. In case there is a very steep decrease in patient severity over time, the learning effect decreases marginally and surgeon heterogeneity also marginally becomes statistically significant. However, such a shift in patient profile is a virtual impossibility in real life scenarios.

Table C.4 Simulation Results for Time Varying Patient Profiles

	No Variation		Moderate Variation		High Variation	
	Posterior Mean	p-Value	Posterior Mean	p-Value	Posterior Mean	p-Value
(Intercept)	3.26E+01	0.64	4.41E+01	0.516	3.29E+01	0.652
Expr	-7.27E-03	0.004**	-8.17E-03	<0.001***	-7.58E-03	0.004**
I(Expr^2)	1.13E-05	<0.001***	1.07E-05	<0.001***	1.11E-05	<0.001***
Doc_Diverse	-2.74E+01	0.688	-3.88E+01	0.576	-2.77E+01	0.698
DocGYNMD03	2.65E+00	0.734	3.87E+00	0.604	2.65E+00	0.722
DocGYNMD04	-1.89E+00	0.698	-2.72E+00	0.546	-1.91E+00	0.714
DocGYNMD05	3.48E-01	0.896	8.67E-01	0.78	3.28E-01	0.896
DocGYNMD06	1.42E+00	0.708	2.08E+00	0.556	1.46E+00	0.69
DocGYNMD10	3.59E+00	0.646	4.89E+00	0.524	3.62E+00	0.648
DocGYNMD11	1.45E+00	0.373	1.51E+00	0.042*	1.49E+00	0.004**
Age	2.17E-04	0.83	-7.67E-04	0.538	9.01E-04	0.384
BMI	2.78E-03	0.59	4.71E-03	0.456	2.08E-03	0.704
Comorb	-4.25E-03	0.682	-4.18E-02	0.106	-2.51E-02	0.146
Severity	-5.42E-05	0.424	5.28E-05	0.408	-1.24E-04	0.042*
Expr:DocGYNMD03	1.62E-02	0.086+	1.60E-02	0.028*	1.64E-02	0.024*
Expr:DocGYNMD04	4.04E-03	0.122	5.02E-03	0.092+	4.38E-03	0.092+
Expr:DocGYNMD05	9.47E-03	0.28	8.43E-03	0.354	9.92E-03	0.274
Expr:DocGYNMD06	2.27E-03	0.548	2.12E-03	0.584	2.07E-03	0.575
Expr:DocGYNMD10	-6.45E-03	0.36	-6.87E-03	0.324	-6.56E-03	0.334
Expr:DocGYNMD11	-2.62E-02	0.173	-2.68E-02	0.098	-2.69E-02	0.072
Expr:DocGYNMD12	-3.24E-03	0.162	-3.60E-03	0.622	-3.39E-03	0.623

Notes:
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Iterations = 3001:12991
Thinning Interval = 10
Sample Size = 1000