EVIDENCE-BASED STRATIFICATION METHODOLOGY FOR NON-PROBABILISTIC SAMPLING SURVEYS

A thesis submitted to
Auckland University of Technology (AUT)
in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

Supervisors

Professor Ajit Narayanan

Associate Professor Russel Pears

FEBRUARY 2018

By

Ali H. Gazala

School of Engineering, Computer and Mathematical Sciences

Abstract

There is increasing use of non-probability sampling methods in large-scale surveys due to the costs involved in ensuring that the sample chosen is representative of the population, as is the case with probability sampling. Conventionally, it has been believed that non-probability sampling does not permit precise estimates of how the statistical properties of the sample differ from the statistical properties of the population due to possible biases in the non-probability sample. However, the increasing growth of big data survey data using non-probability sampling methods may provide an opportunity for researchers to use novel methods for quantifying the amount of bias that may exist in different strata so that within each stratum it may be possible to select respondents through probability sampling or random sampling to create pseudo-controlled samples for estimating population parameters.

In this thesis, we use one of the largest survey databases ever collected in healthcare (Improving Practice Questionnaire IPQ for patients visiting their doctor in UK) through convenience sampling to show it is possible to adopt different stratification strategies in conjunction with machine learning techniques to help researchers to decide on the most appropriate stratification method for estimating population parameters from the chosen strata. Such strategies can enrich our knowledge for an evidence-based stratification methodology to reveal similarities and differences in feedback experience among different smaller sub-populations. This research combines standard statistical and machine learning techniques into a systematic stratification methodology to analysis survey data collected through non-probability sampling.

In summary, the traditional statistical problem of how to estimate population parameters from a study that does not use probability sampling is shown in this thesis to be possible through the use of big data and appropriate use of measures and metrics from machine learning as well as standard statistical methods for analysing population parameters. The implication of this thesis are that it will be possible, in the age of big data, to overcome traditional statistical concerns about the quality of data not obtained through traditional probabilistic techniques and that outcomes of statistical analysis using non-probability sampling methods can be as reliable as from probability sampling, provided that a clear methodology is used to quantify bias at various stratification levels.

Table of Contents

Contents

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

_____

Signature of student

List of Figures

List of Tables

List of Abbreviation

Socio-Demographic (SD)

Quality and Outcomes Framework (QOF)

General Practice Patient Survey (GPPS)

General Practice Assessment Questionnaire (GPAQ)

Rasch Model (RM)

Principal Component Analysis (PCA)

expectancy/disconfirmation paradigm (EDP)

General Practitioner (GP)

Improving Practice Questionnaire (IPQ)

Iterative Dichotomiser 3 (ID3)

Classification and Regression Trees (CART)

Design Science Research Process (DSRP)

Analysis of Variance (ANOVA)

Kaiser-Meyer-Olkin (KMO)

Chi-square Automatic Interaction Detector (CHAID)

Kernel Density Estimation (KDE)

One-Level Signal to Noise Ratio (1LSNR)

Two-Level Signal to Noise Ratio (2LSNR)

Average Variance Ratee (AVS)

Average Aggregated Mean Item Variance (AVI)

Average Variance (VR)

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Ajit Narayanan and Assoc. Prof. Russel Pears for the continuous support, motivation, constructive suggestions, and immense knowledge during my PhD study.

I would also like to express my grateful gratitude to CFEP pty Australia, and to Dr Michael Greco for allowing access to the survey dataset used in this thesis. A special thanks also goes to school of engineering, computing and mathematical sciences staff who facilitated my study at AUT.

I would like also to thank my friends and colleagues for their constructive discussions, especially, Ahmad Wedyan. Also, my friend Chamari I. Kithulgoda for her encouragements and suggestions throughout this PhD journey.

Finally, I am heartily grateful to my family and especially my parents, who were always there to support and provide me with resources and the means with which to complete this thesis. Without their precious support it would not be possible to finish my study.

Chapter 1 Introduction

1.1 Introduction

In recent years, the use of large-scale survey data has significantly increased as organisations seek to uncover new knowledge and information from various sources, such as feedback from customers and other users of organisation services. Such data typically comes from questionnaires on which customers, patients and other service users (‘raters’) answer a series of questions (‘items’) by ticking or selecting an option [1], [2]. Questionnaires are often composed of Likert-scale questions, where raters choose one ‘box’ or select one option among a range of options that are typically laid out in some logically ascending or descending order. The number of options can range from only two (e.g. ‘Yes’, ‘No’) to five (e.g. ‘Very poor’, ‘Poor’, ‘Neutral’, ‘Good’, ‘Very good’) or more. When the questions ask for rater attitudes or feelings, or other responses that reflect psychological aspects, the data are regarded as psychometric. Psychometrics is an area of statistics dealing with the theories and methods of psychological measurement.

In the healthcare domain, psychometric feedback data can motivate the development of actions and policies. In healthcare feedback studies, satisfied patients were found to be more positive about their situation, more compliant, and more willing to actively participate in their treatment plans, than dissatisfied patients [3]. Dissatisfied patients also tend to feel anxious and concerned, and have poorer outcomes [4]. At the policy level, understanding the determinants of healthcare satisfaction and obtaining feedback on patient experience can help decision makers to remove potential disparities among different patients or subgroups of survey raters. Such developments have majorly reduced the waiting times for treatment and improved the access, support, and information provided to patients. Therefore, measuring the quality of

healthcare systems through patients' feedback is crucial for defining areas for improvement and monitoring the impact of change [5], [6].

For example, patients' satisfaction surveys in the United Kingdom (UK) comply with the Quality and Outcomes Framework (QOF), a pay-for-performance scheme introduced in 2004 whereby the income of general practices is partly decided by national survey results of patients' experiences [7]. In 2012, the satisfaction questionnaire also became part of doctors' periodic revalidation process introduced by the UK General Medical Council. Several surveys for revalidation and pay-for-performance programmes have been developed in the past decade, including the Improving Practice Questionnaire [8], the General Practice Patient Survey (GPPS) [9], [10], the General Practice Assessment Questionnaire (GPAQ) and the GPAQ-Revalidation GPAQ-R [11]. The items and scales in these surveys capture patients' views on a wide range of care-related aspects, such as accessibility, communication and doctors' interpersonal skills. Therefore, by understanding the priorities and perspectives of patients in different sub-groups, we can develop an equitable health care service responsive to the needs of diverse population groups.

Although the advantages and applications of large scale psychometric survey data have been demonstrated in many areas and domains, including healthcare, business and education, identifying the feedback profiles of small sub-populations of raters remains a challenging task. Moreover, a rigorous statistical analysis would ensure reliable, generalisable findings and potential bias correction. There is also a need for a novel and evidence-based satisfaction theory that accounts for the inherited sociodemographic sampling biases in patients' satisfaction data.

The following sections of this chapter will address the specific limitations and current research gaps in large-scale survey data research.

1.2 Research Background

The previous section introduced the increasingly popular trend of deriving insights and developing actionable domain policies from large-scale psychometric survey data. The analysis of patients' satisfaction with healthcare systems has gathered momentum in recent decades. The vast majority of patient satisfaction theories and concepts have been influenced by empirical satisfaction studies in the business domain. Customer satisfaction measurements and determinants occupy a central position in marketing and business practice. Businesses and service providers have developed pragmatic operational guidelines based on the long history of customer satisfaction studies. With their focus on customer satisfaction, the business domain has sought to understand the needs of sociodemographic sub-groups of their customers. For example, business studies have highlighted gender differences in their quality and satisfaction judgments. However, the patterns of sociodemographic factors detected in business studies are not always consistent. The impacts of age, gender, and education and income levels differ among studies related to service loyalty and repurchasing behaviour.

Healthcare policies are based on the patient-centred model, which considers patients as customers and physicians and hospital staff as service providers. In the patient-centred model, satisfaction questionnaires are increasingly used to highlight the evaluation behaviour of responders and identify the most important determinants of patient satisfaction. With this increasing emphasis on patient feedback, there is an increasing need to ensure the reliability of patients' survey responses as a performance indicator. Patients can describe high levels of satisfaction while describing suboptimal experiences, and their subjective satisfaction varies systematically with certain socio-demographic characteristics such as age, gender, and ethnicity [12]. Several recent studies have established the importance of socio-demographic characteristics on patient-reported experience [9], [10], [13]–[16]. Although doctors' interpersonal skills are among the most important determinant of patients' satisfaction, patients

who are young, belong to ethnic minorities and report poor self-rated health are known to provide less positive feedback. Nonetheless, the effect of socio-demographic characteristics remains contentious. In a GPAQ analysis, the authors of [13] found an association between patients' age and three satisfaction indicators: access, communication with the doctor, and overall satisfaction, while gender was associated only with the "waiting for appointment" outcome. However, a GP patient survey (GPPS) study [14] found that age is a statistically significant predictor of all health aspects, but gender differences were generally insignificant and inconsistent in their trend directions. Clearly, the impacts of sociodemographic factors in the business and healthcare domains are inconsistent among studies.

'High stakes' policies such as financial reward schemas and physicians (ratees) evaluation are increasingly being decided by large-scale patients (raters) feedback data. Therefore, ensuring data reliability rather than questionnaire reliability is essential for making fair comparisons across subjects. The reliability of questionnaires used for data collection can be determined by well-established statistical techniques, which usually quantify the internal consistency (or reliability) among survey items [17]. However, most of the large-scale healthcare survey studies apply a non-probabilistic sampling methodology whereby questionnaires are handed out to raters until the questionnaires are exhausted or time expires, without any attempt to ensure adequate representation of the various sociodemographic groups in the sample. Current techniques such as case-mix adjustment can adjust for differences in patient socio-demographic characteristics that are not controllable by medical practitioners, allowing equal and fair comparison among healthcare providers. The adjustment technique is usually applied on the raters' sociodemographic factors (e.g., age, gender, education and socio-economic status) and on the healthcare provider characteristics (such as surgical or non-surgical). However, a systematic methodology that splits the rater population into smaller subgroups is lacking in the literature. Such a methodology would support current techniques in identifying

sociodemographic impact and correct potential sampling biases, and provide a tool for detecting the reliability of the satisfaction feedback profiles of small rater sub-populations. The methodology is extendible to many other satisfaction survey areas, such as students assessing the quality of their lecturers, customers rating their broadband services and web users rating the content of their visited websites.

1.3 Problem Statement

The previous section introduced the research background on the use of survey data to study the impact of raters' sociodemographic factors on their satisfaction levels. The background highlighted that survey data are increasingly used in high stakes policies and that analysis results must be validated by rigorous statistical techniques. Satisfaction and survey studies typically rely on standard statistical techniques such as regression analysis, which models the relationship between several independent variables and a dependent variable, and Principal Component Analysis PCA, which condenses a number of highly correlated variables into a smaller subset of independent variables that account for most of the variance in the data. The quality of the results and findings in a survey analysis largely depends on the quality of the collected data and the sampling period. Survey studies usually acquire data by different sampling methods, such as cross-sectional collection (selecting a representative sample from a specified population), successive independent collection (repeatedly drawing many random samples from the same population), longitudinal collection (surveying the same sample at multiple time points) and convenience sampling (approaching random participants and recruiting those willing to participate).

From a statistical perspective, data collection by sampling can be probabilistic or non-probabilistic. Probability sampling methods include simple random sampling (allocating a random number to each member of the population and picking sample members by a random

number generator), systematic random sampling (similar to random sampling but with a random start sample followed by every *n*th sample, where *n* is the sample rate), stratified random sampling (which divides the population into subgroups and then simply or systematically chooses one sample from each subgroup) and cluster sampling (which chooses a naturally occurring cluster and samples all members of that cluster; this method is suitable when a comprehensive list of population members is lacking). Probability sampling methods are typically adopted when the data must have a high level of confidence, and when quantifying possible bias and error.

On the other hand, non-probability sampling is useful for achieving specific research objectives and when samples are not required to represent the population as a whole [18]. Convenience non-probability sampling involves participants who are available and willing to partake in the research. Convenience sampling occurs at the most appropriate location for seeking participants' views, such as lecture theatres (if asking students to evaluate their lecturer) or healthcare centres (after patients have consulted their general practitioner), so is cheaper and more easily administered than other sampling methods. Convenience sampling is also appropriate when researchers cannot access patients' data due to legal constraints. For these reasons, non-probability sampling methodology has become the preferred method for obtaining direct, immediate in-context feedback on the quality of patients' experience. When convenience sampling is adopted in more than one healthcare practice, such as patient feedback at multiple primary health centres, the data are unbalanced (different numbers of raters per practice (ratee)), fully nested (all raters assess a single ratee) and uncrossed (each rater receives only one rating). In the second instance, there no opportunity for raters to rate another ratee; in the third instance, there is no opportunity for a subsequent rating to check the reliability of the first rating. Therefore, the reliability of such data must be demonstrated by a suitable statistical analysis.

Despite its easy practical implementation and popularity, statisticians generally concur that convenience sampling cannot precisely estimate population attitudes, because non-probability sampling is inherently biased. Supporting this position, various studies based on convenience sampling give inconsistent results. Although large-scale studies using the convenience sampling methodology improve the generalizability and estimation of population parameters, the subjects' feedback can be highly overlapped and may not reflect the view of smaller subpopulations. In healthcare satisfaction studies, the patients' (raters) characteristics are usually analysed by considering the entire sample for every socio-demograpic attribute, i.e., by analysing all patients by their gender, then by age-group, then continuing through the remaining attributes. This method allows the multiple analysis of individual patients, but the proportions of patients in different sub-groups may lead to inconsistent findings.

Meanwhile, probability sampling methods are hampered by generalizability problems, because accounting for a large number of sociodemographic factors is prohibitively costly. Moreover, recruiting enough subjects in certain sociodemographic subgroups, such as ethnic minorities, is often difficult or impossible. To avoid these problems, many survey analysts adopt cross-sectional approaches without generalising to the population at large. Instead, their conclusions are restricted to the population measured at that time. The outcomes of such cross-sectional surveys are thought to be generalisable after multiple repeats, and useful for predictive or trend purposes. However, this assumption may underlie major survey failures. For example, when recent surveys in the UK and USA were wrongly generalised, the conclusions falsely predicted the outcomes of Brexit and the US presidential elections [19], [20].

Identifying and handling the potential biases caused by large amounts of missing or incomplete survey responses is also important [21]. For surveys in particular, the missing value problem is significant. Sometimes, imputation of missing values or replacement of the missing values with the grand means are the only options to removing a rater's data from the analysis (case

deletion). Neither of these solutions is appropriate in a 'high stakes' survey where, for instance, satisfaction (e.g. student satisfaction, patient satisfaction) is a factor in promotion or funding. The use of imputation techniques or removal of significant proportions of data can also have financial and legal consequences. Moreover, for technical and ethical reasons, imputation must be appropriately applied and the imputed values must be explicitly flagged in datasets to satisfy the data protection requirements. For these and other reasons, the extraction of robust models based on psychometric data remains problematic generally and specifically in healthcare satisfaction surveys.

As big data are increasingly surveyed by non-probability sampling methods, researchers can develop novel methods for quantifying and correcting the sampling bias in satisfaction studies. The current age of big data will inevitably enable the collection and aggregation of large amounts of raters' feedback data in different domains. In one suggested technique, the patients' feedback and their sociodemographic factors are analysed by a stratified approach. The stratification process computes the performance scores separately for different strata or groupings of raters, such as patients grouped by common characteristics(s). Each healthcare unit then acquires multiple performance scores (one for each stratum) rather than one overall performance score [22].

The concept of dividing the search space into subsets of homogeneous subgroups is well established in the machine-learning domain. Applying a divide-and-conquer approach, the learning process splits the search space into smaller subsets while building a set of knowledge and learning rules. Many of the supervised learning algorithms search the space of possible stratification branches under the guidance of entropy and information gain. The information gain measures the reduction in uncertainty after splitting the dataset based on a selected independent variable [23]. The splitting process continuously selects the independent variable that returns the most homogenous subset. The resulting models are optimized to predict the

future unknown data based on knowledge learned from quantifying the relationships between the independent and target variables. These methods aim to stratify the dataset attributes rather than the raters or subjects.

This thesis investigates the opportunities and challenges of using measures and metrics from machine learning, as well as standard statistical methods, for analysing population parameters from big survey data. Noting that both probability and non-probability sampling techniques have their own advantages and disadvantages, this thesis investigates whether stratification analysis can identify the impact of sociodemographic factors, and provide a set of rules and guidelines for solving satisfaction survey problems with sampling biases and missing-value imputation.

1.4 Research Objective

The objective of this thesis is to investigate the feasibility of combining standard statistical and machine learning techniques into a systematic stratification methodology. A stratification analysis divides the raters ('population') into mutually exclusive and collectively exhaustive subpopulations for individual and comparative analysis. A quantifiable stratification analysis can reveal whether the performance outcomes depend on one or more specific sociodemographic factors. It also identifies and facilitates the reduction of sociodemographic disparities. According to recent studies on patient satisfaction, each healthcare unit can collect multiple performance scores (one for each stratum) rather than one overall performance score [22]. However, no clear guidelines for implementing a systematic stratification approach were identified in the literature.

For the most part, survey data have been confined to small-scale and individual research studies, because a questionnaire will probably change over the long-term. An organisation may prefer to design its own questionnaire for a specific purpose (e.g., to reduce customer turnover by

seeking reasons for customers' leaving, to seek patients' experience of healthcare provided, and to seek students' opinions on the quality of their lecturers), reducing the opportunity to warehouse such data for long-term use. However, large-scale survey data have become more common in recent years, as large organisations accumulate yearly data for auditing, quality control, and enhancement of delivery or provision to their 'customers'. Such data are collected by district or regional health boards, national telecom providers, and educational institutes (schools, colleges and universities) [24]–[27]. As these data become available, researchers gain the opportunity to develop evidence-based concepts of raters' satisfaction by identifying the similarities and differences in experience/satisfaction between different rater groups. In other words, using the knowledge accumulated by numerous convenience non-probability sampling techniques, researchers can estimate the population parameters after quantifying and adjusting the potential sociodemographic biases. This thesis investigates whether combining the machine learning technique with standard statistical techniques can realise a systematic stratification methodology that creates non-homogeneous and mutually exclusive patient subgroups. The stratification methodology could also help to quantify the bias in different sociodemographic factors. Thereby, we could create pseudo-controlled samples for accurate and precise representation of the population parameters. A quantifiable stratification analysis can reveal whether the performance outcomes depend on one or more specific sociodemographic factors. It also identifies and facilitates the reduction of sociodemographic disparities.

The formalisation of the research questions and the thesis methodology are presented in Chapter Three.

1.5 Structure of Thesis

The above sections introduced the research problems and objectives of this thesis. The remainder of this thesis is structured as follows:

- Chapter Two - Literature Review: This chapter explores the roots of patient satisfaction studies, and explains how a significant part of this domain was influenced by 'customer satisfaction' studies in business environments. It highlights the pattern of inconsistent findings on the impacts of sociodemographic factors in the healthcare and business domains. The chapter also reviews supervised machine learning techniques and applications related to stratification and missing-value imputation methods. It concludes that analysing sociodemographic patterns by a stratified approach can help to explain the inconsistencies in non-probability sampling studies.
- Chapter Three – Theoretical Framework: This chapter details the research objectives and the research methodology framework. The chapter addresses the three research questions investigated in this thesis.
- Chapter Four – Exploratory Analysis: This chapter introduces the large-scale study dataset used in this thesis, and implements an exploratory analysis at the zero stratification level (i.e., including all patients and sociodemographic factors). The amount of variance in each sociodemographic factor is assessed by different standard statistical techniques. The potential negative impact of patients' missing answers on survey feedbacks is also discussed.
- Chapter Five – Missing Values Analysis: This chapter investigates the impact of patients' missing answers and how different patterns of missing values are associated with certain sociodemographic subgroups. This chapter highlights the feasibility of a stratification-sensitive data imputation methodology.

- Chapter Six – Stratification Process: In this chapter, we implement the stratification analysis using machine learning and standard statistical techniques.
- Chapter Seven – This chapter investigates the reliability of stratification analysis by highlighting the key differences between questionnaire reliability and data reliability. The results provide guidance for selecting the best stratification path by maximizing or minimizing the variance criteria.
- Chapter Eight - Estimating Population Parameters using Pseudo-Controlled Samples: Using the results of Chapter Six, this chapter creates pseudo-controlled samples for estimating the population parameters from non-probability sampling datasets.
- Chapter Nine – This chapter summarizes the stratification methodology and findings of the thesis, and suggests ideas for future research.

Chapter 2 Literature Review

2.1 Introduction

Studies of patients' satisfaction with healthcare systems began emerging in the second half of the twentieth century, and have gathered momentum in recent decades. The increased interest in patient satisfaction research is partially attributable to the "patient-centred" healthcare policies developed by many national medical councils. In healthcare studies, satisfied patients report feeling more positive about their situation than dissatisfied patients, are more compliant and cooperative, and more likely to participate actively in their treatment regimens [3]. Therefore, assessing the quality of healthcare system from patients' feedback is considered as a vital exercise for service improvement and quality assurance [28]. In recent years, patients' satisfaction feedback has also become part of doctors' periodic revalidation process [5]. Such developments have largely reduced the waiting times for treatment and have improved the access, support, and information provided to patients.

Much patient satisfaction research is dedicated to discovering the important determinants of satisfaction and explaining the evaluation behaviour of patients. As the data accumulated, patterns of patient preferences started to emerge, and systematic variations appeared among certain socio-demographic variables. For example, higher satisfaction level is associated with older age, lower education level, and long-term registeration with the current doctor [29]. Nonetheless, the reported effects of sociodemograpic variables such as gender, age and race on satisfaction level are conflicting and inconsistent. Sociodemographic variables are directly related to satisfaction in some studies, and inversely related or unrelated in other studies. Therefore, by identifying the priorities and personal preferences of patients in different

sociodemographic and socioeconomic groups, we could establish an equitable healthcare system responsive to the needs of diverse population groups [30].

The next section (2.2) explores how patient satisfaction studies were rooted in theoretical and empirical studies of customer satisfaction within job and business environments. It highlights theories of customer satisfaction and how sociodemographic variables are analysed in different business domains. Section 2.3 reviews the multi-dimensional concept of patient satisfaction, and lists some of its main theories and determinants published in the literature. Section 2.4 describes the design and implementation of patient-reporting instruments in the healthcare sector. These instruments are considered as the primary tools for obtaining patient feedback after a consultation or visit to a healthcare provider. The section also highlights some of the most common problems of obtaining questionnaire data by convenience sampling. Finally, section 2.5 discusses the importance of analysing sociodemograpic variables in healthcare studies, and highlights some of the analysis methods that may confuse the results of socio-demographic profiling.

2.2 Customer Satisfaction and Socio-Demographic Effects

2.2.1 Conceptualization of Customer Satisfaction

Patient satisfaction and its measurements are becoming increasingly important to public policy makers, healthcare managers, practitioners and users. Patient satisfaction is one goal of health care delivery and customer satisfaction is the necessary outcome. However, patient satisfaction theories are rooted in theoretical and empirical studies of satisfaction in job and business environments [31]. Today, the concept of consumer satisfaction is widely embraced by manufacturers, retailers and service providers, who benefit from policies that meet customers' needs and preferences. Therefore, the conceptualization and measurement of customer satisfaction has been extensively researched. This section explores the roots of customer satisfaction theories and highlights the effect of sociodemographic variables on satisfaction levels in the business environment.

Customer satisfaction measurements and determinants occupy a central position in marketing and business practice. In business environments, the customer-satisfaction concept assesses consumer activities such as consumption levels, numbers of repeat purchases, attitude changes, and brand loyalty [32]. However, for practical purposes, the psychometric statement of the marketing concept must be converted to operational rules and guidelines. To this end, many researchers have investigated and developed consumer evaluation and satisfaction measures [33]. Since the early 1970s, customer satisfaction has become a legitimate field of inquiry and research into meeting marketing and business needs [34]. Since its inception, the volume of consumer satisfaction research has been impressive. Researchers have proposed numerous theoretical structures for examining the antecedents of satisfaction and developing meaningful measures of the satisfaction construct [35]–[37]. A notable outcome of consumer satisfaction research is the confirmation/disconfirmation paradigm, which hypothesises that consumer

satisfaction is consequent to a comparison process [38]. In the confirmation/disconfirmation framework, consumers compare their perceptions of a product performance with a set of standards (e.g., expectations or some other performance norm). An individual's expectations are defined as: (1) *confirmed* when the product performs as expected, (2) *negatively disconfirmed* when the product performance is lower than expected, and (3) *positively disconfirmed* when the product performance is higher than expected. Therefore, a dissatisfactory output occurs when a consumer's expectations are negatively disconfirmed. The confirmation/disconfirmation paradigm is composed of four main constructs: performance, expectations, disconfirmation, and satisfaction [33].

The expectations component reflects the consumers' anticipations when forming opinions of a product's expected performance. Miller [39] highlighted four types of expectations: ideal, expected, minimum tolerable, and desirable, whereas Day [40] distinguished among product or service expectations, expectations of costs and efforts in obtaining benefits, and social cost–benefit expectations. The performance component provides a comparison standard for assessing disconfirmation [41]. The disconfirmation component emerges from variances between prior expectation and the actual performance. Oliver [42] stressed the importance of separately measuring disconfirmation and expectation. He maintains that each construct exerts an independent, additive effect on satisfaction. Finally, the satisfaction component is the outcome of the purchase and use activities. This component results from consumers' evaluation of the rewards and costs of the purchase, relative to the anticipated consequences [42].

Although the confirmation/disconfirmation paradigm is widely accepted, there are many other frameworks of customer satisfaction, including the expectancy/disconfirmation paradigm (EDP), the equity theory, the attribution theory, the value/percept theory, the dissonance theory, the contrast theory, the comparison level theory, the importance/performance theory, and the evaluative congruity theory [38]. In general, these theories suggest that consumer satisfaction

is a relative concept, and is always judged relative to a standard [43]. Some of these theories (e.g., the value/percept theory) posit that consumers judge satisfaction relative to values and desires; in other theories, the standard is the predictive expectation (the EDP), or the experience-based norms (comparison level theory). Equity theory considers that satisfaction results from comparing consumer inputs and outputs. The EDP, which has become the most widely applied assessment method of consumer satisfaction and dissatisfaction [43], derives from theories that consider satisfaction as resulting from the discrepancy between expectations and perceived performance [44].

In summary, the business and marketing literature provides a wide range of conceptual and theoretical frameworks for assessing customer satisfaction and its determinants. Some studies reported a significant correlation between consumer satisfaction and expectation; others found that consumer satisfaction depends not on expectation, but on perceived performance. The next section discusses the implementation of customer satisfaction theories in the marketing and business environments. The effects of different sociodemographic variables on satisfaction level are also discussed.

2.2.2 Socio-Demographic Effects on Customer Satisfaction

Businesses and service providers have translated the various customer satisfaction theories into pragmatic operational guidelines. Customer satisfaction directly affects the results and profitability of a business. The authors of [45]–[47] reported a positive relationship between customer satisfaction and customer loyalty, and between customer satisfaction and positive word-of-mouth [48]. To enhance customer satisfaction, businesses and service providers must understand the needs of their customers' sub-groups. Individuals in different sociodemographic groups hold particular expectations and focus on particular performance dimensions. For example, gender researchers have highlighted different quality and satisfaction judgments by

male and female customers. Mattila et al. [49] investigated gender effects in service encounters, and found that the server's emotional display differently affects men and women. In particular, men are more outcome focussed, and negatively affective displays do not influence their satisfaction with a successful service encounter. On the other hand, the satisfaction of female customers is decreased by negative emotional displays, even if the service encounter succeeds. Female customers thus seem more focussed on the service process, whereas men place more emphasis on the service outcome.

According to the marketing and business literature, identifying the relationships between customer satisfaction and sociodemographic variables may not be a straightforward process. The authors of [50], [51] studied the relationship between satisfaction level and loyalty (repurchase intention) in the German automobile industry. They investigated the influences of gender, age, and income level on customers' re-ordering decisions. They found that female customers prioritised personal interaction over product satisfaction when deciding whether to repurchase from the same dealer. Meanwhile, young and high-income customers relied more on the information provided by sales personnel than their satisfaction with the product. In addition, female, older (> 60 years), less educated, and married couples with no children were more tolerant (i.e., more likely to repurchase at the same satisfaction level) than other groups.

Other studies of customer loyalty in the service industries highlighted the inconsistency of assessing loyalty by sociodemographic variables [52], [53]. Age, education and income (but not gender) were moderately associated with service loyalty (i.e. repurchase intention and loyalty behaviour). In particular, middle and senior age groups (35–54 and >55 years) displayed significantly more loyal behaviour than their younger counterparts (18–24 and 25–34 years).

Brady et al. [54] also reported inconsistencies in the effects of sociodemographic variables on customers' perceptions [54]. They examined the influence of customer satisfaction on shopping

behaviours and intentions. By analysing several models, they explained the service evaluation process in various settings (fast food outlets, grocery stores, airlines and physicians) in five countries (Australia, China/Hong Kong, Morocco, the Netherlands, and the US). The influence of service quality, satisfaction and value on the behavioural intentions of shopping were best fitted by the "comprehensive" model. However, these results were not consistent across the different sub-samples of the study. Quality, service value and satisfaction majorly affected shopping behaviour only in the United States and Australian samples [55]. In the Netherlands, the effect of satisfaction was insignificant, and in Hong Kong and Morocco, the effect of service quality was insignificant. Service value alone significantly affected the behavioural intentions in all samples.

Sharma et al. [56] extended the previous study to explore how two customer demographics (age and gender) moderate the relationships between service quality, sacrifice, value, satisfaction, and behavioural intentions. They conducted a mall-intercept survey in major shopping areas in different parts of Hong Kong during the March–April period of 2009. Shopping intentions were significantly moderated by gender and age in the service evaluation process. However, unlike Brady et al. [54], they also found that service quality influences the shopping behavioural intensions in the Hong Kong sample. After careful inspection, the authors found different proportions of sociodemographic variables in the two studies. The Hong Kong sample in [54] was much younger (all < 30 years old) and more biased towards females (63.5%) than that in [56]. In contrast, Sharma et al.'s [56] sample contained a roughly equal distribution of younger and older participants (47% >30 years old; 53% < 30 years), of whom 58% were female. The authors concluded that variances in some sociodemographic variables may account for the inconsistent results across the samples.

Many studies within the marketing and businesses literature have highlighted the importance of sociodemographic variables on customer satisfaction level. However, many studies present

inconsistent findings among studies, as explained in the above examples. Identifying whether these differences arise from the type of business or service, value-consciousness, or some other factors requires further studies. Noting the inconsistent effects of sociodemographic variables in the customer satisfaction literature, other studies have evaluated products in the post-purchase period. In these studies, product perforamce was consistently influenced by expectation (or some other comparison standard) and confirmation/disconfirmation.

2.3 History of Patient Satisfaction Theories

Throughout the past few decades, increasing numbers of national medical councils have developed "patient-centred" healthcare policies. This demand was influenced by the proliferation of interest among academic researchers, policy makers and health service professionals in integrating patients' perspectives when formulating, monitoring and improving their health policies and services [57]. Earlier studies indicated that satisfied patients are more positive about their situation than dissatisfied patients, and more likely to comply with, cooperate with, and actively participate in their treatment regimens [31]. On the other hand, frustrated or stressed patients whose basic expectations are not being met may not respond fully to therapeutic interventions [58]. Therefore, measuring the quality of healthcare system through patients' feedback reveals weak services, areas of substandard quality, and the impacts of change [28]. Such developments have largely reduced the waiting times for treatment and improved the access, support, and information provided to patients.

Patient satisfaction is researched not only for its clinical and psychological benefits, but also for political reasons. Assessing public satisfaction with the overall extent and quality of services enables higher-level healthcare policy planning [3]. For instance, Kenagy et al. [59] shows that improving the service quality significantly reduces the costs of care. They highlighted that the dynamics of poor service often involve wasted effort, repetition, and

misuse of skilled employees. As another example, William et al. [60] investigated the relationship between patient satisfaction and hospital readmission. They concluded that patients reporting higher overall patient satisfaction with their clinical cares and discharge planning have lower 30-day risk-standardized hospital readmission rates. Therefore, governments and national medical councils have increasingly developed patient-centred policies focussed on respect, choice, empowerment, patient involvement in the health policy, access and support, and information provision.

Despite the accumulated evidence of the importance of patient satisfaction, the literature lacks a well-defined framework that outline the patient-satisfaction concept. This is partially due to the different definitions and reference contexts adopted by various studies, at both the hospital and health system levels [3]. Another factor is the multidimensional and subjective nature of the patient satisfaction concept, which is usually affected by the needs, desires and expectations of individuals. This section reviews the different models and theoretical frameworks that conceptualize the patient satisfaction concept.

2.3.1 Conceptualization of Patient Satisfaction

Despite the growing number of healthcare studies, attempts to conceptualize the construct of ‘patient satisfaction’ are rare. The lack of any clear definition of ‘patient satisfaction’ was identified by Locker and Dunt in 1978 [61]. They remarked that neither the researchers who employ it nor the respondents who respond to it can agree on what ‘patient satisfaction’ actually describes. The lack of attention to the meaning of the ‘Patient Satisfaction’ construct has been the greatest flaw in patient satisfaction research [62]. Nonetheless, a widely accepted definition is the reaction of a health care recipient to salient aspects of the context, process, and result of their service experience [63]. Over the past twenty years, the healthcare service environment has shifted towards a more consumerist ethos. In this mode, patients are treated as customers

while physicians and hospital are considered as service providers. Overall consumer satisfaction is the desired outcome of the healthcare process [1]. As mentioned previously, theories of patient satisfaction have been largely influenced by theoretical and empirical studies of satisfaction within the job and business environments [31]. Many of the updated definitions in healthcare and patient satisfaction theories, which proliferated during the 1980s, were restatements of theories first published in the business literature [64]. Gill et al [65] summarizes the literature into five main groups of theories:

1. Discrepancy and transgression theories [66] advocate that dissatisfaction arises when the healthcare orientations and care provider conditions differ. When the orientations and conditions are congruent, patients are satisfied; otherwise, patients are dissatisfied.
2. Expectancy/value theory by Linder-Pelz [31] postulates that satisfaction is mediated by care-related personal beliefs and values, and by prior expectations of care. The study highlighted the important relationship between expectations and variance in satisfaction ratings, and offered an operational definition of patient satisfaction as "positive evaluations of distinct dimensions of healthcare". Pascoe [59] developed the Linder-Pelz model, which considers how expectations influence satisfaction, and created a psychological model with six factors: cognitive and affective perception formation, multidimensional construct, dynamic process, attitudinal response, iterative process, and amelioration by individual differences.
3. Determinants and components theory by Ware et al [67] propounds that patient satisfaction depends on patients' subjective responses to their care, which are mediated by their personal preferences and expectations.
4. Multiple-models theory by Fitzpatrick et al [64] argues that expectations are socially mediated, reflecting the health goals of the patient and the extent to which illness and healthcare violate the patient's personal sense of self.

5. Healthcare quality theory by Donabedian [68] proposes that satisfaction is the principal outcome of the interpersonal process of care. He claimed that satisfaction or dissatisfaction expresses the patient's judgement on the quality of all care aspects, but particularly the quality of the interpersonal component of care.

Patient satisfaction and its determinants is often explained by the expectations/disconfirmation paradigm. As explained in the previous section, this paradigm considers whether the perceived performance falls below, equals or exceeds the prior expectations. In the healthcare environment, this paradigm is viewed as an attitude or feeling determined by the patient's cognitive belief, or their perceptions and affective evaluations of health care attributes [31]. The expectations/disconfirmation framework has been supported in a number of patient satisfaction studies [31], [61], [64], [69]–[74], but other studies have reported conceptual difficulties with this framework.

Brody et al. [75] reported that patients' satisfaction levels are only indirectly affected by their expectations, but directly relate to the physician's efforts to meet the patient's personal needs for information, control, support and advice regarding stressful situations. Other authors were more cautious in proposing a relationship between satisfaction and expectation. For example, Linder et al. [31] reported that expectations account for only 8% of the variance in satisfaction, and suggested that patients' satisfaction was more determined by their background beliefs than their perceptions of the care received. Although many authors have acknowledged a relationship between patients' expectation and their satisfaction levels, this relationship appears to be complex and indecisive [31], [59], [67], [76].

2.3.2 Factors of Patient Satisfaction

Besides the conceptual theories, various studies have identified the important factors that constitute satisfaction. These factors were highlighted as determinants of patient satisfaction in healthcare studies published throughout the last several decades. The authors of [67], [77], [78] classified satisfaction factors after a content analysis of the items included in published patient satisfaction questionnaires and patient responses to open-ended questions, which were posed to identify satisfaction. According to another meta-analysis in the patient satisfaction literature [31] and summarized by [79], there is a growing general acceptance that (irrespective of healthcare context) satisfaction can be measured along the following 10 dimensions:

1. Accessibility/convenience,
2. Availability of resources,
3. Continuity of care,
4. Outcomes of care,
5. Financial Arrangement,
6. Humaneness and interpersonal aspects of care,
7. Information gathering,
8. Information giving,
9. Pleasantness of surroundings,
10. Competence

The theories and the ten dimensions highlighted above must provide a reliable indicator of patient satisfaction. Patients sometimes report high levels of satisfaction while also describing suboptimal experiences [80]. Moreover, patients' personal satisfaction levels can depend on their socio-demographic variables such as age, gender, and ethnicity [30], [81]. Whether such differences arise from variations in expectations, differences in services provided to patients

with different backgrounds, or differences in the ways that patients express their experiences, is unclear [82]. Therefore, healthcare analysis studies are trending towards measuring patients' experiences rather than their satisfaction [80]. By measuring patients' experiences, researchers can remove the effects of value judgments and expectations. This idea assumes that expressed experiences are less influenced by subjective expectations than reported satisfaction levels [13]. However, whether reports of patients' experiences are also systematically associated with sociodemographic variables is not clear [83]. Figure 2-1 displays the patient satisfaction factors reported in [84].

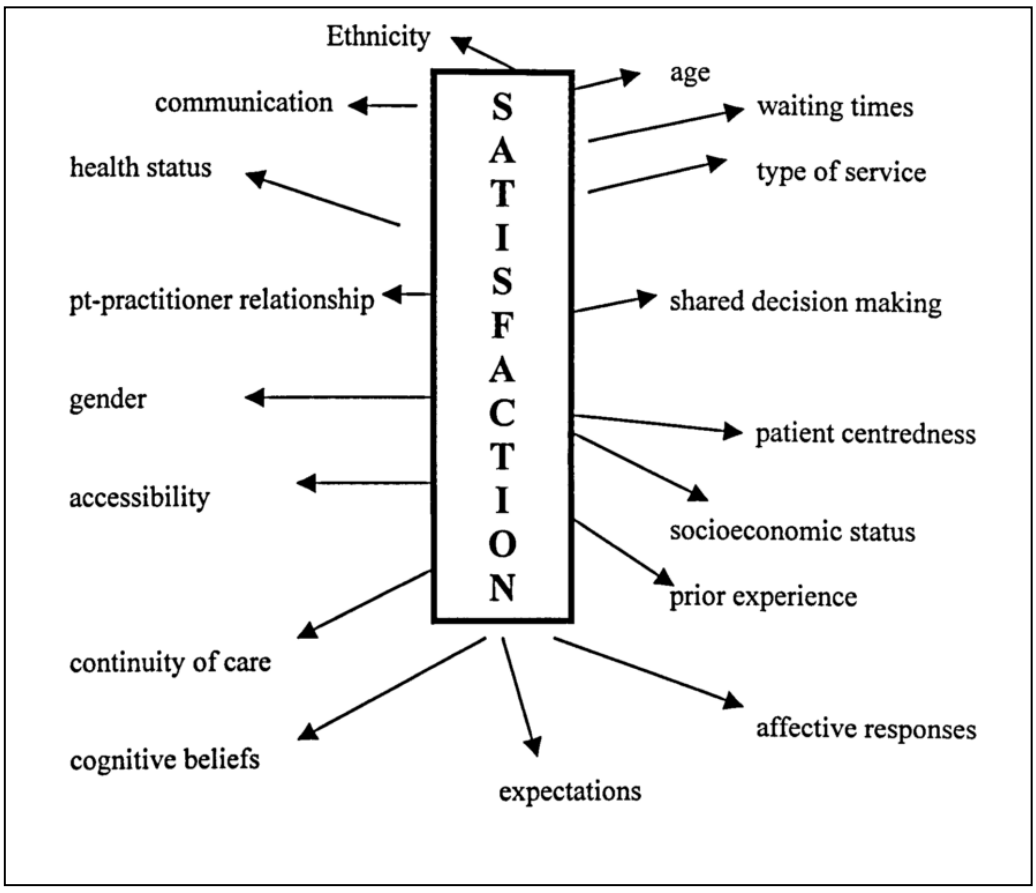

Figure 2-1: Patient satisfaction factors reported in [84]

2.3.3 Limitations and Problems of Patient Satisfaction

Collecting patient's feedback through satisfaction surveys is now well established in healthcare practices, with increasing evidence of positive impact on medical and clinical outcomes [85]. One challenge faced by healthcare researchers is that patients generally report high levels of patient satisfaction, with very few expressing dissatisfaction [61], [69]. Possible reasons for this tendency are genuine satisfaction with their practices, non-confidence in their ability to evaluate their practices, and unwillingness to criticise their practices [79]. A meta-analysis of more than 200 studies reported average satisfaction levels of 76% [86], whereas Fitzpatrick [87] identified satisfaction levels of 80% or higher for any given question in respondents' questionnaires.

Researchers have proposed different explanations for the pattern of skewed Likert scales in satisfaction questionnaires. Some studies have showed that raters tend to agree with almost any statement, regardless of its content, and report greater satisfaction than they actually feel. Patients may also consider positive comments as more adequate, a phenomenon known as 'social desirability response bias' [88]. Moreover, some patients fear that by criticising a service on which they depend, they risk unfavourable treatment in the future [89]. The acquiescence bias problem, also known as the 'tendency to agree with statements of opinion regardless of content', is highlighted in all satisfaction research [88].

Another group of researchers have highlighted that the wording of questions and the scales type can affect the degree of satisfaction and dissatisfaction expressed by patients. When scale questionnaire items are used, attitudinal response scales (e.g., from 'very satisfied' to 'very dissatisfied' or from 'strongly agree' to 'strongly disagree') appear to be more subject to acquiescence response bias than evaluative response scales ('poor' to 'excellent') [90]. However, despite acquiescence bias, patients can express high satisfaction levels while

describing suboptimal experiences [80]. In other words, patients may report negative experiences as subtle differences in their responses, while their overall response remains within the upper scales. Therefore, patient dissatisfaction should be inferred from statistical techniques that are sensitive to 'relative' rather than 'absolute' performance [79].

2.4 Patient Satisfaction Surveys

As highlighted in the previous section, evaluating raters' feedback in healthcare systems is crucial for defining areas needing improvement and for monitoring the impact of change. Although "satisfaction" still lacks a tangible definition in the literature and is difficult to measure, the concept continues to be widely used. Patient feedback is primarily collected by questionnaires that patients complete after a consultation or a visit to the healthcare provider. These questionnaires are designed to quantify patients' experiences of different aspects of their clinical services, such as booking an appointment, continuity of treatment, and communication with physicians and staff. The feedback reported on the satisfaction questionnaires provides a service quality indicator for the personal development of healthcare professionals. Therefore, many national medical councils and healthcare employers currently recommend or require patient feedback as part of an ongoing personal development program between healthcare professionals and their mentors [11]. In the United Kingdom, for example, patients' satisfaction surveys are used within the Quality and Outcomes Framework (QOF), a pay-for-performance scheme introduced in 2004. In this scheme, national survey results of patients' experiences detemine part of a GP's income [91]. In 2012, the satisfaction questionnaire introduced by the UK General Medical Council became part of doctors' periodic revalidation process [92].

Satisfaction questionnaires are designed to measure patients' perceptions of different dimensions of healthcare services. In survey research, a dimension (also called a construct) is the abstract idea, underlying theme, or subject matter to be measured through survey questions

[93]. Some dimensions (such as political party affiliation) are relatively simple and can be measured by asking one or a few questions, while other constructs (such as confidence in a physician) are more complex, requiring a battery of questions to fully operationalise the dimension to suit the research needs [94]. Many studies have attempted to identify the underlying dimensions of patient satisfaction [95], [96]. These attributes, which are usually suggested by a panel of experts, provide an evaluation lexicon for assessing healthcare services, and guidance for instrument selection [97]. Wong et al. [28] reviewed the processes of identifying patient-satisfaction attributes, and summarized the results in six dimensions and 15 sub-dimensions (Table 2-1).

Table 2-1: Dimensions of patient satisfaction attributes, reported by [28]

Dimension	Sub-dimension	Definition
Access	First contact accessibility	The availability of care (including advice and support) required by the patient or client from the provider of choice within a time frame appropriate to the urgency of the problem [97].
	Accommodation	The ease with which resources accommodating patients or clients (appointment systems, hours of operation, walk-in facilities, telephone services) can be accommodated by the patient or client [97].
	Economic accessibility	The extent to which required or recommended cares are impeded by direct or indirect costs.
Interpersonal communication	General communication	The ability of the provider to address patient or client concerns, and to explain health and health care issues [97], [98].
	Respectfulness	The ability of the primary care organization and practitioners to treat users with dignity, provide adequate privacy, and meet users' other expectations regarding respect [97], [98].
	Shared decision-making	The extent to which practitioners involve patients or clients in their treatment decisions [98].
	Whole-person Care	The extent to which providers address the physical, emotional and social aspects of a patient's or client's health in both individual and community contexts [97].
Continuity and Coordination	Relational continuity	The development of a therapeutic relationship between a patient or client and one or more identified providers spanning separate health care episodes, and the delivery of care that meets the patient's or client's biopsychosocial needs [97].
	Information continuity	The extent to which information provides appropriate care to the patient or client.
	Coordination	The provision and organization of various health services and information that meets a patient's or client's health needs, including services available from other community health service providers.
	Team functioning	The ability of primary health care providers to work effectively as a collaborative team to manage and deliver quality care to patients or clients.
Comprehensiveness of services	Services provided	The direct or indirect provision of a full range of services that meet the healthcare needs of patients or clients. Services includes health promotion, prevention, diagnosis and treatment of common conditions, referral to other clinicians, management of chronic conditions, rehabilitation, palliative care and (in some models) social services [97].
	Health Promotion and primary prevention	Empowers patients or clients to better control and improve their own health [99] Primary prevention aims to prevent the initial occurrence of a disorder [73].
Trust		An expectation that one person's behaviour will benefit the other. This dimension allows for risks; for example, if patients or clients trust their physician, they are more likely to divulge personal information [100].
Patient reported impacts of care	Patient activation	The ability or willingness of the patient or client to engage in health behaviours that will maintain or improve their health status [101], [102].
	Patient safety	Patients' or clients' reports of their medication errors (whether they have been given or taken the wrong drug or dose), incorrect medical or laboratory reports, and communication with the provider when prescribed medications are not taken or have side effects.
	Confidence in the PHC system	The perception that a healthcare provider will deliver safe and technically competent care to the patients or clients, encouraging decision making by the patients or clients [103].

As mentioned above, satisfaction questionnaires have become the main tool for measuring patients' feedback on healthcare services. When designing a questionnaire, several rounds of discussions are usually required, administered by systematic group decision-making such as the Delphi method [104]. For example, the length of the survey ultimately depends on the dimensions and sub-dimensions of interest, and the purpose of the survey. Therefore, the dimensions and sub-dimensions of interest, rather than the specific items, should be identified before administering the survey [105]. The design process also involves the wording of questionnaire items and the scales over which the respondents will express their level of satisfaction. Healthcare satisfaction questionnaires are frequently based on the Likert scale, and the response scales can be attitudinal (e.g., 'very satisfied' to 'very dissatisfied', 'strongly agree' to 'strongly disagree') or evaluative ('poor' to 'excellent') [79]. Finally, the patients may be asked to self-complete a questionnaire or be interviewed. Questionnaires can be distributed on site at a computer terminal or as hard copies, or mailed to the homes of potential respondents. Personal interviews may be conducted face-to-face (on site or at home) or by telephone [3].

An early instrument for measuring patient satisfaction, the 'Satisfaction with Physician and Primary Care Scale', was developed in 1970 by Hulka et al. [105]. Since then, numerous instruments have been developed based on ad-hoc patient-satisfaction tools. In the past decade, several surveys have been developed for revalidation and pay-for-performance programs. Among these are the Improving Practice Questionnaire (IPQ) [8], the General Practice Patient Survey (GPPS) [106], the General Practice Assessment Questionnaire (GPAQ) and the GPAQ-Revalidation (GPAQ-R) [11]. The items included in these surveys, and their scales, are designed to capture patients' views over a wide range of care aspects, such as accessibility, communication, and doctors' interpersonal skills.

With the increasing emphasis on patient feedback, questionnaire designers must ensure that the data obtained from patients is reliable [70], especially in high-stake applications such as revalidation, recertification and ongoing accreditation [107]. The social sciences and communication fields apply rigorous and well-established statistical analysis techniques to validate the reliability of their data-collection instruments. One popular statistical procedure is principal component analysis, which verifies whether the raters' response patterns reflect the main construct topics that were intended when designing the survey. Another commonly used measure is the Cronbach's alpha, which measures the internal consistency (or reliability) among survey items [17]. However, the reliability coefficient provided by Cronbach's alpha test must satisfy several sampling assumptions (e.g. the data are balanced, crossed or not fully nested). Studies that violate these assumptions require more complex methods based on analysis of variance, which can generalise the results to different populations of raters and ratees [108]. Most of the recent healthcare satisfaction studies relay on non-probabilistic sampling methodologies: questionnaires are distributed to raters until the copy supply or the survey time runs out. This sampling methodology is cost effective but might not adequately represent the various sociodemographic groups in the sample. More recent studies have highlighted the necessity for a data-reliability measure rather than a questionnaire-reliability measure [109].

Data reliability measures are designed for nonstandard research design problems usually associated with convenience sampling techniques. Many patient satisfaction studies based on convenience sampling have a hierarchical or multilevel data structure. In this model, patients represent the raw-scores "raters" level, while practitioners represent the aggregated-scores "ratees" level. Although this structure allows multilevel analysis and reveals the nested relationship between raters and ratees, it embodies three research design aspects that are problematic from a statistical reliability perspective [110]:

1. Healthcare studies based on convenience sampling techniques may not adequately represent the number of raters and ratees in the multilevel data structure. For instance, medical professionals (ratees) may treat a varying number of patients (raters). Therefore, the numbers of response cases might be insufficient for generating a reliable set of performance scores for each service provider.
2. By its very nature, convenience sampling cannot allow multiple evaluations of a healthcare provider by the same patient (uncrossed responses). Therefore, there is no opportunity for validating the consistency (reliability) of the first rating.
3. In most real applications, patients usually evaluate only one healthcare provider during the study period. In this case, the raters' feedback is fully nested within a unique ratee, meaning that the ratees' performance scores depend on the subjective feedback of raters' responses on a single questionnaire.

Current data reliability measures can tackle convenient sampling research problems and ensure that the minimum numbers of raters and ratees required for deeper drill-down analysis are available [111]. Thus far, no study has attempted a stratification analysis of reliability by measuring the feedback consistency among the different rater subpopulations. Measuring the data reliability is critical for identifying the reliability of subpopulation responses in different stratification levels. If the variance of the subpopulation feedback decreases with increasing stratification level, increasing amounts of noise are being removed as the drill-down proceeds. Conversely, if the variance in the subpopulation feedback increases, the stratification is introducing noise. Chapter Seven of this thesis investigates the use of data reliability measures to validate the reliability of the stratification analysis.

Another ubiquitous feature of surveys and rater feedback studies is the lack of detailed answers in questionnaire forms. Missing data in survey studies are usually classified into three main categories: *noncoverage* (certain elements in the targeted population are missing from the survey sample), *total nonresponse* (the sampled raters do not participate in the study), and *item nonresponse* (the sampled raters provide non-acceptable responses to some or all of the questionnaire items) [112]. Survey analyses based on convenience sampling typically ignore

the noncoverage and total nonresponse biases by including only the raters who are available and welling to participate in the study. However, nonresponse bias remains a critical problem in survey and social science datasets [113].

A large number of missing answers may change the distribution shape of the questionnaire items. If ignored, these missing responses will lead to biased and inconsistent survey estimates. The potential impact largely depends on the non-response rate of the survey items. If the non-response rate of an item is low, the bias in the univariate analyses of that item will be small, so removing the missing answers is reasonable and acceptable. However, most survey data require multivariate analysis, in which the combined effect of many low non-response rates may considerably reduce the dataset size and any information contained in the incomplete answers [114]. Despite the potential impact of missing data, many researchers choose to ignore missing value cases by applying case-wise deletion. In some studies, the handling of missing data values is not reported [115]. The missing data problem can introduce large bias if certain sub-populations are more likely than others to provide incomplete answers.

To compensate the item nonresponse bias, several imputation methods that assign values to missing responses cases are available. Imputation enables all relevant records to be retained without further consideration of the missing data. Current imputation strategies include *mean imputation*, which replaces all missing values with an overall mean, and *class mean imputation*, which divides the dataset into imputation groups based on auxiliary variables [116]. However, imputation strategies do not necessarily reduce the bias from that of the incomplete dataset; in fact, depending on the imputation procedure and the estimate distribution, the imputation can enlarge the bias [114]. Therefore, increasing the availability of large-scale survey data will open opportunities for developing stratification-based imputation strategies that are sensitive to sociodemographic differences. The knowledge gained from such strategies will highlight whether certain sociodemographic sub-populations are more likely than other groups to provide

missing answers, and whether the differences between rater groups with and without missing answers are statistically significant. Chapter Five of this thesis investigates the challenges and opportunities in developing evidence-based imputation strategies based on drill-down stratification analysis.

2.5 Socio-Demographic Analysis

The previous sections highlighted the complexity of the human and patient satisfaction concept. Patient satisfaction depends on numerous factors including lifestyle, past experiences, future expectations and the values of both individuals and society. In the 1970s and 1980s, researchers established the theoretical framework of patient satisfaction, and determined its detriments and dimensions. However, despite the large effort to standardise the satisfaction concept, many healthcare studies have concluded that satisfaction means different things to different people. In other words, patients' subjective satisfaction varies systematically with their socio-demographic characteristics such as age, gender, and ethnicity [13]. The literature on healthcare satisfaction presents contradictory findings on the effects of sociodemographic variables. Some studies report a minor effect of patient demographics on patient satisfaction, while in others, demographics accounted for 90–95% of the variance in satisfaction rates [117]. Despite this discrepancy, the increasing emphasis on healthcare satisfaction has unearthed some consistent relationships between satisfaction levels and sociodemographic variables. Patients from certain age groups, those belonging to ethnic minorities, and those with poor self-rated health, are known to give less positive feedback. Studies also agree that doctors' interpersonal and communication skills are among the most important determinants of patients' satisfaction. This section reviews the effects of sociodemographic variables in the healthcare satisfaction literature, and the current techniques for identifying and correcting the sociodemographic bias.

2.5.1 Sociodemographic Effects

Associations between patients' sociodemographic characteristics and their satisfaction levels have been reported since the early healthcare studies conducted in the 70s and 80s. In 1981, Fox et al. [66] highlighted the contradictory findings on the sociodemographic characteristics of healthcare satisfaction reported in the literature. Variables such as ethnicity, age, gender, and income level can directly relate to satisfaction in one study, inversely relate in another, and be unrelated in a third study. As also mentioned by Fox et al., the situation is so chaotic that some researchers have dismissed sociodemographic variables as reliable predictors of satisfaction. Williams [70] opined that most researchers identify sociodemographic correlates of satisfaction, rather than developing a solid sociopsychological theory of satisfaction.

In the past few decades, the effects of sociodemographic variables on patient satisfaction level have been reported in many studies [13]–[15], [30], [90], [118]–[124]. Herman et al. [118] related satisfaction feedback to the individual characteristics of patients and healthcare providers. They considered three aspects of healthcare services: accessibility, interpersonal relationships, and the information given to patients. For a multi-level analysis, they nested all patients' data within the patients' own GPs. In all three dimensions of patient satisfaction, the study found no statistical evidence that gender and age influence patient satisfaction levels, indicating the presence of cross-level effects for these demographic factors. The satisfaction scores of female patients rating female GPs did not differ significantly from those of female patients rating male GPs. Similarly, male patients were equally satisfied with male and female GPs.

Campbell et al. [30] examined whether assessments of primary care depend on age, gender, socioeconomic, and ethnicity variables. They surveyed 7692 patients using the GPAS instrument, which assesses 13 dimensions of primary care provision. Older patients rated their

care more favourably than younger patients in all GPAS domains. These results are consistent with earlier studies, in which age was positively associated with favourable perception of care [94, 95]. Older patients, who generally interact more frequently with healthcare practitioners than younger patients, have more opportunity for favourably evaluating the services provided. Alternatively, health practitioners may convey to younger patients, or younger patients may interpret, that they are less entitled to primary care services than older patients [30].

Chris et al. [13] conducted a multilevel modelling of satisfaction data at the practice, doctor, and patients levels. They investigated patients' feedback along three dimensions: waiting for an appointment, access to care, and healthcare practitioners' communication skills, in addition to overall satisfaction. Patients expressed high levels of satisfaction while describing suboptimal experiences, and their subjective satisfaction levels systematically varied with age, sex, and ethnicity. For this reason, recent questionnaires focus on patients' experiences, which should be less influenced by subjective expectations than satisfaction. Chris et al. also investigated whether patients' reported experiences are systematically associated with sociodemographic variables. They found that questioning patients' on their experience with a practice more discriminately measured the practice's performance than subjectively questioning their satisfaction with the practice.

Moret et al. [120] investigated the relationship between patient age and satisfaction. Patient age has controversially been described as the most significant sociodemographic variable in healthcare satisfaction results. Moret et al. collected data from two satisfaction studies conducted in 27 short-stay teaching hospitals. A total of 9171 patient responses were collected from self-report questionnaires and telephone interviews. The authors found a nonlinear relationship between age and satisfaction. Patient age was positively and linearly correlated with satisfaction before 65 years, and negatively correlated thereafter. However, the authors

mentioned that the threshold around 65 years, beyond which satisfaction scores for the quality of medical and nursing care decrease, requires verifying in future study.

Elliott et al. [121] examined the gender differences in inpatient experiences and their dependence on care dimensions and other patient characteristics. They compared the experiences of male and female inpatients along 10 healthcare dimensions in multiple linear regression models. They analysed a large dataset of 1,971,632 patients (medical and surgical service lines) discharged from 3,830 hospitals between July 2007 and June 2008. The female patients gave lower positive scores than the male patients, especially for Communication about Medicines, Discharge Information, and Cleanliness of the Hospital Environment. Women reported a more positive experience than men only for Doctor Communication. These differences might reflect differences in both patient expectations and the behaviour of hospital staff.

The above studies highlight the inconsistencies of sociodemographic characteristics in determining patients' satisfaction. Nonetheless, these characteristics are considered important and their effects are being adjusted in healthcare studies for fairer comparisons among healthcare providers. Meanwhile, a well-defined theory that explains the impact of sociodemographic patterns in healthcare satisfaction studies remains lacking. As more large-scale survey data become available, researchers will better understand patients' experience of healthcare provided, and will develop a novel and evidence based satisfaction theory that accounts for the inherited sociodemographic sampling biases in patients' satisfaction data. An evidence based theory based on accumulated large-scale survey data would accept the statistical limitations usually associated with non-probability sampling. Such a theory would increase our knowledge of the similarities and differences in the satisfaction experiences of different rater sub populations. Figure 2-2 schematizes a framework that derives new evidence-based knowledge of patients' satisfaction theory from large scale survey data.

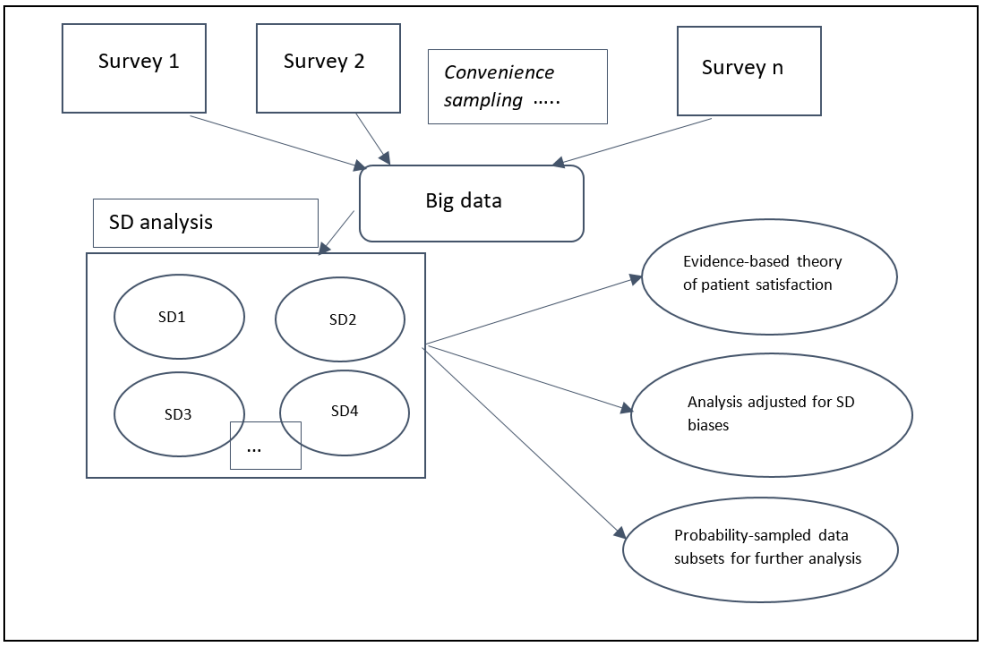

Figure 2-2: Evidence-based theoretical framework of patient satisfaction

When developing an evidence-based patient satisfaction theory, we require statistical techniques that provide useful results from the drill-down data analysis into sociodemographic factors, and avoid potentially ambiguous findings when the results are reported in multiple ways. The knowledge derived from an evidence based theory would improve current estimates of population parameters by quantifying and adjusting the potential sociodemographic biases. A widely used method called *case-mix adjustment* [125] facilitates fair comparisons among healthcare providers by adjusting forpatients' sociodemographic characteristics (age, gender, education and socio-economic status) that are beyond the control of practitioners. The healthcare provider characteristics (surgical or non-surgical) are also usually adjusted.

The debate surrounding case-mix adjustment highlights the advantages and disadvantages of this approach. Advocates argue that physicians caring for disadvantaged patient populations generally perform less well on the satisfaction measures that are commonly used in pay-for-performance schemas [126]–[129]. Therefore, adjusting for economic and demographic

characteristics may improve the physician-level profiles by levelling the playing field and attenuating the effect of potential unintended consequences. Without these adjustments, physicians might exclude patients that will likely worsen their measured quality [128]. Opponents argue that adjusting for patient socio-demographic characteristics might obscure thee differences in quality of care among healthcare providers, and reduce incentives for raising the quality of care of vulnerable populations [128]. However, identifying potential disparities in raters' feedback requires grouping the raters into sociodemographic groups, requiring additional information and analysis. In other words, any risk adjustment practice that involves the sociodemographic characteristics of raters requires additional methods (e.g., stratification) to identify disparities [22].

Alternatively, to case-mix adjustment, one can measure multiple performance scores on different clinical and non-clinical factors. The stratification process computes separate performance scores for different strata or patient groups based on some characteristics(s), meaning that each healthcare unit receives per-stratum scores rather than an overall performance score [22]. This analysis method reveals whether the performance outcomes depend on one or more specific sociodemographic factors, and facilitates the identification and reduction of sociodemographic disparities. Advocates of the stratification approach claim that this method helps to 'unmask' healthcare disparities, because it compares the performance outcomes of groups that have been historically advantaged and historically disadvantaged. Currently, hospitals and healthcare providers are stratified into multiple groups based on the proportion of disadvantaged beneficiaries. Under this policy, healthcare providers are compared in a "like-with-like" manner, and the impacts of unknown social difference that are reasonably beyond the control of the healthcare provider are avoided [130]. Many research studies in the healthcare domain have presented stratification based solutions to identify smaller patients subpopulations and optimize medical care services [131]–[133]. The work of

Parsonnet et al presented a method to stratify open-heart operations into levels of predicted operative mortality using publicly available administrative data. The study identified several risk factors using a dataset of 3,500 consecutive open-heart operations. The suggested stratification methodology can place patients into five groups of increasing risk with some of the high-risk factors include operative complication rates and length of hospital stay. The work of House et al studied the impact of sociodemographic and socioeconomic status on advance care planning for senior patients. The study involved several patients' factors such as education, income level and occupation on the reported health plans. The results suggest that economically advantaged persons engage in end-of-life planning as a two-pronged strategy entailing financial and health-related preparations. These studies demonstrated the feasibility stratification-based solutions to identify smaller patients' subpopulations and facilitate comparison between subgroups.

A healthcare expert panel recently suggested that case-mix adjustment and stratification are not mutually exclusive [22], and that both methods can be combined into a given performance measure using a specific analytic approach. However, combining the two methods is a non-trivial task. There are currently no clear instructions for constructing the strata and developing a top-down or bottom-up stratification process given the potential use of multiple sociodemographic factors. Moreover, most healthcare studies adopt a non-probabilistic sampling methodology that cannot ensure a representative distribution of all sociodemographic factors targeted in the study. Therefore, a systematic set of rules by which researchers can unearth the similarities and differences among large and small sub-populations is needed. For example, when analysing whether sociodemographic disparities exist in a care, stratification can provide instructions for identifying the factors that minimize or maximize the differences among the sub-populations.

2.5.2 Sociodemographic Effects and Patterns Recognition

The previous sections of this thesis presented review and evidence of inconsistent findings about the effect of raters sociodemographic characteristics' within the domains of customers and patients satisfaction studies. Chapter one presented the research objective of this thesis is to integrate traditional statistical analysis and machine learning techniques to highlight hidden patterns about smaller populations subgroups that otherwise would remain unknown. The applications of machine learning are designed to provide automated learning and improvements from historical experience without being explicitly programmed. The learning process relies on observations and historical data in order to identify hidden patterns in data and make better predictions for future unseen examples [134].

The domain of machine learning algorithms and techniques can be generally explained into two different subgroups known as supervised and unsupervised learning [135]. The process of supervised learning is focused on the concept of learning from ex- supervised examples. The learning algorithm is provided with sets of training and test datasets. The goal of the learning algorithm is to develop a set of rules that describe the relationships between the dataset features and the prelabelled output so that it can identify unlabelled examples in the test set with the highest possible accuracy. Unsupervised machine learning is focused on the concept of applying a learning algorithm to automatically identify complex and hidden patterns without a human to provide guidance along the way [136]. All the machine learning techniques applied and discussed in this thesis can be described under the type of supervised learning.

A popular learning approach in supervised machine learning is known as divide-and-conquer where the search space is splits into smaller homogeneous subsets while building a set of knowledge and learning rules. The learning model is usually represented as a top-down decision tree constructed following a greedy search approach through the dataset variables at

each tree node. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain as the test attribute of current node. Information gain is an impurity-based criterion that uses the entropy measure (origin from information theory) as the impurity measure [23][137]. Each leaf in the tree model is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Researchers in the domain of machine learning presented several variations of the top-down decision trees model such as ID3 (Iterative Dichotomiser 3) and C4.5 [23][138], Classification and Regression Trees CART [139].

The decision tree learning approach presented above is mainly designed to predict a class of a future unknown examples. However, when the task of predicting future unknown values involves a range of numeric values, the applied machine learning model is commonly as regression. One of the most widely used regression models is known as linear regression that applies a mathematical formula of a straight line ($y = mx + b$) to estimate a relationship between two variables, while more advanced techniques, such as multiple regression, is used to predict a relationship between multiple variables - for example, is there a correlation between raters satisfaction scale and raters sociodemographic variables. The addition of more variables considerably increases the complexity of the prediction [140]. The stratification analysis presented in this thesis combined several aspects from the classification and regression learning algorithms into a systematic process to stratify convenience sampling raters population into mutually exclusive subpopulations.

2.6 Summary

This chapter overviewed the current and previous literature on healthcare satisfaction studies, focussing on how socio-demographic variables influence satisfaction feedback. The chapter first highlighted that most healthcare satisfaction theories and concepts are extensions of customer satisfaction and loyalty studies in the business field. Understanding the different factors affecting customer loyalty is crucial for business success. Sales and marketing studies have extensively investigated the effect of sociodemographic factors on customers' satisfaction levels. Such research has clarified the satisfaction profiles of customers' sociodemographic sub-populations, which have been translated into pragmatic operational guidelines by businesses and service providers. However, the effects of certain demographic factors are inconsistent among studies, and are sometimes contradictory. This pattern of inconsistent findings is feasibly attributable to the sampling method; convenience sampling may not capture the biases in the different sociodemographic characteristics of the raters. Consequently, some researchers have dismissed sociodemographic variables as a reliable satisfaction indicator, and have instead evaluated the post-purchase behaviour of customers.

The theoretical framework of patient satisfaction, and the challenges in defining this multi-dimensional concept, were then discussed. The 'patient satisfaction' construct has been conceptualized by numerous theories, factors and determinants. According to the literature review, patient satisfaction questionnaires have become the preferred method in healthcare studies, as they provide direct and immediate patient feedback at low cost with low effort. However, the sampling methodology is limited to patients that can conveniently participate in the study. At present, no standardized set of rules has been developed for analytical techniques such as stratification and segmentation in convenient research designs.

The last section of this chapter reviewed the effect of sociodemographic variables on healthcare satisfaction feedback. Originally used for public reporting and quality improvement, satisfaction measures have now entered high-risk accountability applications that demand accuracy and consistency. Therefore, whether patients' responses to satisfaction surveys are influenced by their sociodemographic characteristics, and the effect magnitude of variables such as age and gender, must be discerned. With the increasing focus on these details, the rating-behaviour patterns of certain sociodemographic sub-population have emerged. However, the sociodemographic rating patterns reported by various researchers are inconsistent and contradictory. Such inconsistency when sociodemographic characteristics are factored into rating behaviour appears in both healthcare and business satisfaction studies.

After several decades of research in healthcare services satisfaction, a clear theory that explains the influences of sociodemographic variables on patient satisfaction is still lacking. A novel evidence-based satisfaction theory would help researchers to design suitable data stratification and segmentation strategies that account for the inherited sociodemographic sampling biases in patient satisfaction data. Such a theory would accept the statistical limitations of non-probability sampling in a rigorous drill-down data stratification and reliability analysis. Evidence-based theory aims to reveal the similarities and differences in satisfaction experience among different sub-populations of raters. In the high stratification levels, patients' feedback can be highly overlapped and may obscure the views of smaller subpopulations. At lower stratification levels, the rating patterns of some or all of the performance measures may vary among the subpopulations. Therefore, the importance of individual socio-demographic subgroups must be identified by a systematic stratification strategy. The results of this study provide a first set of rules for handling raters' sociodemographic factors in analyses of conveniently sampled data. The next chapter will introduce the systematic analysis methodology and the research questions addressed in this thesis.

Chapter 3 Formulating a Methodological Framework

3.1 Introduction

The previous chapter in this thesis provided an overview of the current and previous research work in healthcare satisfaction studies with a focus on the effect of socio-demographic variables on satisfaction feedback. The literature review indicated that the majority of healthcare satisfaction theories and concepts were extended from the business domain in the context of customers' satisfaction and loyalty studies. One finding from the literature review is the inconsistent and contradictory reports about the effect of socio-demographic characteristics on the reported satisfaction levels. As patients feedback data become increasingly important for the purpose of physicians evaluation and re-credential purposes, there is a need to ensure performance measures provide fair comparisons across subjects. The literature review chapter also highlighted the lack of an evidence-based theory to explain the impact of raters sociodemographic patterns in healthcare satisfaction studies. The advantage of such theory is to help researchers consider potential problems and sampling biases when analysing patients' satisfaction data, and to increase our knowledge about similarities and differences in satisfaction experience among different raters subpopulations. The development of an evidence-based patients' satisfaction theory would require a stratified analysis approach to allow drill-down data analysis into raters sociodemographic factors. Therefore, the research objective of this thesis is to investigate the feasibility of combining standard statistical and machine learning techniques to generate a systematic stratification methodology. Section 3.2 explains the theoretical methodology framework adapted for this research while section 3.3 highlights the three main research questions investigated in this thesis.

3.2 Methodology Framework

This research investigated different methodology frameworks to implement the stratification analysis including Rasch Model (RM) [141], Generalizability Theory [142], and Design Science Research Process (DSRP) proposed by Peffers et al. [143]. The rasch model is a psychometric model that transforms raw categorical data into abstract, equal interval scale. The Rasch Model was first proposed in the 60s to evaluate education ability tests. In recent years the model has been employed in the evaluation of services; such as customers and patient's satisfaction. The model generates a new scale called Logit that can map subjects and items against each other. In this research work, a preliminary clustering analysis was implemented to investigate if results generated using Rasch analysis would differ from applying standard statistical technique without transferring the data into the logit scale. The results showed a similar clustering output between the two analysis approaches. Therefore, for the analysis and results presented in this thesis, Likert scale values are treated as ordinal and scalar on the assumption that the distances between each Likert point is equal. This assumption allows the use of standard measures of dispersion for comparing sub-group scores given to subjects as well as informative parametric techniques.

Generalisability (G) theory [142] is designed to quantify possible sources of rater and item variance to obtain a measure of the reliability of obtained subject scores, especially when it is not possible to repeat the measurements. The classical G theory (and the associated G reliability coefficient) attempts to quantify this extraneous variance in two ways: by assuming that the same raters are used for all ratees (a balanced and unnested design), and by assuming that it is possible to repeat the measurement for the same ratee (a crossed design). However, many patient satisfaction studies based on convenience sampling have a hierarchical or multilevel data structure. In this model, patients represent the raw-scores "raters" level, while practitioners represent the aggregated-scores "ratees" level. Although this structure allows multilevel

analysis and reveals the nested relationship between raters and ratees, it embodies some research design aspects that are problematic from a statistical reliability perspective. Therefore, A variance-based, two-level signal-to-noise ratio formula was used to determine reliability in such convenience sampling contexts. Further details about this question can be found in chapter seven of this thesis.

The research objective presented in this thesis is to verify the validity of a systematic data stratification methodology. To achieve this objective, the research follows the design science research process (DSRP) proposed by Peffers et al. [143]. The methodology is a wildly applied research framework in information systems. The DSRP framework incorporate the principles, practices and procedures required to carry out and present information systems research. The steps of DSRP are shown in Figure 3-1.

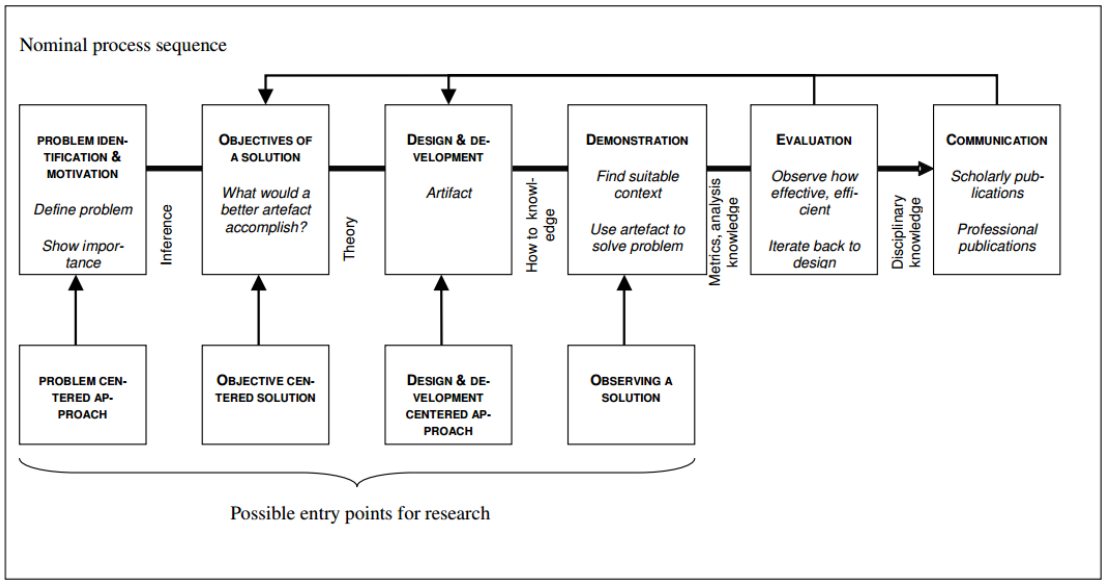

Figure 3-1: Design science research process (DSRP) framework

3.2.1 Problem Identification and Motivation

Previous and recent researches have highlighted the increasing importance of survey data in various areas, including public health, market research and customer satisfaction. The measurement of raters' satisfaction from survey data was initiated by business researchers, who

sought to understand the influencing factors of customer loyalty. Satisfaction surveys have gradually extended to other domains, including student satisfaction surveys that assess the quality of teaching, and patient satisfaction surveys that indicate the quality of primary health care. Traditional survey research usually relies on data collected by probability sampling techniques such as random sampling, stratified random sampling, and cluster sampling. By applying these sampling techniques, researchers can obtain the confidence level of the collected data and the reliability of the survey outcome. However, probability sampling is of limited generalisability and the patterns of smaller sub-populations are obscured. Probabilistic survey data typically collect samples from a specified population, usually based on findings described for that population only. Recent surveys in the UK and USA based on probability sampling yielded incorrect predictions for the outcomes of Brexit and the US presidential elections.

In recent years, non-probability sampling methods have been popularly used in large-scale satisfaction surveys because they are low-cost and easily administered. Convenience non-probability sampling obtains feedback from participants who are available and willing to partake in the research. Convenience sampling methodology collects a relatively large amount of data at suitable locations for seeking the participants' views, such as shopping centres (when evaluating customers' perceptions of their service provider) or healthcare centres (when evaluating patients' experiences with their general practitioner). Convenience sampling also provides a practical means of overcoming the legal and administration limitations that may prevent access to raters' personal data. Despite its popularity, many statisticians believe that biases in non-probability sampling degrade the accuracy of convenience sampling. The literature review identified several examples of inconsistent findings when sociodemographic variables influenced the results of convenience sampling studies. Although large-scale convenience sampling better estimates the population attitudes than small-scale sampling, the

overall feedback can be highly overlapped and may obscure the views of smaller subpopulations.

In the current age of big data, large amounts of rater feedback data can be collected and aggregated in different domains. This thesis investigates the opportunities and challenges of combining big data with measures and metrics from machine learning, along with standard statistical methods for analysing population parameters. Given that probability and non-probability sampling techniques have their own advantages and disadvantages, this thesis investigates whether data stratification analysis under a possible set of rules can reveal the hidden patterns of sociodemographic factors and smaller subpopulations.

3.2.2 Solution Objectives

According to recent studies, stratified data analysis can help us to understand healthcare disparities and identify the feedback profiles of smaller subpopulations. However, how to implement a stratification analysis and construct strata by a top-down or bottom-up stratification process based on multiple sociodemographic factors has not been reported. This thesis develops the first set of rules that will guide researchers towards selecting appropriate sociodemographic factors during the stratification process. The stratification methodology will help to quantify the bias level in different sociodemographic factors. The population parameters can then be accurately and precisely represented by creating pseudo-controlled samples.

3.2.3 Design and Development

The development of a stratification methodology starts by implementing an exploratory analysis at the zeroth stratification level (including the entire patient dataset and all sociodemographic factors). The central tendencies of all survey items are then determined by descriptive statistical analyses, such as the mean and standard deviation. Other statistical

techniques such as dimensionality reduction and analysis of variance then identify the statistically significant differences among the different sub-populations. Stratification analysis divides the entire dataset into subsets of homogenous subpopulations. This concept is widely applied in hierarchical supervised learning techniques such as ID3 and J48 algorithms. When dividing the data into smaller subsets, the learning process follows a divide-and-conquer approach and develops a set of learning rules. Most of the hierarchical learning algorithms guide the search through the space of possible branches using entropy and information-gain measures. The information gain measures the uncertainty reduction after splitting the dataset based on a selected independent variable. The splitting process continues using the independent variable that returns the most homogenous subpopulation.

Hierarchical supervised learning algorithms generate prediction models for future unknown data. The results of these algorithms are usually presented as top-down decision trees. The first node in the tree contains the independent variable with the highest information gain. The splitting process continues until all records belong to the same leaf node or the information gain is below a specified small limit. Hierarchical learning algorithms construct the best tree that predicts the target variable. In survey analyses, the hierarchical learning concept can identify homogenous subpopulations with different sociodemographic factors. By combining the machine learning approach with standard statistical techniques, researchers can elucidate previously unknown patterns of the sociodemographic factors. The hierarchical learning approach is applicable to both ordinal and categorical data.

Survey analysts may need a mechanism that identifies homogenous subpopulations from numerical variables. The supervised learning approach can model the relationship between a dependent variable and a set of independent sociodemographic variables. Linear regression analysis detects the change in the mean value of the dependent variable with a one-unit change in each of the sociodemographic variables. A large (small) regression coefficient, whether

positive or negative, indicates a large (small) difference among the sub-populations in a top-down stratification approach.

The intercept of a regression model with categorical independent variables must be within the context of the required task. For example, when seeking the variable that maximises or minimises the difference among the sub-populations, the intercept has no real meaning and can be removed from the model. However, removing the intercept will force the regression model through the coordinate origin ($x = y = 0$), and may bias the coefficient values. As another example, when analysing a particular subject characteristic, such as 'female' or 'senior', the intercept provides a reference for determining the regression coefficient.

Since surveys and questionnaires usually represent sociodemographic factors as categorical variables, their coding schema require special care when interpreting the regression coefficient. One option is to assign ordinal values to categorical variables (e.g., females = 1, males = 2, or usual = 1, unusual = 2). Other categorical variables with sequential meaning (such as 'young', 'middle-aged', and 'senior', or 'under 5 years', '5 to 10 years', and 'over 10 years') can be implicitly modelled as ordinal values. Alternatively, categorical variables can be recoded into a number of separate dichotomous variables. For example, the gender variable can be recoded into a new variable called 'Is_Female', in which all female and male patients are assigned values of 1 and 0, respectively. Similarly, the age variable can be recoded into two new variables 'Is_Middle_Age' and 'Is_Senior'. In this process, the zero code always refers to the reference group of each dichotomous variable, and the intercept is the mean of the *y* coordinates of the regressions of each predictor reference group.

Statistical and machine learning techniques such as information gain and linear regression can identify the homogeneous subpopulations in non-probability samplings.

3.2.4 Demonstration and Evaluation

To demonstrate the design and implementation of the stratification process, we constructed a real-life convenient sampling survey. This large-scale survey (2.5 million records) obtains reliable results because each stratum contains a large number of records after splitting the data. Using the stratification methodology, researchers can quantify the amount of bias in different demographic factors, thereby creating pseudo-controlled samples for accurate and precise representation of the population parameters.

3.3 Research Questions

The stratification process is a natural first step for clarifying the performance information for a particular sub-population. When the patient numbers in the various categories are diverse, the large sub-groups in the population dominate the performance measure scores, obscuring the views of smaller sub-groups. Therefore, the stratification methodology is useful for examining feedback by groups giving substantially different responses in their satisfaction reports. This finer-grained information is particularly useful for assessing and addressing the disparities among smaller sub-groups, and can reveal patterns that otherwise remain unknown.

This thesis investigates the following research questions:

- **Can traditional statistical methods combined with machine learning techniques create a systematic stratification methodology?**

  As big survey data obtained by non-probability sampling methods become increasingly available, researchers are granted the opportunity to investigate conflicting and inconsistent findings on the effects of sociodemographic variables (such as gender, age and race) on the satisfaction level of primary health care. Recent healthcare studies have highlighted the need for a stratified approach, whereby each healthcare ratee gives multiple performance scores (one for each homogeneous stratum) rather than an overall

performance score. How to implement a stratification methodology on large-scale non-probability survey datasets has not been reported in the literature.

Supervised machine learning algorithms such as linear regression and hierarchical modelling can identify the independent variables yielding the best prediction model. This research question investigates whether a systematic stratification methodology can be constructed from an adaptive machine learning technique (based on entropy and information gain) and standard statistical techniques (analysis of variance and principal component analysis). Ideally, the stratification methodology will create homogeneous and mutually exclusive subgroups of patients. Answering this research question would provide a clear procedure for identifying the important sociodemographic factors and stratifying the sampled population into mutually exclusive rater sub populations. Using this procedure, researchers could implement a drill-down stratification analysis guided by rigorous statistical techniques.

- **Can the proposed data stratification methodology create pseudo-controlled samples for estimating population parameters?**

  There is a consensus belief that non-probability sampling yields imprecise estimates of population parameters, because convenient sampling techniques are inherently biased. However, a systematic stratification methodology can reveal patterns of differences between the feedbacks of smaller sub-groups and the overall population. This research question investigates whether researchers can identify the sociodemographic factors that maximise or minimise the variance, and thereby create pseudo-controlled samples for estimating population parameters from a conveniently sampled dataset. The advantage of this outcome is its scalability to other survey analyses based on conveniently sampled data. This technique will derive important insights from large-

scale survey studies and increase the opportunity for warehousing survey data for long-term use.

- **Can a proposed data stratification methodology create a missing-values imputation strategy sensitive to sociodemographic sub groups?**

  When answering satisfaction surveys, raters sometimes provide no answers to a subset of the questionnaire items. A large number of survey responses with missing answers may change the distribution shapes of the questionnaire items and any derived summative items. Many data analysis and statistical techniques are designed only for complete-answer datasets, and respond to missing items by eliminating the subject from the analysis. The statistical requirement of complete-case datasets can significantly reduce the sample size, and will increase the bias if certain subpopulations are more likely than others to return an incomplete survey. Some statistical analyses requiring complete datasets impute the missing values or replace the missing values with grand means. However, these solutions are inappropriate in 'high stakes' survey analyses, in which the satisfaction measure determines promotional or funding outcomes. This research question investigates whether certain sociodemographic sub-populations are associated with higher rates of missing answers, and whether the differences between the rater groups returning complete and incomplete surveys are statistically significant. Such knowledge would reveal whether removing the incomplete cases will likely bias the statistical analysis techniques requiring complete datasets, and whether a stratification analysis can enable imputation strategies that are sensitive to sociodemographic differences.

3.4 Summary

This chapter presented the theoretical background and research motivation for this thesis. It identified many contradictory and inconsistent findings on which sociodemographic factors affect patients in the healthcare domain and customers in the business environment. Recent studies have proposed the stratified approach for analysing large-scale data collected by the convenient sampling methodology (the most popular methodology). The stratification approach provides survey ratees with multiple performance scores (one for each homogeneous stratum) rather than an overall performance score. The research proposal integrates standard statistical and machine learning techniques into a systematic stratification methodology. This chapter also outlined the implementation steps of the research, following the design science research process. Finally, the thesis objectives were presented as three research questions. The next chapter implements an exploratory analysis on a real-life healthcare satisfaction dataset.

Chapter 4 Exploratory Analysis

4.1 Introduction

Previous chapters in this study have highlighted the need for a data stratification methodology to support researchers trying to understand subjects' feedback in convenient sampling methods. The increasing growth of big survey data using non-probability sampling methods have resulted in inconsistent results about the effect of different sociodemographic factors. A systematic stratification methodology can quantify the amount of bias exist in each sociodemographic factor and help researchers to use probability or random sampling to create pseudo-controlled samples for estimating accuracy population parameters.

This chapter starts the process by conducting an exploratory analysis at the zero level of data stratification. The analysis at this level involves the entire population of a non-probability sampling survey. A frequency and descriptive analysis techniques is used to identify global statistics such as mean and standard deviation for all survey items. In addition to that, the dimensionality reduction technique, Principal Component Analysis (PCA) was used to minimize related survey items into a smaller number of uncorrelated variables to account for maximum variance in the data. Applying PCA analysis at different stratification levels can confirm if different population sub-groups have identified the same underline questionnaire components (construct validity). Although the statistical features and construct validity analysis of the dataset used in this research were already established in earlier research work, the analysis presented in this thesis focused on a novel stratification technique to divide survey raters into smaller subpopulations. The PCA was repeated at each subpopulation level to explore whether the results found at a higher stratification level will still be visible at lower stratification levels. Analysis of variance ANOVA was used to confirm if differences

highlighted among sub-populations are statistically significant. The exploratory analysis techniques have helped to identify the different satisfaction profiles that exist among the population large sub-groups such as female vs male and young vs senior.

The exploratory analysis also revealed that despite having statistically significant differences in satisfaction levels among larger sub-groups, different supervised machine learning algorithms were unable to generate accurate models to predict sociodemographic profile based on satisfaction feedback. Further descriptive analysis has identified a considerable overlap in reported satisfaction feedback that may hide differences among smaller population sub-groups. Section 4.2 describes the characteristics of patients' satisfaction survey dataset used in this study, sections 4.3 and 4.4 describes the results of different statistical and supervised machine learning techniques used in the exploratory analysis. Finally, section 4.5 provides a discussion of the chapter results and highlights the analytical motivation for the subsequent stratification analysis.

4.2 Dataset

Improving Practice Questionnaire or IPQ dataset was provided by CFEP pty Australia. The survey was designed to quantify patients' satisfaction and experience about their doctor visit. It consists of 27 performance evaluation questions formed using Likert scale method where categories ranging from (1 – Poor) to (5 – Excellent), null values are represented by zero [85]. There are also two more free text questions to allow comments about how the practice and doctor could improve. The statistical analysis presented in this thesis does not include the free text items. The survey is focused on obtaining information primarily in three core areas: access and booking (Q1 – Q8), practitioners interpersonal skills (Q9 – Q20) and communication with staff (Q21 – Q27). The questionnaire is designed to generate data that have a hierarchical or multilevel stricture. Patients represent the raw-scores "raters" level while practitioners

represent the aggregated-scores “ratees" level. This structure allows multilevel analysis and enables better understanding of the nested relationship between patients and practitioners. The dataset contains 2,546,182 patients’ responses to evaluate 33,203 different physicians (average 77 per physician, minimum 1, and maximum 940). Patients who answered all questions represent 55% or 1,412,588 of IPQ raw scores data. Table 4-1 describes the mean, standard deviation and the available number of records for each question at the patients’ level.

Patients were asked to report 4 socio-demographic variables; gender (2 levels), age (3 levels: young, middle age, and senior), whether the visit was to the usual doctor (2 levels: yes, no), and how many years the patient had visiting the healthcare provider (3 levels: less than 5 years, 5 to 10 years, and more than 10 years). Female patients represent the majority of the dataset 62% while male patients count for 33.6% and 4.5% with gender values un-identified. Middle-age patients cover almost half of the dataset with 53% while senior and young patients’ counts for 32.2% and 9.2% respectively. Unknown age group patients represent 5.6% of the data. Almost 57% of the population have been attending the practice for more than 10 years and 63% of patients are visiting their usual doctor. Patients who did not provide information for the usual doctor variable represent almost 10% of the population followed by 5.6% for age group and 4.5% for both gender and years attending variables. Figure 4-1 shows data distribution for the 4 socio-demographic variables while Table 4-2 describes the four socio-demographic subgroups with the percentage of valid cases available for analysis in each sub-group.

Table 4-1: Descriptive Statistics at the Patients and Practitioners Level

Patients Level				Questions	Practitioners Level			
N	**Mean**	**Std.**	**Range**		**N**	**Mean**	**Std.**	**Range**
2487539	3.61	.957	4	Your level of satisfaction with the practices opening hours	33197	3.63	.295	4
2491304	3.40	1.148	4	Ease of contacting the practice on the telephone	33196	3.45	.517	4
2495582	3.66	1.059	4	Satisfaction with the day and time arranged for your appointment	33198	3.70	.354	4
2471493	3.48	1.198	4	Chances of seeing the doctor within 48-24 hours	33177	3.52	.470	4
2439991	3.27	1.194	4	Chances of seeing a doctor of your choice	33184	3.31	.477	4
2230348	3.31	1.111	4	Opportunity of speaking to a doctor on the telephone	33178	3.35	.398	4
2498718	3.60	.995	4	Comfort level of waiting room	33188	3.63	.365	4
2445860	3.20	1.084	4	Length of time waiting in the practice to see the doctor	33178	3.23	.416	4
2488031	4.18	.878	4	My overall satisfaction with this visit to the doctor	33193	4.23	.293	4
2487155	4.23	.860	4	The warmth of the doctors greeting to me	33191	4.28	.297	4
2478202	4.25	.870	4	On this visit I would rate the doctors ability to really listen	33194	4.30	.293	4
2475948	4.19	.877	4	The doctors explanations of things to me were	33192	4.24	.292	4
2468541	4.14	.904	4	The extent to which I felt reassured by this doctor was	33193	4.19	.300	4
2479558	4.27	.862	4	My confidence in this doctor’s ability	33193	4.32	.286	4
2466781	4.17	.894	4	The opportunity the doctor gave me to express my concerns	33193	4.23	.292	4
2479840	4.33	.830	4	The respect shown to me by this doctor was	33191	4.38	.269	4
2439126	3.91	.939	4	The amount of time given to me for this visit was	33187	3.97	.295	4
2404854	4.09	.904	4	This doctor’s consideration of my personal situation	33185	4.15	.292	4
2415940	4.13	.903	4	The doctors concern for measapersonin this visit was	33188	4.18	.299	4
2420164	4.22	.901	4	The recommendation I would give to my friends about this doctor	33189	4.27	.308	4
2509129	3.99	.939	4	The manner in which you are treated by the reception staff	33190	4.03	.295	4
2477533	3.97	.950	4	Respect shown for your privacy and confidentiality	33185	4.00	.281	4
2418140	3.83	.986	4	Information provided by the practice about its service	33182	3.86	.286	4
2184651	3.58	.978	4	The opportunity for making compliments or complaints	33173	3.61	.291	4
2336693	3.73	.946	4	The information provided by this practice	33179	3.76	.265	4
2270202	3.64	1.020	4	The availability and administration of reminder systems	33181	3.67	.296	4
1914084	3.64	.972	4	The practices respect to your right to seek a second opinion	33161	3.67	.287	4

Table 4-2: Valid and Missing Values by Sociodemographic

Sociodemographic Factor	Valid	Missing
Gender	2432851	113331
Age	2403143	143039
Usual GP	2299774	246408
Years Attending	2430667	115515

Figure 4-1: Socio-Demographic Distribution

4.3 Data Analysis

The statistical analysis at zero level involves the entire sample population (2,546,182 patients) with all 4 socio-demographic variables. The descriptive statistics on IPQ dataset shows that patients' feedback was skewed toward the positive end of the Likert scale for all 27 items with global mean value of 3.85 for all patients. Question 8 (Length of time waiting in the practice to see the doctor) has the lowest mean value of 3.2 while question 16 (The respect shown to me by this doctor) has the highest mean value of 4.33. The skewness was more obvious for practitioners interpersonal skills items (Q9-Q20) with lowest mean values 3.91 for question seventeen "The amount of time given to me for this visit was". Practitioners' interpersonal skills items also show the lowest standard deviation values among all other IPQ items.

Examining the wording of all 27 items in addition to inter-correlation analysis showed that no single item can qualify to be a summative variable for IPQ dataset. A new overall scale item with a range from 0 to 100 was created to summarise all 27 items into a single value. Examining the overall scale revealed a negatively skewed distribution with the majority of patients report a high satisfaction results (Mean = 73, std = 15.9). The skewness in the overall scale reflects a similar pattern in the original distribution shape of the 27 items.

The exploratory analysis on IPQ 27 items also showed a large number of missing values where patients did not provide full answers in questionnaire forms. The number of missing values range from (37053) for question 21 (The manner in which you were treated by the reception staff) to (632098) for question 27 (The right to seek a second opinion). The missing answers can increase the negative skewness of the overall scale and lead to a large portion of patients to be considered as "outlier". Therefore, a "clean" version of IPQ dataset with non-missing values for all 27 items and socio-demographic information was generated. The new dataset

contains (1,251,357 patients). Figure 4-2 shows the distribution of overall scale with complete and clean IPQ datasets.

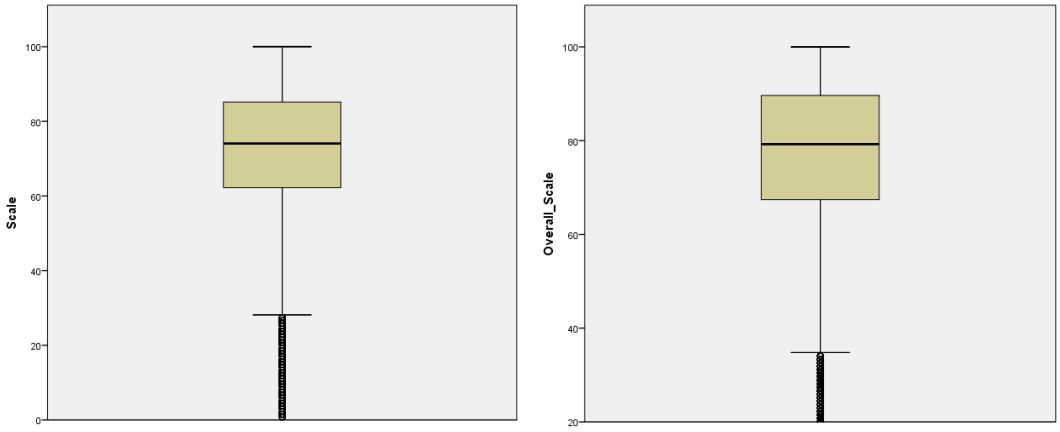

Figure 4-2: Overall Scale Distribution with all patients (Left) and non-missing Patients (Right)

The construct validity of the three core components of IPQ was evaluated using PCA with orthogonal (varimax) rotation. The Kaiser-Meyer-Olkin (KMO) and bartlett's test verified the sampling adequacy for the analysis, KMO = .978 and all KMO values for individual items were > 0.90, which is well above the accepted limit of (0.5). At the patients' level, results show that the entire population of IPQ dataset and each of its individual socio-demographic sub-groups clearly identified the three underlining components of IPQ dataset with eigenvalues over Kaiser's criterion of 1 and in combination explained about 74% of the variance. The practitioners' communication and interpersonal skills stand up as the most significant component with highest Eigenvalues. Communication with staff and access to clinic appear to be switching places between the second and third most significant components across different IPQ patients' sub-groups. Two questions (Q7, Q8) tend to show loading values that are between access to the clinic and communication with staff components. The PCA analysis was also repeated at the aggregated practitioners level. The aggregated mean score values were calculated for all 33,203 doctors. The analysis showed that removing patients

sociodemographic factors at the higher practitioners level only identified two of the original three components with initial eigenvalues of more than one. In other words, the analysis shows that access to the clinic and communication with staff items were merged together as the second most significant component while practitioners' interpersonal skills continues to hold its position with the highest Eigenvalues. The scree plot shown in Figure 4-3 displays the inflexion point that will justify a 3 components solution at the patients level while Table 4-3 shows the factor loadings after rotation.

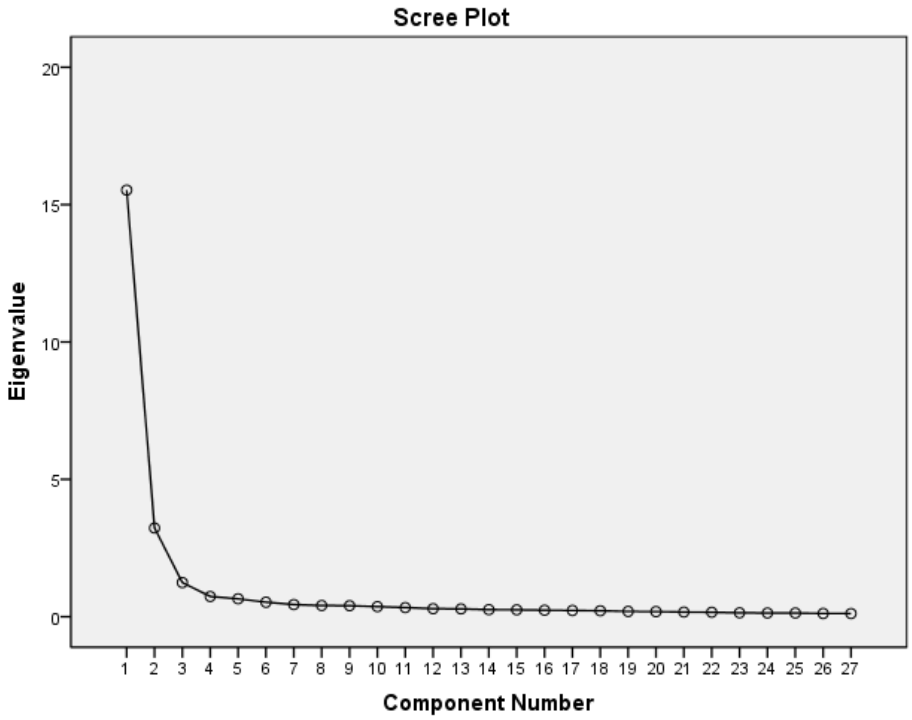

Figure 4-3: Scree Plot for PCA at the Patients Level

Table 4-3: Rotated Loadings at Zero Level

Qs	Communication	Access	Staff
Q01		0.65	
Q02		0.70	
Q03		0.76	
Q04		0.77	
Q05		0.74	
Q06		0.68	
Q07		0.49	
Q08		0.61	
Q09	0.79		
Q10	0.81		
Q11	0.86		
Q12	0.84		
Q13	0.85		
Q14	0.85		
Q15	0.85		
Q16	0.85		
Q17	0.71		
Q18	0.81		
Q19	0.82		
Q20	0.83		
Q21			0.68
Q22			0.71
Q23			0.74
Q24			0.71
Q25			0.74
Q26			0.73
Q27			0.70

The rotated loading values were used to calculate patients scores for each of the three underlining components of IPQ dataset. In order to maximize the differences between the three components, only questions related to the corresponding component were used to calculate patients component scores according to the following formula:

$$Y_i = b_1X_{1i} + b_2X_{2i} + \dots b_n X_{ni}$$

Where

Y = the component

b = the Loading Factor value

X = the individual score for that question

The process above added three new uncorrelated score values for each patient record; “Clinic Access” based on questions 1-8; “Practitioner Communication” based on questions 9-20; and “Staff Information” based on questions 21-27. The values presented in each score will differ based on the number of items involved in each component. To provide fair comparison among the new component scores, three new standardized components scores are also added to the dataset. These score values were used to investigate differences in satisfaction profiles among multiple sociodemographic subpopulations. The next section provides details investigation about satisfaction feedback among the different sociodemographic groups in IPQ dataset.

4.4 Subpopulations Analysis

The previous sections presented exploratory analysis for IPQ dataset and explained several steps of data cleaning and preparation. The process involved creating several new summative and standardized attributes at the row scores level. This section investigates satisfaction feedbacks among patients subpopulations based on the hypothesis that there are real differences among sociodemographic subgroups such as females vs males and young vs senior. The ANOVA test was used to investigates if there are statistically significant differences among sociodemographic subgroups. The test was repeated with different percentages of patients records included. Following that, supervised machine learning models are built to investigate if patients sociodemographic characteristics can be predicted based on satisfaction feedback data. Section 4.4.1 presents the sociodemographic analysis between smaller patients sub-groups while section 4.4.2 presents the results of predicting sociodemographic characteristics using supervised machine learning algorithms.

4.4.1 Sociodemographic Analysis

This section presents the average score values for the four sociodemographic factors and their underlining levels on the different items, summative variable, and standardized components of IPQ dataset. Female patients represent the majority (62%) of the valid IPQ dataset responses. Overall, female patients gave lower scores than male patients in all 27 Likert scale items as well as the overall-scale summative item. However, the difference between females and males mean scores was smaller for doctor communication questions. The results are also reflected in the corresponding principal components; clinic access (Qs 1-8), doctor interpersonal skills (Qs 9-20), staff information (Qs 21-27). For both clinic access and staff information components, female patients gave below average scores (17.97, 17.25) compare to male patients who gave above average scores (18.7, 18.02) respectively. Figure 4-4 show IPQ items scores differences between all females / males subgroups and the entire dataset; figures 4-5 to 4-7 show gender scores for 3 components and standardized components.

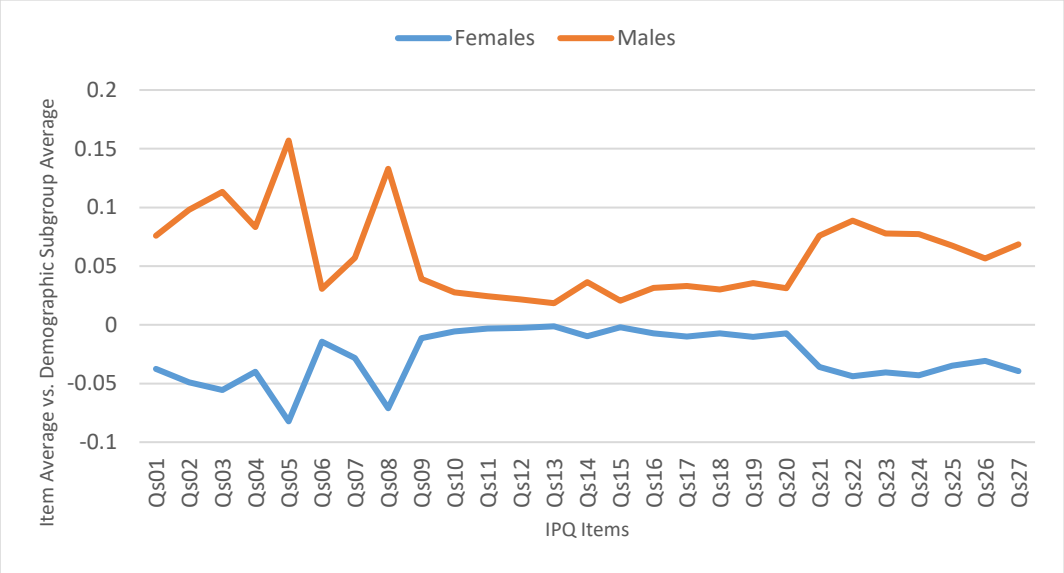

Figure 4-4: IPQ Items Scores Differences between all Females / Males Subgroups and the Entire Dataset

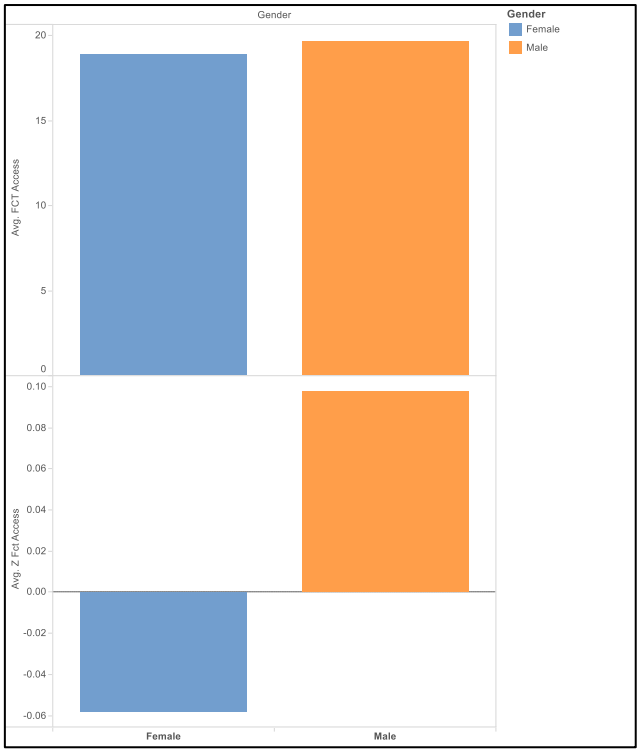

Figure 4-5: Females and Males Scores on the Original (above) and Standardized (below)Access Component

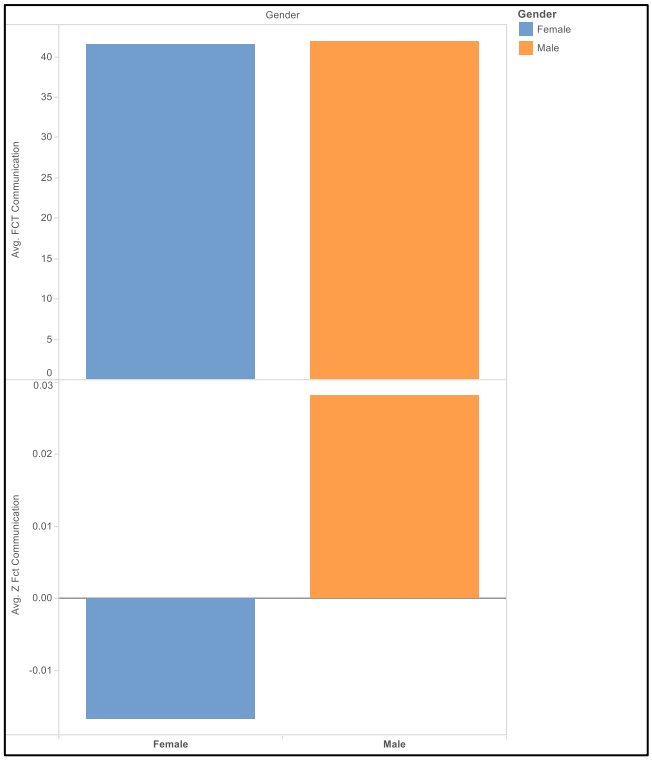

Figure 4-6 Females and Males Scores on the Original (above) and Standardized (below) Communication Component

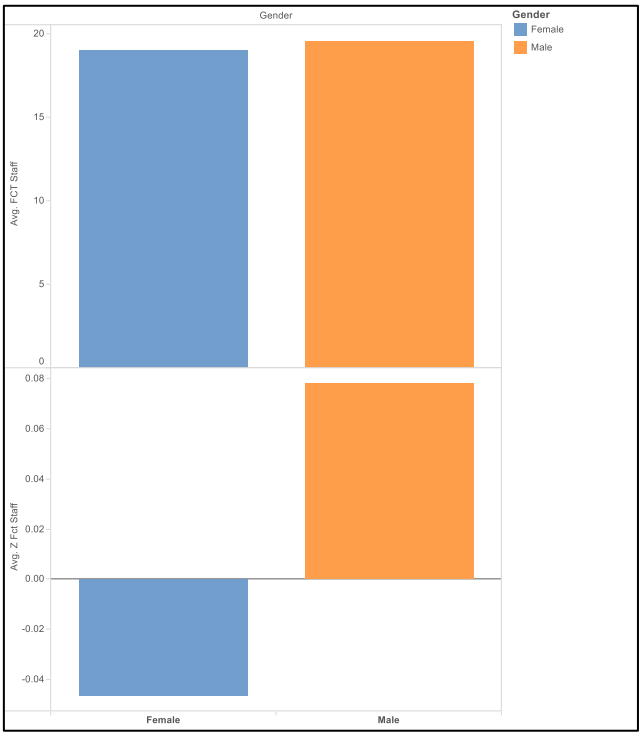

Figure 4-7: Females and Males Scores on the Original (above) and Standardized (below) Staff Information Component

Analysis of variance ANOVA test implemented between females and males' groups showed statistically significant differences ($p \leq 0.01$) in all 27 items. The analysis used the entire IPQ sample population of 2,546,182 patients; and was also repeated to include different percentages (50%, 20%, 10%, 5%, 2%, and 1%) of IPQ dataset. The repeated analysis showed that differences between females and males patients for doctor communication questions are becoming insignificant as the smaller the dataset gets. However, gender differences were still statistically significant for clinic access and staff information with the exception of question six "Opportunity of speaking to doctor by phone".

For age groups, middle age patients represent more than half (53%) of the valid IPQ population. Senior patients gave considerably higher feedback scores for all IPQ questions followed by middle-age and young patients respectively. The raw scores results are also reflected in the corresponding principal components; clinic access (Qs 1-8), doctor interpersonal skills (Qs 9-20), staff information (Qs 21-27). For all principal components, young and middle age patients gave below average scores Figures 4-8 to 4-11 shows gender scores for 3 components and standardized components.

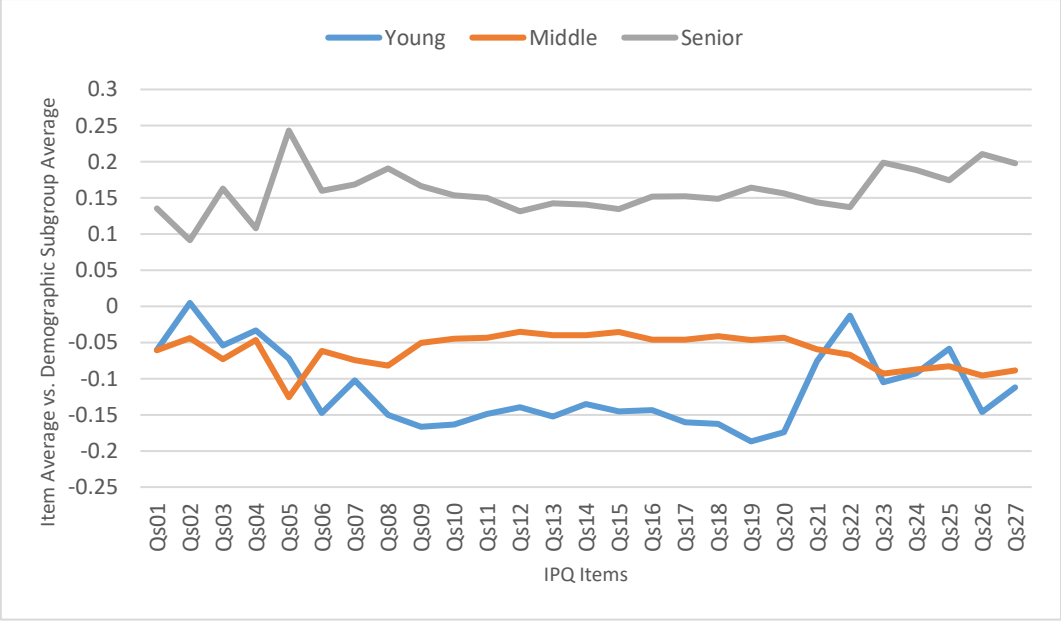

Figure 4-8: IPQ Items Scores Differences between all Young, Middle-Age, and Senior Subgroups and the Entire Dataset

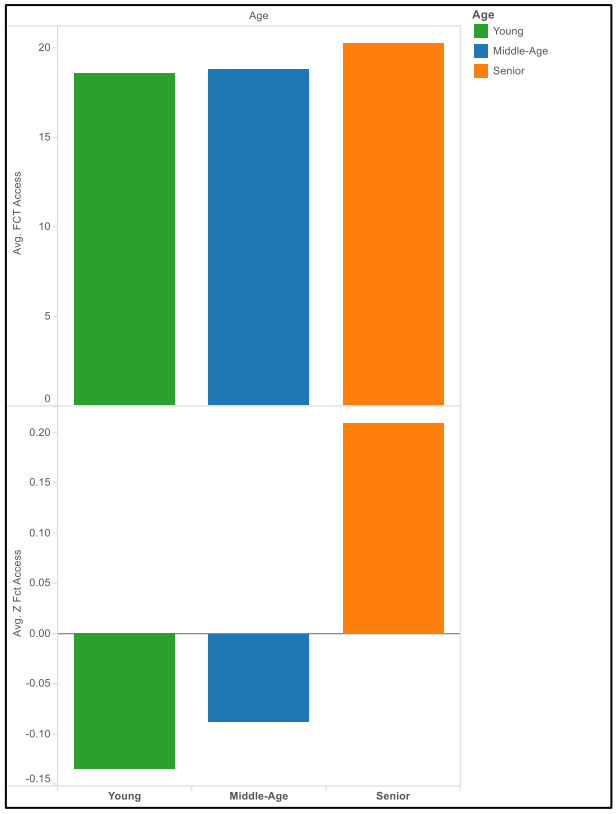

Figure 4-9: Young, Middle-Age and Senior Scores on the Original (above) and Standardized (below) Access Component

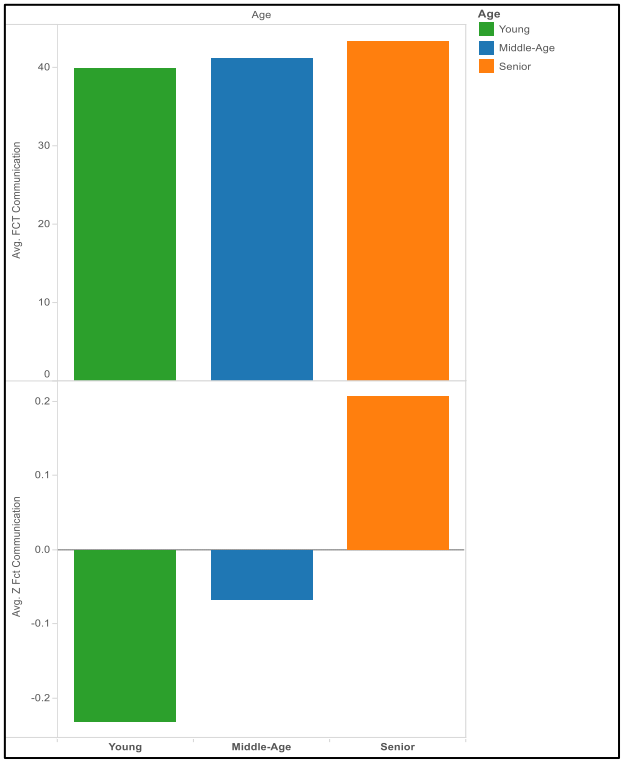

Figure 4-10: Age Groups Scores on the Original (above) and Standardized (below) Communication Component

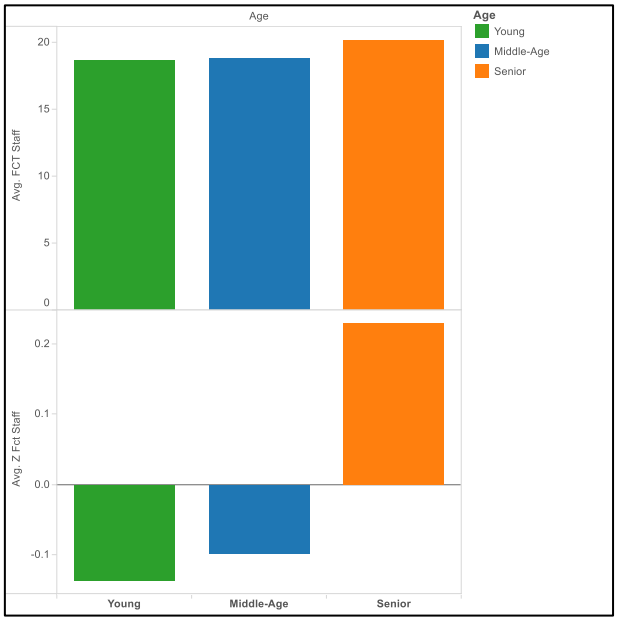

Figure 4-11: Age Groups Scores on the Original (above) and Standardized (below) Staff Information Component

The ANOVA test implemented between young, middle-age and senior groups showed statistically significant differences ($p \leq 0.01$) in all 27 items of IPQ dataset with the exception of question five "Seeing a doctor of your choice". The ANOVA test was also repeated with different percentages (50%, 10%, 5%, 2%, and 1%) of IPQ dataset. The repeated analysis showed that differences between young, middle-age, and senior patients for doctor communication questions remained significant even with smaller samples sizes. On the other hand, differences between age groups for clinic access and staff information are becoming insignificant as the smaller the dataset gets.

For years attending group, almost 57% of the valid IPQ responses come from patients who have been seeing their doctors for more than 10 years. Patients' seeing the doctor for less than 5 years gave the highest scores for almost all clinic access questions while patients seeing the doctor for more than 10 years gave the highest scores for doctor communication questions. The results are also reflected in the corresponding principal components; clinic access (Qs 1-8), doctor interpersonal skills (Qs 9-20), staff information (Qs 21-27). For clinic access component, patients seeing the doctor for less than 5 years gave the highest and above average scores (19.24) compared to (18.91, 19.02) for 5 to 10 years and more than 10 years respectively. For doctor communication component, patients who are seeing the doctor for less than 10 years gave below average scores. However, the scores pattern shows a positive relationship between the time spent with the doctor and the feedback scores. Figures 12-15 shows years attending scores for 3 components and standardized components.

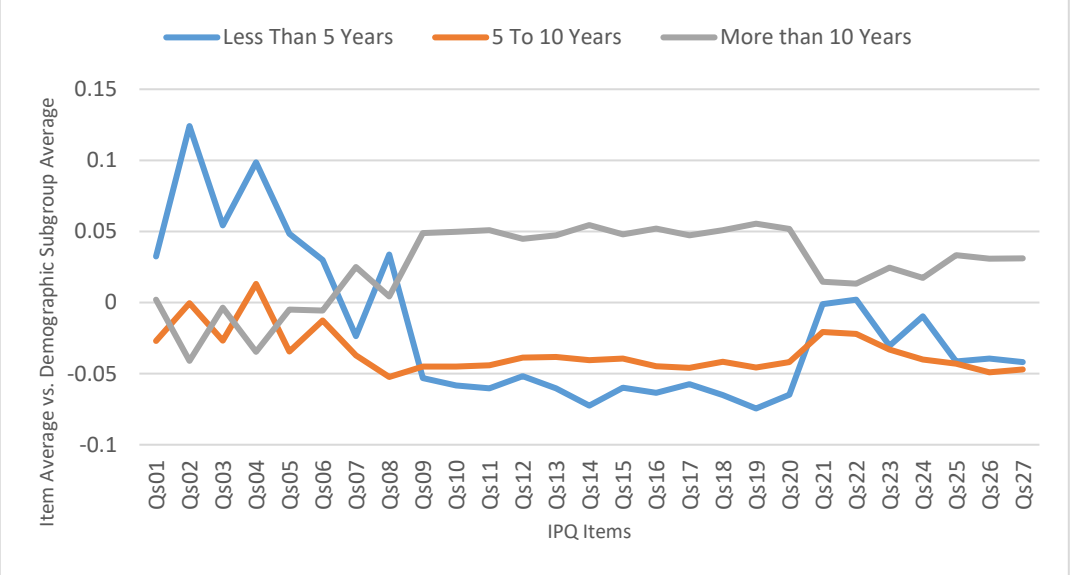

Figure 4-12: IPQ Items Scores Differences Between Years Attending Subgroups and the Entire Dataset

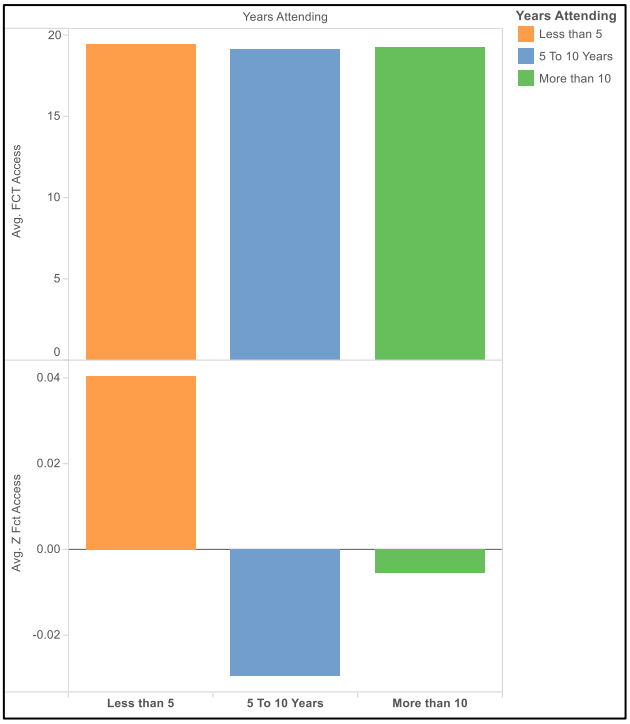

Figure 4-13: Years Attending Scores on the Original (above) and Standardized (below) Access Component

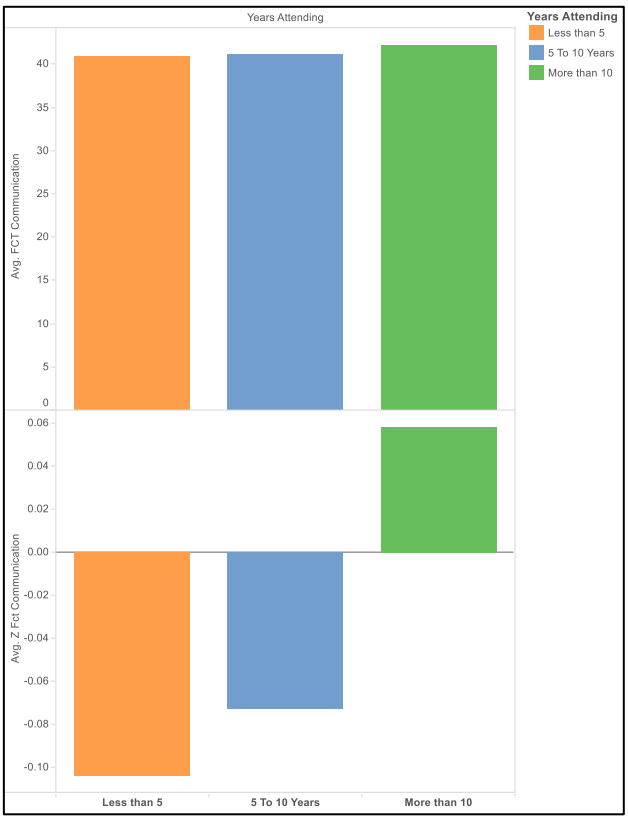

Figure 4-14: Years Attending Scores on the Original (above) and Standardized (below) Communication Component

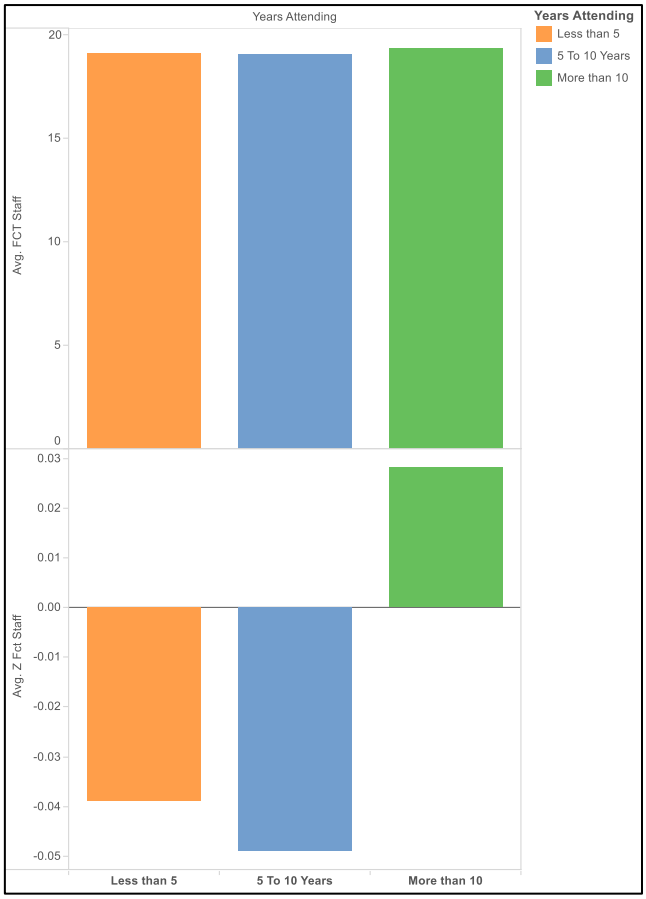

Figure 4-15: Years Attending Scores on the Original (above) and Standardized (below) Staff Component

The ANOVA test implemented between the three years-attending groups showed statistically significant differences ($p \leq 0.01$) in all clinic access and doctor communication items. Several staff information questions showed statistically insignificant differences among at least two homogeneous subgroups (less than 5 years and 5 to 10 years). The ANOVA test was also repeated with different percentages (50%, 10%, 5%, 2%, and 1%) of IPQ dataset. The repeated analysis showed that differences between years attending patients subgroups become statistically insignificant with smaller samples size. The Tukey post-hoc revealed that several items showed three homogeneous subgroups at the sample sizes between one and two percent of IPQ dataset.

Finally, Patients who are evaluating their usual doctor represent 70% of the valid IPQ population. Patients' seeing their usual doctor gave higher scores for all questions. The results are reflected in the corresponding principal components; clinic access (Qs 1-8), doctor interpersonal skills (Qs 9-20), staff information (Qs 21-27). For all 3 components, non-usual patients gave below average scores than usual patients. Figures 16-19 shows usual and non-usual doctor scores for 3 components and standardized components. The ANOVA test showed statistically significant differences between the usual and non-usual doctor groups ($p \leq 0.01$) in all 27 items and with different percentages (50%, 10%, 5%, 2%, and 1%) of IPQ dataset.

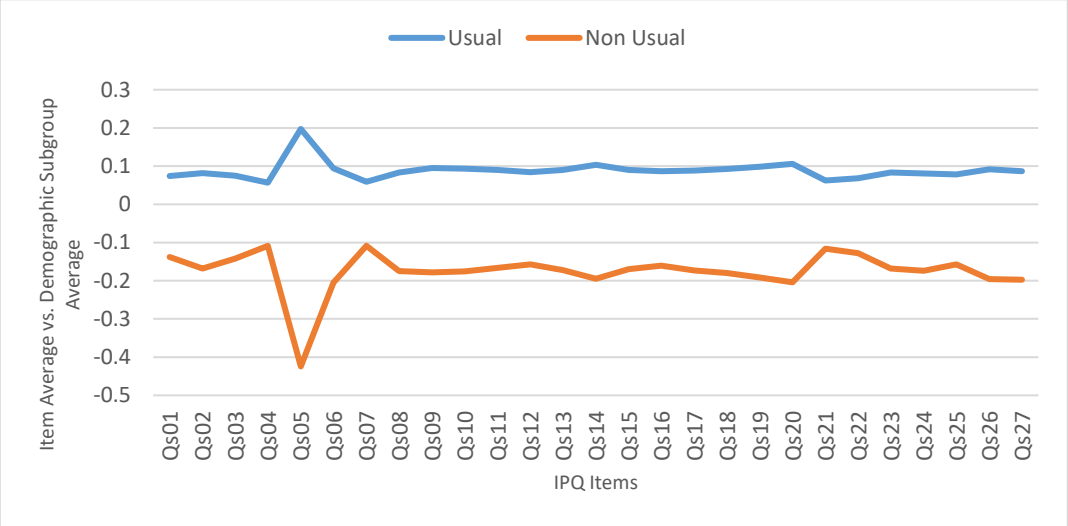

Figure 4-16: IPQ Items Scores Differences Between Usual and Non-Usual Subgroups and the Entire Dataset

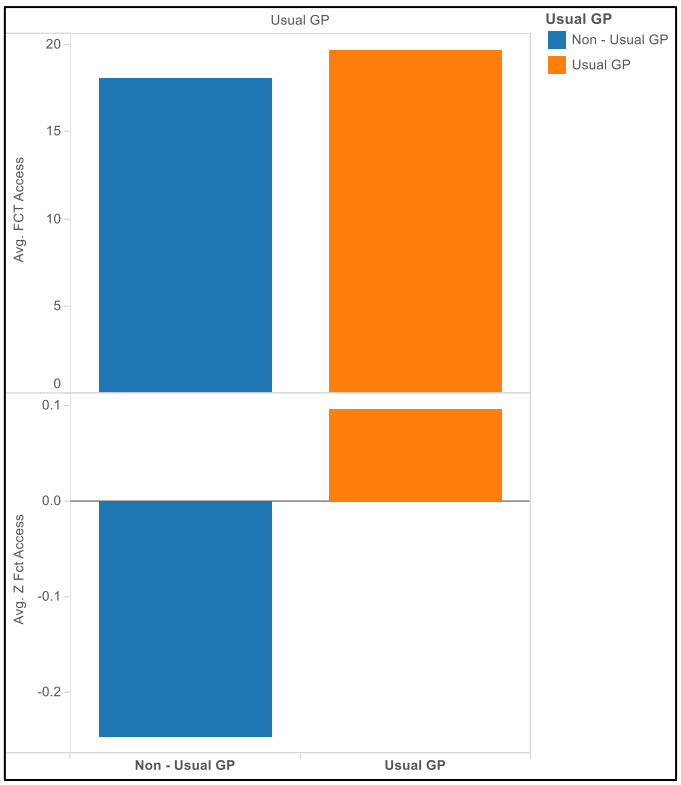

Figure 4-17: Usual and Non-Usual Scores on the Original (above) and Standardized (below) Access Component

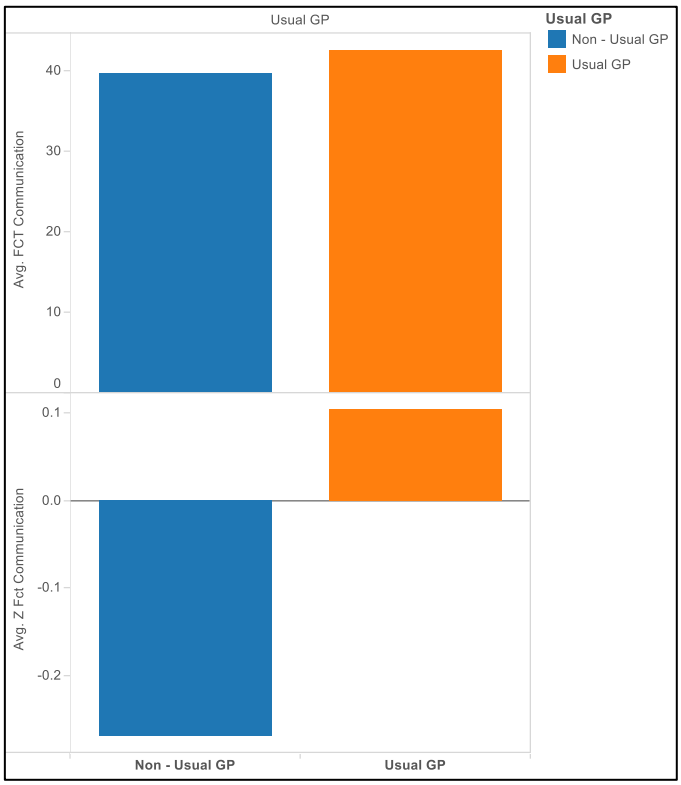

Figure 4-18: Usual and Non-Usual Scores on the Original (above) and Standardized (below) Communication Component

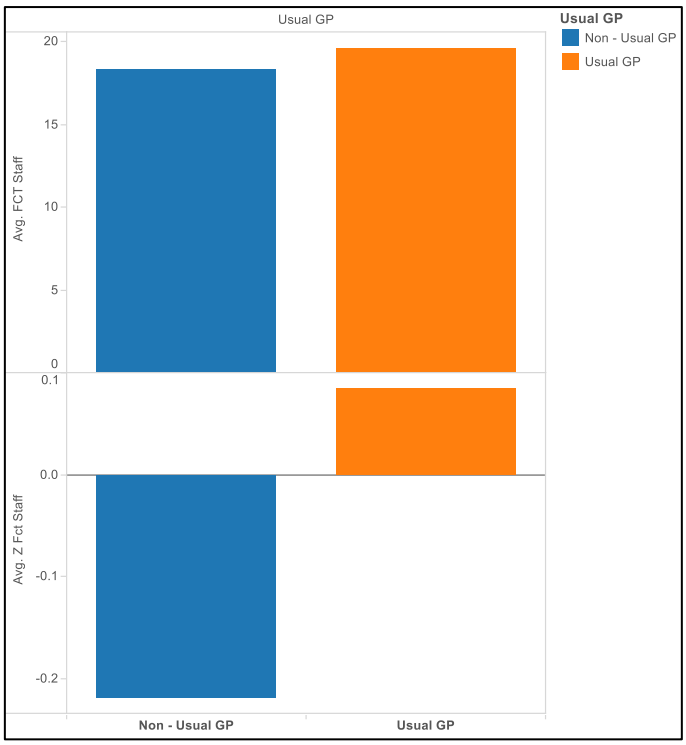

Figure 4-19: Usual and Non-Usual Scores on the Original (above) and Standardized (below) Staff Component

Table 4-4 shows the mean and standard deviation values for the 6 components at the raw scores level and irrespective of any patients sub-group while Table 4-5 shows mean and standard deviation values for the 6 components by each socio-demographic sub-group.

Table 4-4: Descriptive Statistics for the 6 Components

Components	Minimum	Maximum	Mean	Std. Deviation
Access	5.44	27.2	19.056	4.6977
Communication	9.939	49.695	41.825	7.82
Staff	5.047	25.238	19.284	4.179
Access - Standardized	-2.897	1.735	0.00	1.00
Communication - Standardized	-4.0736	1.0054	0.00	1.00
Staff - Standardized	-3.406	1.4242	0.00	1.00

Table 4-5: Descriptive Statistics for the 6 Components by Different Sub-Groups

	Overall Scale	Factor Access	Factor Access - Standardized	Factor Communication	Factor Communication - Standardized	Factor Staff	Factor Staff - Standardized
Gender							
Female	72.7544	18.78	-0.05	41.69	-0.016	19.09	-0.046
Male	74.8705	19.51	0.09	42.04	0.028	19.61	0.078
Age group							
Young	71.8024	18.42	-0.13	40	-0.23	18.7	-0.13
Middle Age	72.4445	18.64	-0.08	41.3	-0.06	18.87	-0.09
Senior	76.0363	20.03	0.2	43.43	0.2	20.24	0.23
Usual GP							
Yes	75.6730	19.5	0.09	42.64	0.10	19.63	0.08
No	69.6644	17.89	-0.24	39.7	-0.269	18.37	-0.218
# years with doctor							
Less than five	72.6315	19.24	0.04	41.01	-0.10	19.12	-0.038
Five to ten	73.0450	18.91	-0.02	41.26	-0.07	19.08	-0.048
More than ten	74.0451	19.02	-0.005	42.27	0.05	19.4	0.028

For each sociodemographic factor, the ANOVA test was repeated using decreasing sample sizes (100%, 90% … 1%). The goal of this approach is to investigate if the chosen significance level and sample size would highlight any insights about the stratification effect of each sociodemographic factor. At each sample size, the ANOVA test was used to identify if a given item or component score provided statistically significant evidence of a non-zero effect size. The results showed that as sample size decrease, more differences among items start to appear statistically insignificant. Observing the number of statistically significant or insignificant items at certain sample size points would help researchers' to identify sociodemographic factors with higher or lower effect size. The results also highlighted clear differences in the "Type" of items that are considered significant with different sample sizes. Items related to doctors communication skills start to appear statistically insignificant between females and males patients with sample sizes around 20% of IPQ dataset. The ANOVA test was repeated for ten times at each sample size level to identify if each item would appear as statistically insignificant in more than five times. Table 4-6 and Table 4-7 shows a representation about at

which sample size level IPQ items would appear statically insignificant. Figure 4-20 shows a representation for the number of statistically insignificant items associated with each sample size. Figure 4-21 shows a logarithmic model to predict the number of statistically insignificant items at each sample size. The results show sociodemographic factors have unequal effect size that would need different sample size to appear statistically significant. The sociodemographic factor "Usual Doctor" appears to have the highest impact factor while "Years Attending" shows the lowest impact factor.

Table 4-6: Sample Size and ANOVA Test for Gender and Age Groups

ANOVA Test for Gender						**ANOVA Test for Age Group**					
Sample Size						**Sample Size**					
100%	50%	5%	2%	1%	Questions	100%	50%	10%	5%	2%	1%
					Qs01						
					Qs02			X	X	X	X
					Qs03			X	X	X	X
					Qs04				X	X	X
					Qs05	X	X	X	X	X	X
			X	X	Qs06						
					Qs07						X
					Qs08						
				X	Qs09						
				X	Qs10						
				X	Qs11						
		X	X	X	Qs12						
		X	X	X	Qs13						
				X	Qs14						
		X	X	X	Qs15						
				X	Qs16						
				X	Qs17						
				X	Qs18						
				X	Qs19						
				X	Qs20						
					Qs21						
					Qs22				X	X	X
					Qs23						
					Qs24						X
					Qs25		X	X	X	X	X
					Qs26						
					Qs27						X
					Access						
				X	Communication						
					Staff						
X = At Least Two Homogeneous Subsets											

Table 4-7: Sample Size and ANOVA Test for Usual Doctor and Years Attending

ANOVA Test for Usual GP							ANOVA Test for Years Attending					
Sample Size							Sample Size					
100%	50%	10%	5%	2%	1%	Questions	100%	50%	10%	5%	2%	1%
						Qs01				X	X	X
						Qs02			X	X	X	X
						Qs03				X	X	X
						Qs04			X		X	X
						Qs05				X	X	X
						Qs06			X	X	X*	X*
						Qs07			X		X	X
						Qs08					X	X
						Qs09				X	X	X
						Qs10				X	X	X
						Qs11				X	X	X
						Qs12				X	X	X
						Qs13				X	X	X
						Qs14					X	X
						Qs15				X	X	X
						Qs16				X	X	X
						Qs17				X	X	X
						Qs18				X	X	X
						Qs19				X	X	X
						Qs20				X	X	X
						Qs21				X	X	X*
						Qs22				X	X	X
						Qs23	X	X	X	X	X	X
						Qs24				X	X	X
						Qs25	X	X	X	X	X	X
						Qs26	X	X	X	X	X	X
						Qs27	X	X	X		X	X
						Access				X	X	X
						Communication				X	X	X
						Staff					X	X

X = At Least Two Homogeneous Subsets
X* = Three Homogeneous Subsets

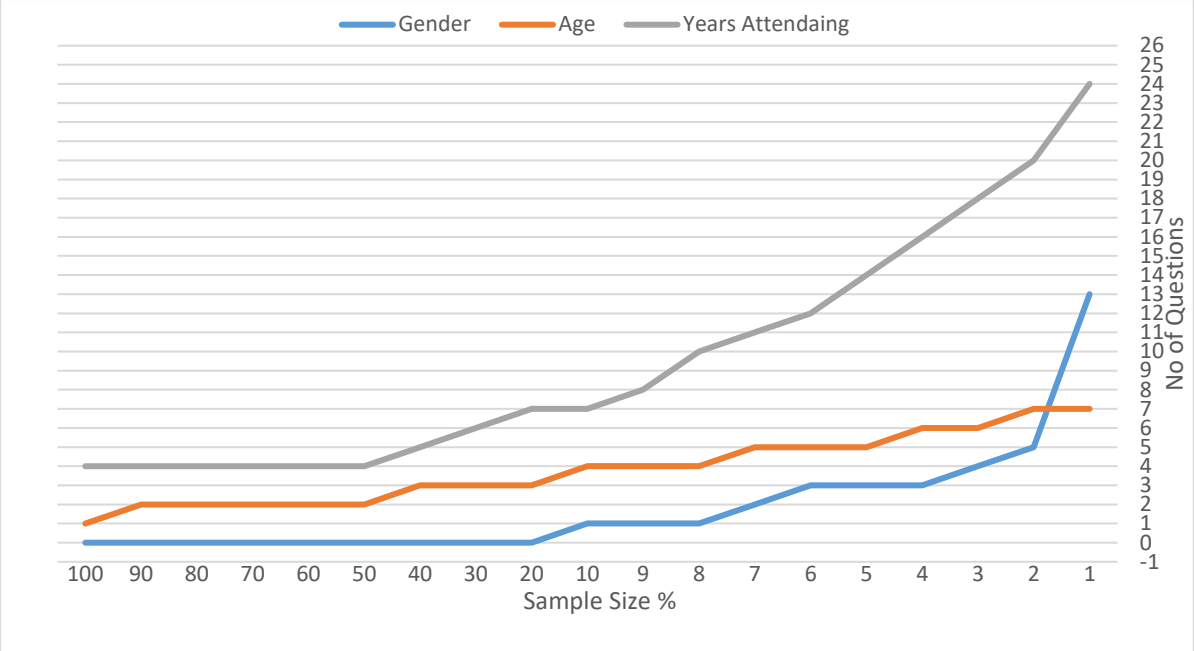

Figure 4-20: No. of statically Insignificant Items in for Each Sociodemographic Factor

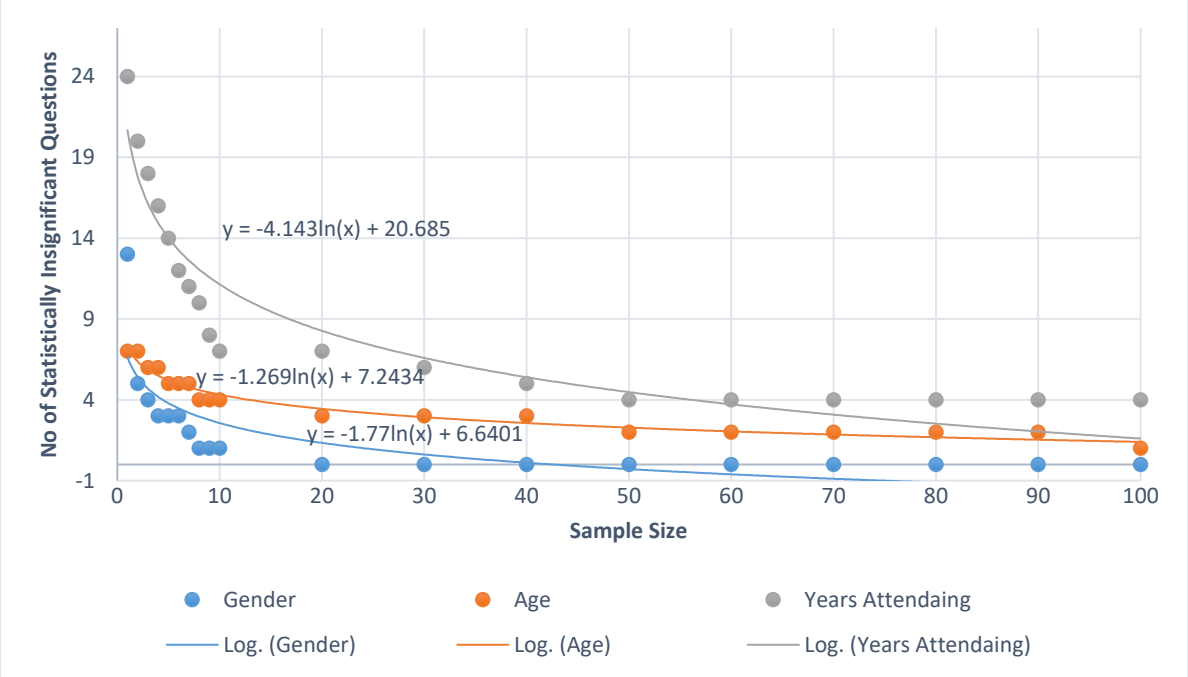

Figure 4-21: Logarithmic Model to predict the number of Statistically Insignificant Items at Each Sample Size

The analysis above used the significant level alpha 0.05 as the statistically acceptable standard cut-off point in social science studies. However, this standard is often set in studies where sample size is relatively small. With the availability of large-scale datasets, researchers can investigate if items and components would appear insignificant at different alpha level values. For example, examining the p-value of question two “Ease of contacting the clinic by telephone” appears statistically significant among all three age groups. However, adjusting the alpha level to 0.02 reveal a homogeneous subgroup between young and middle-aged patients for this item. On the other hand, question five “Chance of seeing the usual doctor” was statistically insignificant between young and middle age patients at 0.05 alpha level. It needed adjusting the significance level to 0.36 to get statistically significant differences among the three age groups. The significant levels are also affected by the available sample size. If the effect size is small, a large sample size is needed to detect the difference. Similarly, if the effect size is large, even a small sample size can show statistically significant differences among subgroups.

4.4.2 Predicting Sociodemographic Characteristics using Supervised Learning

The previous section highlighted that all sociodemographic factors measured in IPQ dataset showed statistically significant differences among their respective smaller subgroups. Senior patients and patients who are seeing their usual doctor gave significantly higher sores that other subgroups. Similarly, female and young patients gave lower scores that other subgroups. This section investigates if supervised machine learning algorithms can predict sociodemographic characteristics based on the patient feedback scores.

A decision tree classification algorithm was built to predict sociodemographic factors using the three component scores as independent variables. The classification algorithm was repeated using different growing methods (CHAID, Exhaustive CHAID, CRT, and QUEST) and holding 80% of data for model training and 20% for testing. In all models, the accuracy of the classification algorithm was affected by the heavily skewed distribution of class labels. For example, the usual doctor model achieved 70% prediction accuracy due to the imbalanced distribution of class labels. Therefore, all classification models used equal-sized label scores. The resulted prediction accuracy ranges between 35% and 59% for all sociodemographic subgroups. Despite having a statistically significant different among sociodemographic subgroups, the distribution overlap makes it highly challenging to predict the label group based on patients' feedback values. The highest prediction accuracy of 59% achieved in "Usual Doctor" model due to large different between patients who are evaluating their usual or non-usual doctors. This difference was also highlighted in the previous section when ANOVA test showed statistically significant differences among the two subgroups in all 27 items even when using smaller sample sizes. On the other hand, the lowest accuracy result of 35% achieved by the years attending models. The previous section revealed a large amount if similarity among its subgroup through ANOVA test. However, there is currently no systematic methods to quantify these differences among multiple sociodemographic subgroups in non-probability

sampling studies. Figures 22-25 shows the Kernel Density Estimation KDE graphs for the four sociodemographic factors on the overall scale variable. Tables 8 - 11 shows the results of classification accuracy models using different growing methods.

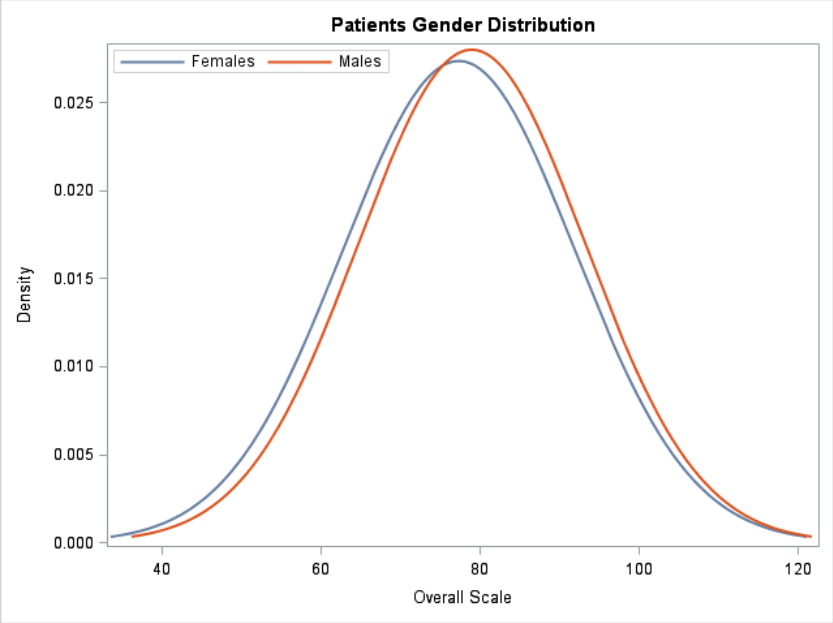

Figure 4-22: Patients Gender Distribution

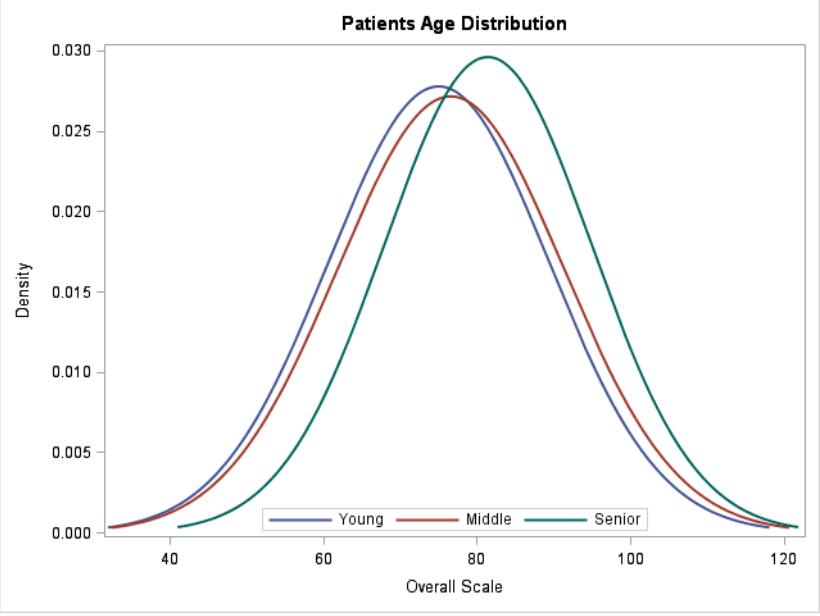

Figure 4-23: Patients Age Groups Distribution

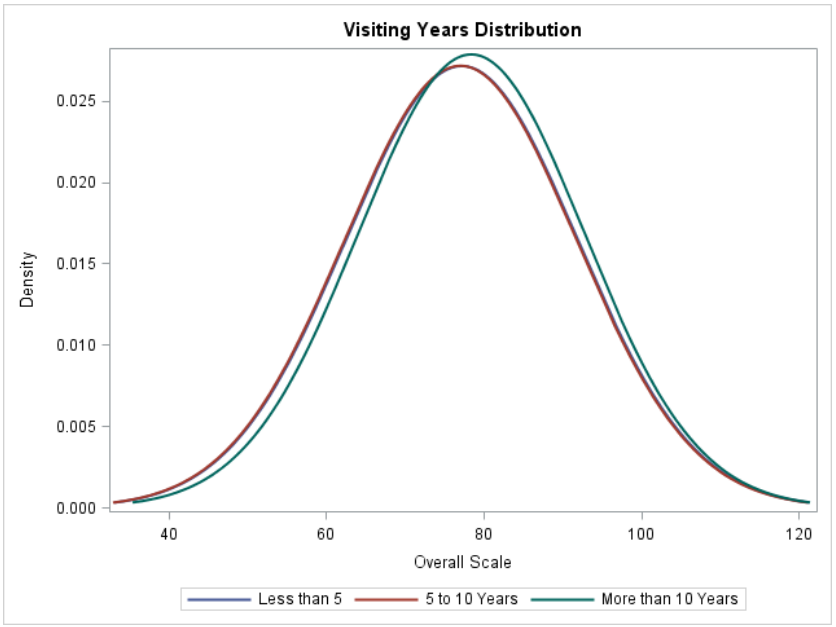

Figure 4-24: Patients Years Attending Groups Distribution

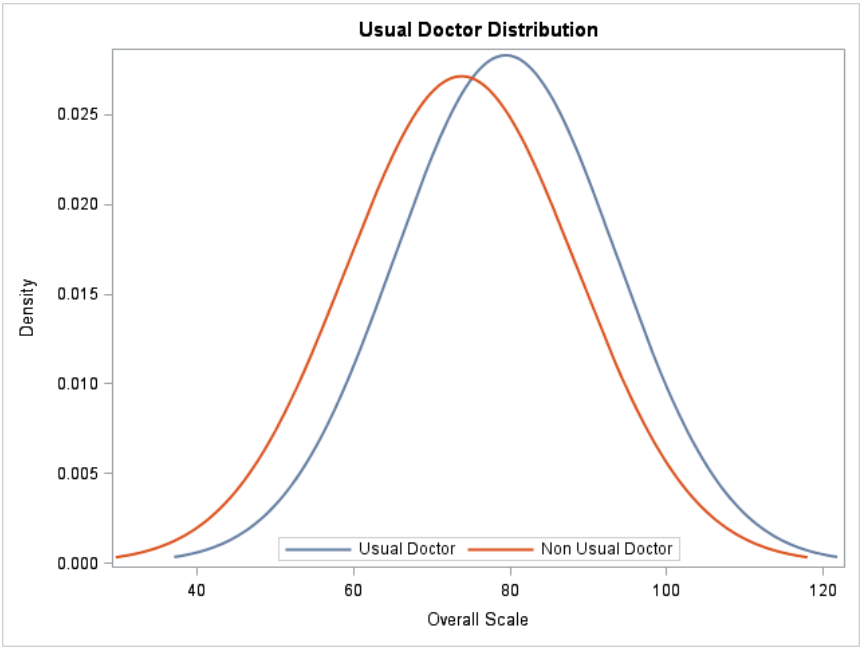

Figure 4-25: Patients usual Doctor Groups Distribution

Table 4-8: Classification Model for Gender

Sample		Predicted		
		Female	Male	Correct %
Training	Female	4013	3994	50.1%
	Male	3234	4800	59.7%
	Overall %	45.2%	54.8%	54.9%
Test	Female	952	1041	47.8%
	Male	839	1127	57.3%
	Overall %	45.2%	54.8%	52.5%

Growing Method: CHAID

Sample		Predicted		
		Female	Male	Correct %
Training	Female	3595	4416	44.9%
	Male	2880	5105	63.9%
	Overall %	40.5%	59.5%	54.4%
Test	Female	860	1129	43.2%
	Male	749	1266	62.8%
	Overall %	40.2%	59.8%	53.1%

Growing Method: EXHAUSTIVE CHAID

Sample		Predicted		
		Female	Male	Correct %
Training	Female	4606	3350	57.9%
	Male	3845	4205	52.2%
	Overall %	52.8%	47.2%	55.0%
Test	Female	1188	856	58.1%
	Male	950	1000	51.3%
	Overall %	53.5%	46.5%	54.8%

Growing Method: CRT

Sample		Predicted		
		Female	Male	Correct %
Training	Female	5361	2698	66.5%
	Male	4622	3379	42.2%
	Overall %	62.2%	37.8%	54.4%
Test	Female	1270	671	65.4%
	Male	1159	840	42.0%
	Overall %	61.6%	38.4%	53.6%

Growing Method: QUEST

Table 4-9: Classification Model for Age

Sample		Predicted			
		Young	Middle Age	Senior	Correct %
Training	Young	4203	1365	2484	52.2%
	Middle Age	3296	1870	2787	23.5%
	Senior	2952	1374	3644	45.7%
	Overall %	43.6%	19.2%	37.2%	40.5%
Test	Young	980	347	620	50.3%
	Middle Age	869	468	709	22.9%
	Senior	781	377	870	42.9%
	Overall %	43.7%	19.8%	36.5%	38.5%

Growing Method: CHAID

Sample		Predicted			
		Young	Middle Age	Senior	Correct %
Training	Young	3290	2142	2531	41.3%
	Middle Age	2591	2877	2564	35.8%
	Senior	2276	2200	3465	43.6%
	Overall %	34.1%	30.2%	35.8%	40.2%
Test	Young	823	584	629	40.4%
	Middle Age	671	649	647	33.0%
	Senior	565	621	871	42.3%
	Overall %	34.0%	30.6%	35.4%	38.7%

Growing Method: EXHAUSTIVE CHAID

Sample		Predicted			
		Young	Middle Age	Senior	Correct %
Training	Young	4123	1711	2096	52.0%
	Middle Age	3252	2531	2248	31.5%
	Senior	2895	2024	3135	38.9%
	Overall %	42.8%	26.1%	31.1%	40.8%
Test	Young	1036	481	552	50.1%
	Middle Age	814	592	562	30.1%
	Senior	743	468	733	37.7%
	Overall %	43.4%	25.8%	30.9%	39.5%

Growing Method: CRT

Sample		Predicted			
		Young	Middle Age	Senior	Correct %
Training	Young	2706	2907	2419	33.7%
	Middle Age	2079	3341	2544	42.0%
	Senior	2062	2654	3290	41.1%
	Overall %	28.5%	37.1%	34.4%	38.9%
Test	Young	692	699	576	35.2%
	Middle Age	516	863	656	42.4%
	Senior	505	658	829	41.6%
	Overall %	28.6%	37.0%	34.4%	39.8%

Growing Method: QUEST

Table 4-10: Classification Model for Years Attending

Sample		Predicted			
		Less Than 5 Ys	5 To 10 Ys	More than 10 Ys	Correct %
Training	Less Than 5	2161	849	5042	26.8%
	5 To 10 Ys	1816	867	5270	10.9%
	More than 10 Ys	1593	669	5710	71.6%
	Overall %	23.2%	9.9%	66.8%	36.4%
Test	Less Than 5	513	205	1229	26.3%
	5 To 10 Ys	487	201	1358	9.8%
	More than 10 Ys	410	180	1438	70.9%
	Overall %	23.4%	9.7%	66.8%	35.7%

Growing Method: CHAID

Sample		Predicted			
		Less Than 5 Ys	5 To 10 Ys	More than 10 Ys	Correct %
Training	Less Than 5	1727	1558	4678	21.7%
	5 To 10 Ys	1540	1637	4855	20.4%
	More than 10 Ys	1284	1408	5251	66.1%
	Overall %	19.0%	19.2%	61.8%	36.0%
Test	Less Than 5	477	420	1139	23.4%
	5 To 10 Ys	366	423	1178	21.5%
	More than 10 Ys	376	408	1273	61.9%
	Overall %	20.1%	20.6%	59.2%	35.9%

Growing Method: EXHAUSTIVE CHAID

Sample		Predicted			
		Less Than 5 Ys	5 To 10 Ys	More than 10 Ys	Correct %
Training	Less Than 5	3788	910	3232	47.8%
	5 To 10 Ys	3462	999	3570	12.4%
	More than 10 Ys	3217	708	4131	51.3%
	Overall %	43.6%	10.9%	45.5%	37.1%
Test	Less Than 5	948	238	883	45.8%
	5 To 10 Ys	853	207	908	10.5%
	More than 10 Ys	812	186	946	48.7%
	Overall %	43.7%	10.6%	45.8%	35.1%

Growing Method: CRT

Sample		Predicted			
		Less Than 5 Ys	5 To 10 Ys	More than 10 Ys	Correct %
Training	Less Than 5	2104	950	4978	26.2%
	5 To 10 Ys	1938	951	5075	11.9%
	More than 10 Ys	1682	853	5473	68.3%
	Overall %	23.8%	11.5%	64.7%	35.5%
Test	Less Than 5	542	224	1201	27.6%
	5 To 10 Ys	525	244	1266	12.0%
	More than 10 Ys	413	193	1386	69.6%
	Overall %	24.7%	11.0%	64.3%	36.2%

Growing Method: QUEST

Table 4-11: Classification Model for Usual Doctor

Sample		Predicted		
		Usual	Non-Usual	Correct %
Training	Usual	4913	3139	61.0%
	Non-Usual	3522	4431	55.7%
	Overall %	52.7%	47.3%	58.4%
Test	Usual	1191	756	61.2%
	Non-Usual	919	1127	55.1%
	Overall %	52.8%	47.2%	58.1%

Growing Method: CHAID

Sample		Predicted		
		Usual	Non-Usual	Correct %
Training	Usual	4530	3433	56.9%
	Non-Usual	3218	4814	59.9%
	Overall %	48.4%	51.6%	58.4%
Test	Usual	1102	934	54.1%
	Non-Usual	832	1135	57.7%
	Overall %	48.3%	51.7%	55.9%

Growing Method: EXHAUSTIVE CHAID

Sample		Predicted		
		Usual	Non-Usual	Correct %
Training	Usual	3982	3948	50.2%
	Non-Usual	2685	5346	66.6%
	Overall %	41.8%	58.2%	58.4%
Test	Usual	1001	1068	48.4%
	Non-Usual	701	1267	64.4%
	Overall %	42.2%	57.8%	56.2%

Growing Method: CRT

Sample		Predicted		
		Usual	Non-Usual	Correct %
Training	Usual	4445	3587	55.3%
	Non-Usual	3155	4809	60.4%
	Overall %	47.5%	52.5%	57.9%
Test	Usual	1059	908	53.8%
	Non-Usual	797	1238	60.8%
	Overall %	46.4%	53.6%	57.4%

Growing Method: QUEST

4.5 Discussion

This chapter presented an exploratory analysis of patients satisfaction feedback measured in a large scale survey through a non-probability sampling method. The IPQ dataset has more than 2.5 million response records and contain information for four patients sociodemographic factors. The exploratory analysis performed at the zero stratification level (entire population and all sociodemographic factors). The goal of the analysis is to investigate the possibility of a systematic stratification method to split patients into a homogenous sub-population. Parametric statistical tests such as analysis of variance in addition to supervised machine learning techniques like decision tree classification and principal component analysis are used to identify and model differences among sociodemographic characteristics.

Patients who are female, young, evaluating their non-usual doctor, and have been visiting the practice for less than five years gave systematically higher scores than other patients. Analysis of variance test indicated statistically significant differences among the 4 sociodemographic factors for all 27 items ($p \leq 0.05$). The large scale of IPQ dataset gave the opportunity to perform the analysis with different percentages (50%, 10%, 5%, 2%, and 1%) patients cases. The results showed that some demographic differences such as "evaluating doctors communication skills between females and males patients" can only appear statistically significant with large scale datasets. As the sample size get smaller, some differences start to appear insignificant for certain sociodemographic factors. For instance, gender differences became insignificant for doctor communication items when using a sample size that is about 10% of the original IPQ dataset while age differences became insignificant for clinic access and staff information items when using a sample size of 50% of the original IPQ dataset. Other sociodemographic factors like "Usual Doctor" remained significant for all items using different sample sizes. These results show that sociodemographic factors can have different effects in highlighting satisfaction feedback differences among multiple subpopulations. Therefore,

understanding the feedback profiles of smaller subgroups such as “Young Females” or “Middle-Aged Males visiting the Usual Doctor” requires a systematic stratification methodology to create homogeneous subpopulation groups. However, there is currently no clear methodology in the literature to guide researchers in identifying homogeneous groups.

Another common pattern of satisfaction surveys studies highlighted in this analysis is the patients tendency not answer all the required questionnaire items. The large volume of unanswered items can lead to misleading conclusions about patients feedback profiles. Removing all missing values would also lead to almost 50% loss of available sample size. However, it can be necessary to only include complete response cases to avoid having a significant number of patients considered as outlier. Chapter five of this thesis is dedicated to investigate missing answers patterns within IPQ sociodemographic factors.

supervised machine learning techniques were applied to investigate if sociodemographic characteristics can be predicted based on patients feedback profiles. The preliminary analysis showed a high level of inter-correlation among IPQ items with no clear candidate to act as a summative item. Therefore, a summative scale item with a set of uncorrelated survey components were added to the dataset and used as independent variables in a decision tree algorithm. Despite having statistically significant differences between sociodemographic subpopulations, the algorithm generated low accuracy models (Gender = 52.5%; Age = 38.5%; Years Attending = 35.7%; Usual GP = 58.1%). Examining the sociodemographic distributions using the overall summative scale revealed a large overlap between small subpopulations groups. The observed overlap among all sociodemographic factors at zero stratification level makes it very challenging to identify the feedback pattern of smaller sub-populations such as young-females or senior-males.

In summary, the relatively cheap cost of collecting large scale surveys dataset through non-probability sampling methods can provide opportunities for researchers to highlight feedback differences among smaller subpopulations. However, there are challenges about what is the best method to identify smaller subgroups. The analysis presented in this chapter also demonstrated that while IPQ dataset only contains four sociodemographic variables, the results show enough evidence to implement data stratification analysis. All results and data stratification techniques can be scaled for larger datasets with many sociodemographic variables. The next chapters of this thesis discuss the use of statistical and machine learning techniques to develop a systematic data stratification methodology.

Chapter 5 Missing Values Analysis

5.1 Introduction

The previous chapter has highlighted the difficulty of identifying patient's sociodemographic characteristics based on their feedback. The results showed a high level of overlap in satisfaction distribution among sociodemographic factors at zero stratification level. Another common feature of satisfaction datasets is that subjects do not always provide full answers in questionnaire forms. A large number of missing answers may change the distribution shape of different questionnaire items and overall scales. In addition to that, a substantial number of statistical analyses techniques requires the use of complete cases datasets and therefore, eliminate all subjects with any missing values from the analysis. The approach of using complete cases datasets can cause several disadvantages to the statistical analysis. One disadvantage is that it can significantly reduce sample size leading to the loss in information that is contained in the incomplete answers. Another disadvantage of removing missing answers is that it can lead to a larger amount of bias if certain sub-populations have a higher chance of having incomplete survey answers.

The purpose of this chapter is to investigate the effect of missing values on the distribution of IPQ items and whether certain sociodemographic factors are associated with higher missing answers rates. As the general tendency in healthcare satisfaction questionnaires is to give a high satisfaction feedback, the chapter investigates the hypothesis if missing answers are associated with lower feedbacks or a replacement for a low score for certain patients' sub-populations. For example, if patients want to express a satisfaction evaluation that is sub-optimal, they would prefer to leave out a score than provide a low one. Identifying these patterns can help to identify whether missing values are genuinely missing or contain another

low score effect that could be attributed to differences in sociodemographic groupings. Such knowledge would reveal if removing incomplete answers would leave statistical analysis techniques to work on biased datasets and if stratification analysis can be used to develop imputation strategies that are sensitive to sociodemographic differences. The knowledge would also help our understanding about the effect of missing response patterns especially for smaller subpopulation.

5.2 Missing Values Analysis

The statistical analysis at zero level involves the entire sample population (2,546,182 patients) with all four socio-demographic variables. A new item scale from 0 to 100 was created to summarize all 27 items into single overall value. Examining the overall scale revealed a negatively skewed distribution with majority of patients report a high satisfaction results. The skewness is similar to the original distribution shape of the 27 items. An exploratory analysis on IPQ 27 items shows a large number of missing values where patients did not provide full answers in questionnaire forms. The analysis revealed an average 5.2% of missing value across all 27 items with lowest of 1.4% for question 21 “The manner in which you are treated by the reception staff” and highest of 25% for question 27 “The practice's respect of your right to seek a second opinion or complementary medicine was”. The analysis also revealed that 742 patients did not answer any question. Table 5-1 shows items mean values with the number of valid answers and percentage of missing values; while Figure 5-1 shows the percentage of missing values.

The next step of analysis was to remove the missing answers to create a complete dataset of non-missing values of items and sociodemographic factors. The new dataset contains “1251357” cases which represent 50% of the original dataset. The analysis was repeated to remove cases from only the questions with a high percentage of null values (Qs 6, Qs 24, Qs 26, Qs 27). This

option created a dataset that contains 67% or “1709561” cases of the original dataset. Investigating the new smaller datasets highlighted that removing missing values would increase items mean values to different levels. The average increase in mean values was around 0.04 for all 27 items; however, the effect was more clear on clinic access and staff information questions with an average increase of 0.06 and 0.05 respectively. The practitioner’s communication questions (Qs9 – Qs20) showed a modest of 0.03 average increase. The same pattern was also reflected in the overall scale with the increased mean value from 73.2 to 77.9. The effect of missing values on the overall scale is noticeable on the left side of the distribution with the number of patients who are considered “outliers” due to not providing answers or to reporting the lowest feedback value of 1. Table 5-2 shows items mean values after removing messing values with an increased mean difference while Figure 5-2 Overall Scale Distribution with all patients (Left) and non-missing Patients (Right).

Table 5-1: Items Mean Values with the Number of Valid and Missing Answers

Items	Mean	No of Valid Answers	No of Missing	% of Missing
Qs01	3.60	2487539	58643	2.30
Qs02	3.40	2491304	54878	2.15
Qs03	3.65	2495582	50600	1.98
Qs04	3.47	2471493	74689	2.93
Qs05	3.26	2439991	106191	4.17
Qs06	3.31	2230348	315834	12.40
Qs07	3.60	2498718	47464	1.86
Qs08	3.20	2445860	100322	3.94
Qs09	4.18	2488031	58151	2.28
Qs10	4.23	2487155	59027	2.31
Qs11	4.25	2478202	67980	2.66
Qs12	4.18	2475948	70234	2.75
Qs13	4.13	2468541	77641	3.04
Qs14	4.27	2479558	66624	2.61
Qs15	4.17	2466781	79401	3.11
Qs16	4.33	2479840	66342	2.60
Qs17	3.91	2439126	107056	4.20
Qs18	4.09	2404854	141328	5.55
Qs19	4.12	2415940	130242	5.11
Qs20	4.21	2420164	126018	4.94
Qs21	3.99	2509129	37053	1.45
Qs22	3.97	2477533	68649	2.69
Qs23	3.83	2418140	128042	5.02
Qs24	3.57	2184651	361531	14.19
Qs25	3.73	2336693	209489	8.22
Qs26	3.63	2270202	275980	10.83
Qs27	3.64	1914084	632098	24.82

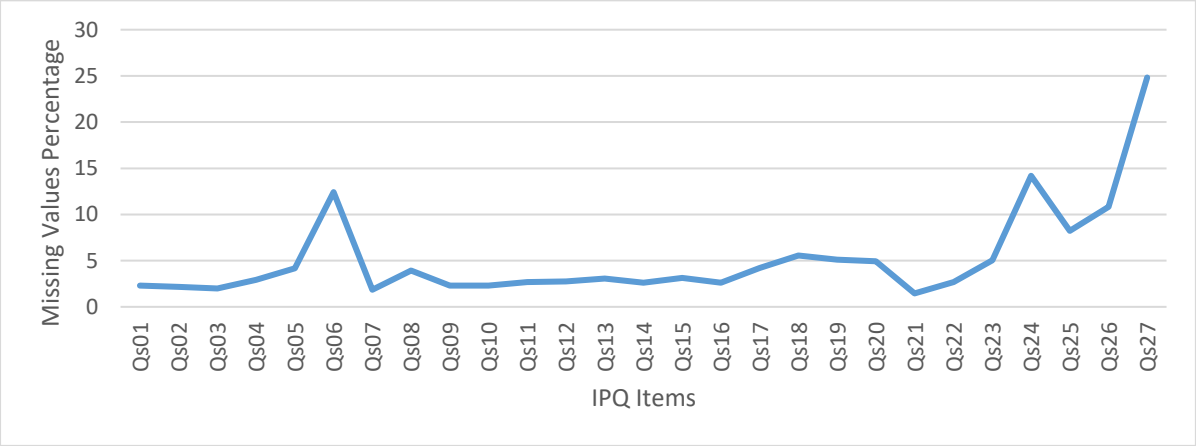

Figure 5-1: Percentage of Missing Answers

Table 5-2: Mean Values After Removing Missing Answers

Variable	Mean Before	Mean After	Mean Difference	No of Missing	% of Missing
Qs01	3.60	3.67	0.06	58643	2.30
Qs02	3.40	3.47	0.06	54878	2.15
Qs03	3.65	3.71	0.05	50600	1.98
Qs04	3.47	3.53	0.05	74689	2.93
Qs05	3.26	3.35	0.08	106191	4.17
Qs06	3.31	3.35	0.04	315834	12.40
Qs07	3.60	3.6	0.05	47464	1.86
Qs08	3.20	3.26	0.06	100322	3.94
Qs09	4.18	4.20	0.02	58151	2.28
Qs10	4.23	4.26	0.02	59027	2.31
Qs11	4.25	4.27	0.02	67980	2.66
Qs12	4.18	4.21	0.02	70234	2.75
Qs13	4.13	4.16	0.02	77641	3.04
Qs14	4.27	4.30	0.03	66624	2.61
Qs15	4.17	4.20	0.02	79401	3.11
Qs16	4.33	4.35	0.02	66342	2.60
Qs17	3.91	3.9	0.03	107056	4.20
Qs18	4.09	4.11	0.02	141328	5.55
Qs19	4.12	4.15	0.02	130242	5.11
Qs20	4.21	4.24	0.02	126018	4.94
Qs21	3.99	4.04	0.04	37053	1.45
Qs22	3.97	4.03	0.05	68649	2.69
Qs23	3.83	3.89	0.06	128042	5.02
Qs24	3.57	3.62	0.05	361531	14.19
Qs25	3.73	3.79	0.05	209489	8.22
Qs26	3.63	3.69	0.05	275980	10.83
Qs27	3.64	3.66	0.01	632098	24.82

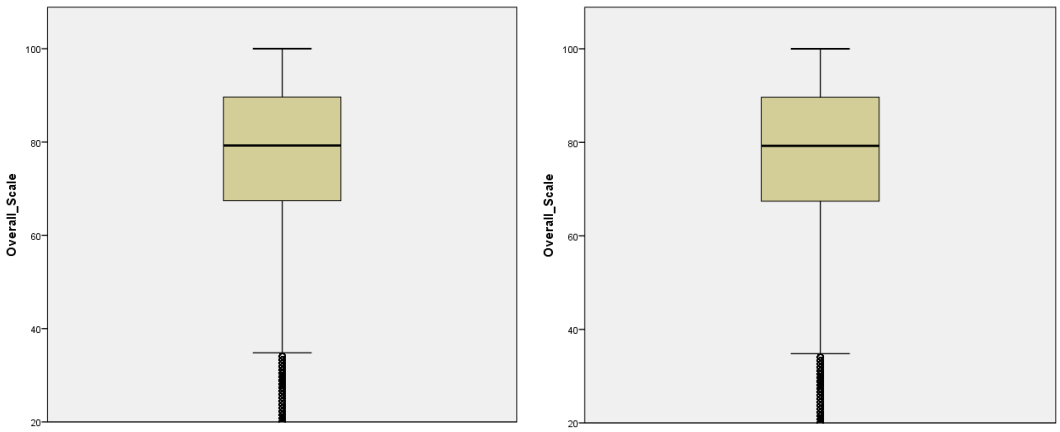

Figure 5-2: Overall Scale Distribution with all patients (Left) and non-missing Patients (Right)

Analysing missing answers can reveal patterns about how certain sociodemographic subgroups provide uncompleted answers to satisfaction questioners. The following tables 3 - 6 compares missing values percentage among different sociodemographic subgroups.

Table 5-3: Missing Values Comparison by Gender

	Overall		Females		Males	
Variable	**No of Missing**	**% of Missing**	**No of Missing**	**% of Missing**	**No of Missing**	**% of Missing**
Qs01	58643	2.30	35190	2.23	16760	1.95
Qs02	54878	2.15	30707	1.94	17453	2.04
Qs03	50600	1.98	29633	1.87	14752	1.72
Qs04	74689	2.93	42758	2.71	24314	2.84
Qs05	106191	4.17	62139	3.93	34635	4.04
Qs06	315834	12.42	185611	11.76	112430	13.14
Qs07	47464	1.86	26878	1.70	13571	1.58
Qs08	100322	3.94	62548	3.96	28378	3.317
Qs09	58151	2.28	34396	2.18	15643	1.82
Qs10	59027	2.31	33650	2.13	16198	1.89
Qs11	67980	2.66	38817	2.46	19139	2.23
Qs12	70234	2.75	40625	2.57	19423	2.27
Qs13	77641	3.04	44884	2.84	21880	2.55
Qs14	66624	2.61	38403	2.43	18133	2.12
Qs15	79401	3.11	45969	2.91	22469	2.62
Qs16	66342	2.60	38348	2.43	17586	2.05
Qs17	107056	4.20	66015	4.18	27056	3.16
Qs18	141328	5.55	87057	5.51	37532	4.38
Qs19	130242	5.11	79312	5.02	34239	4.00
Qs20	126018	4.94	75892	4.81	33395	3.90
Qs21	37053	1.45	18090	1.14	8655	1.01
Qs22	68649	2.69	37358	2.36	18507	2.16
Qs23	128042	5.02	73649	4.66	37789	4.41
Qs24	361531	14.19	230374	14.60	103234	12.04
Qs25	209489	8.22	128817	8.16	56864	6.64
Qs26	275980	10.83	169174	10.72	80872	9.45
Qs27	632098	24.82	410847	26.04	183614	21.46

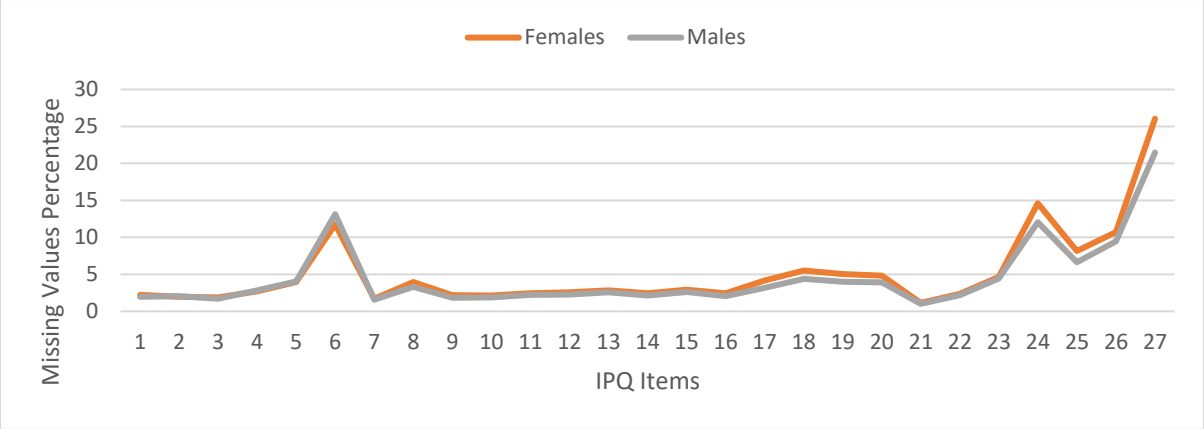

Figure 5-3: Missing Values Percentage by Gender

Table 5-4: Missing Values Comparison by Age Group

Variable	Overall		Young		Middle Age		Senior	
	# Missing	% of Missing	# Missing	% of Missing	# Missing	% of Missing	# missing	% of Missing
Qs01	58643	2.30	4363	1.855	25767	1.91	21231	2.59
Qs02	54878	2.16	3946	1.678	21682	1.61	21601	2.64
Qs03	50600	1.99	2971	1.263	21058	1.56	19181	2.34
Qs04	74689	2.93	4794	2.039	31707	2.35	29244	3.57
Qs05	106191	4.17	8951	3.806	52947	3.92	33092	4.04
Qs06	315834	12.40	25443	10.820	157851	11.70	111404	13.60
Qs07	47464	1.86	3150	1.340	18740	1.39	17513	2.14
Qs08	100322	3.94	8211	3.492	52181	3.87	28692	3.50
Qs09	58151	2.28	5105	2.171	24133	1.79	19810	2.42
Qs10	59027	2.32	5537	2.355	24922	1.85	18412	2.25
Qs11	67980	2.67	5933	2.523	28313	2.10	22617	2.76
Qs12	70234	2.76	6036	2.567	28796	2.13	24205	2.96
Qs13	77641	3.05	6500	2.764	32239	2.39	26629	3.25
Qs14	66624	2.62	5862	2.493	27690	2.05	21915	2.68
Qs15	79401	3.12	6618	2.814	32401	2.40	28131	3.43
Qs16	66342	2.61	6226	2.648	27927	2.07	20838	2.54
Qs17	107056	4.20	10085	4.289	50442	3.74	31282	3.82
Qs18	141328	5.55	12018	5.111	65029	4.82	45527	5.56
Qs19	130242	5.12	11533	4.904	60835	4.51	39289	4.80
Qs20	126018	4.95	11472	4.878	59468	4.41	36313	4.43
Qs21	37053	1.46	2410	1.025	12798	0.95	10908	1.33
Qs22	68649	2.70	3753	1.596	26345	1.95	24286	2.97
Qs23	128042	5.03	9918	4.218	62726	4.65	36172	4.42
Qs24	361531	14.20	19322	8.217	179475	13.30	128918	15.74
Qs25	209489	8.23	11283	4.798	93494	6.93	76420	9.33
Qs26	275980	10.84	20394	8.673	143443	10.63	81037	9.89
Qs27	632098	24.83	36866	15.677	330239	24.48	219432	26.79

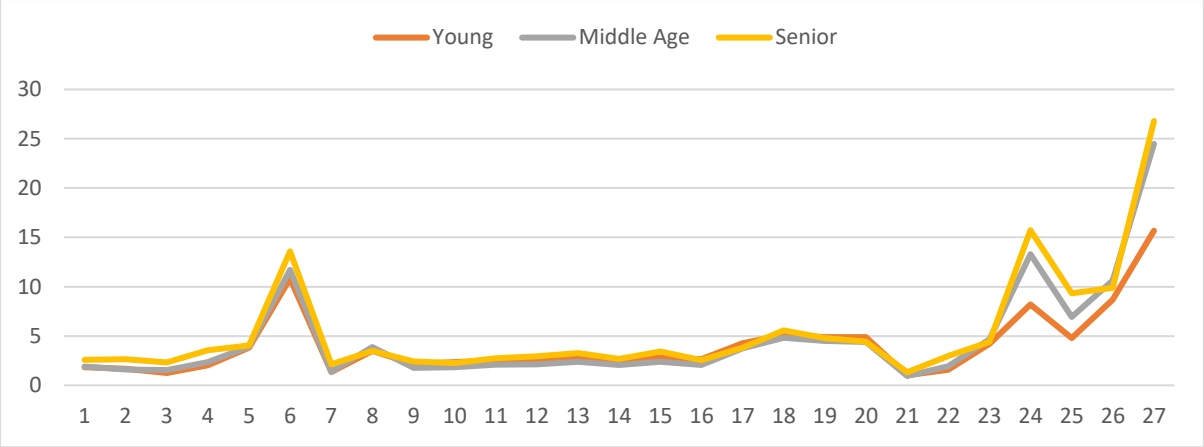

Figure 5-4: Missing Values Percentage by Age Group

Table 5-5: Missing Values Comparison by Usual Doctor Group

	Overall		Usual Doctor		Non - Usual Doctor	
Variable	**No of Missing**	**% of Missing**	**No of Missing**	**% of Missing**	**No of Missing**	**% of Missing**
Qs01	58643	2.30	32275	2.01	14782	2.14
Qs02	54878	2.16	30670	1.91	12362	1.79
Qs03	50600	1.99	28923	1.80	11613	1.68
Qs04	74689	2.93	41963	2.61	16228	2.35
Qs05	106191	4.17	48642	3.02	30010	4.34
Qs06	315834	12.40	175423	10.90	89809	13.00
Qs07	47464	1.86	26240	1.63	10426	1.51
Qs08	100322	3.94	50071	3.11	30290	4.38
Qs09	58151	2.28	25870	1.61	17438	2.52
Qs10	59027	2.32	25569	1.59	17332	2.51
Qs11	67980	2.67	30174	1.88	19667	2.85
Qs12	70234	2.76	31200	1.94	20325	2.94
Qs13	77641	3.05	35908	2.23	21772	3.15
Qs14	66624	2.62	29179	1.81	19499	2.82
Qs15	79401	3.12	35228	2.19	23191	3.36
Qs16	66342	2.61	28518	1.77	19523	2.83
Qs17	107056	4.20	48116	2.99	34072	4.93
Qs18	141328	5.55	64977	4.04	44272	6.41
Qs19	130242	5.12	59655	3.71	40589	5.87
Qs20	126018	4.95	57098	3.55	39676	5.74
Qs21	37053	1.46	15877	0.99	7701	1.11
Qs22	68649	2.70	31288	1.94	16739	2.42
Qs23	128042	5.03	58087	3.61	37443	5.42
Qs24	361531	14.20	196677	12.23	101567	14.70
Qs25	209489	8.23	105216	6.54	57388	8.31
Qs26	275980	10.84	137832	8.57	80927	11.71
Qs27	632098	24.83	353108	21.95	185417	26.83

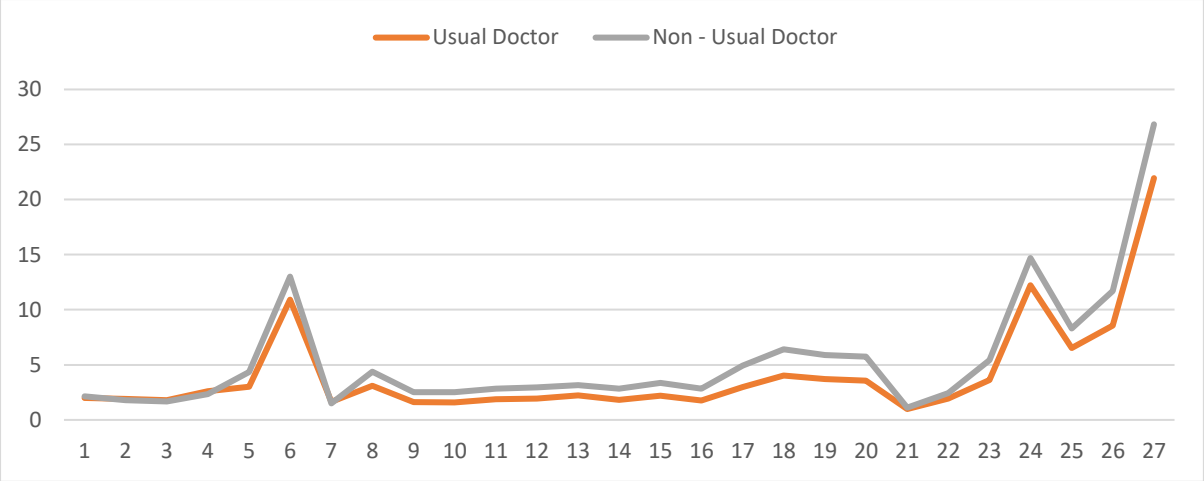

Figure 5-5: Missing Values Percentage by Usual Doctor Group

Table 5-6: Missing Values Comparison by Years Attending Group

Variable	Overall		Less than 5 Years		5 to 10 Years		More than 10 Years	
	# Missing	% of Missing	# Missing	% of Missing	# Missing	% of Missing	# Missing	% of Missing
Qs01	58643	2.30	11589	2.12	8478	2.00	31457	2.15
Qs02	54878	2.16	11970	2.19	7437	1.75	28212	1.93
Qs03	50600	1.99	9019	1.65	7263	1.71	27730	1.90
Qs04	74689	2.93	19046	3.49	9668	2.28	37549	2.57
Qs05	106191	4.17	33938	6.22	14466	3.40	47300	3.24
Qs06	315834	12.40	89286	16.36	45610	10.73	162615	11.14
Qs07	47464	1.86	9012	1.65	6403	1.51	24411	1.67
Qs08	100322	3.94	21762	3.99	14982	3.53	53153	3.64
Qs09	58151	2.28	11501	2.11	8052	1.90	30069	2.06
Qs10	59027	2.32	11987	2.20	8229	1.94	29013	1.99
Qs11	67980	2.67	13676	2.51	9392	2.21	34200	2.34
Qs12	70234	2.76	13836	2.53	9568	2.25	35882	2.46
Qs13	77641	3.05	15694	2.88	10907	2.57	39365	2.70
Qs14	66624	2.62	13446	2.46	9127	2.15	33195	2.27
Qs15	79401	3.12	15920	2.92	10911	2.57	40856	2.80
Qs16	66342	2.61	13347	2.45	9217	2.17	32511	2.23
Qs17	107056	4.20	21443	3.93	15734	3.70	55362	3.79
Qs18	141328	5.55	28599	5.24	20822	4.90	74662	5.11
Qs19	130242	5.12	26189	4.80	19097	4.49	67353	4.61
Qs20	126018	4.95	25413	4.66	18606	4.38	64328	4.41
Qs21	37053	1.46	5646	1.03	4507	1.06	15964	1.09
Qs22	68649	2.70	13008	2.38	8756	2.06	33173	2.27
Qs23	128042	5.03	30204	5.53	17971	4.23	61794	4.23
Qs24	361531	14.20	80657	14.78	53476	12.59	198789	13.62
Qs25	209489	8.23	47829	8.76	28703	6.76	107682	7.38
Qs26	275980	10.84	72108	13.21	40214	9.46	135973	9.31
Qs27	632098	24.83	144239	26.42	97359	22.91	352866	24.17

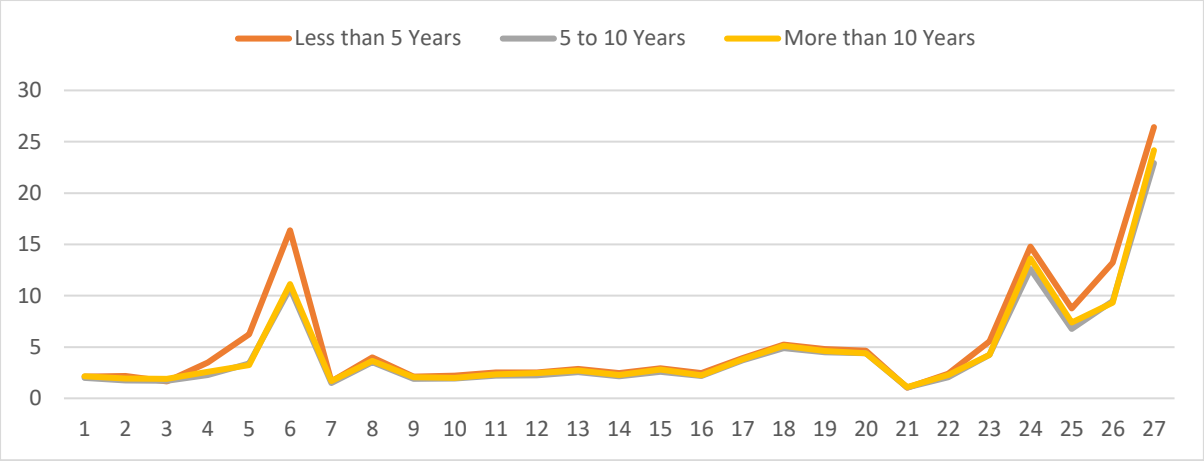

Figure 5-6: Missing Values Percentage by Years Attending Group

5.3 Sociodemographic Evaluation Profiles

This section introduces the analysis of missing answers within each sociodemographic factor. For each factor, the dataset is split into two subsets (complete and incomplete answers). An expletory analysis and ANOVA test were used to identify if items have a statistically significant difference after removing all incomplete answers. Figure 5-7 shows the process for factor "Gender" while tables 7-9 compare mean values with and without incomplete answers.

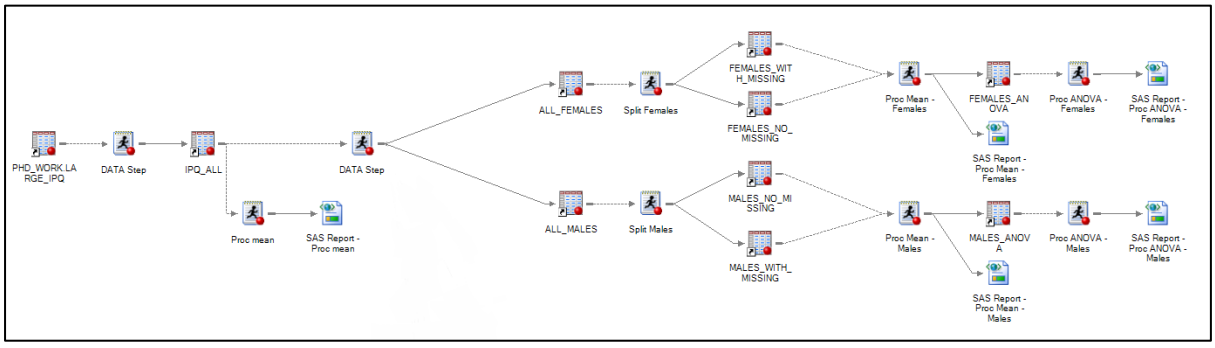

Figure 5-7: Missing Values Analysis for Gender

Table 5-7: Mean Values Comparison by Gender

Variable	Overall Mean	Non Missing Cases	
		Females Mean N = 857687	Males Mean N = 507996
Qs01	3.61	3.63	3.74
Qs02	3.40	3.42	3.56
Qs03	3.66	3.64	3.81
Qs04	3.48	3.49	3.60
Qs05	3.27	3.27	3.48
Qs06	3.31	3.34	3.38
Qs07	3.60	3.62	3.71
Qs08	3.20	3.19	3.39
Qs09	4.18	4.18	4.23
Qs10	4.23	4.24	4.27
Qs11	4.25	4.25	4.28
Qs12	4.19	4.19	4.22
Qs13	4.14	4.15	4.17
Qs14	4.27	4.27	4.32
Qs15	4.17	4.18	4.20
Qs16	4.33	4.33	4.37
Qs17	3.91	3.92	3.96
Qs18	4.09	4.09	4.13
Qs19	4.13	4.13	4.17
Qs20	4.22	4.22	4.26
Qs21	3.99	4.00	4.10
Qs22	3.97	3.98	4.10
Qs23	3.83	3.86	3.96
Qs24	3.58	3.59	3.70
Qs25	3.73	3.75	3.85
Qs26	3.64	3.66	3.75
Qs27	3.64	3.62	3.72
Scale	73.19	77.19	78.83

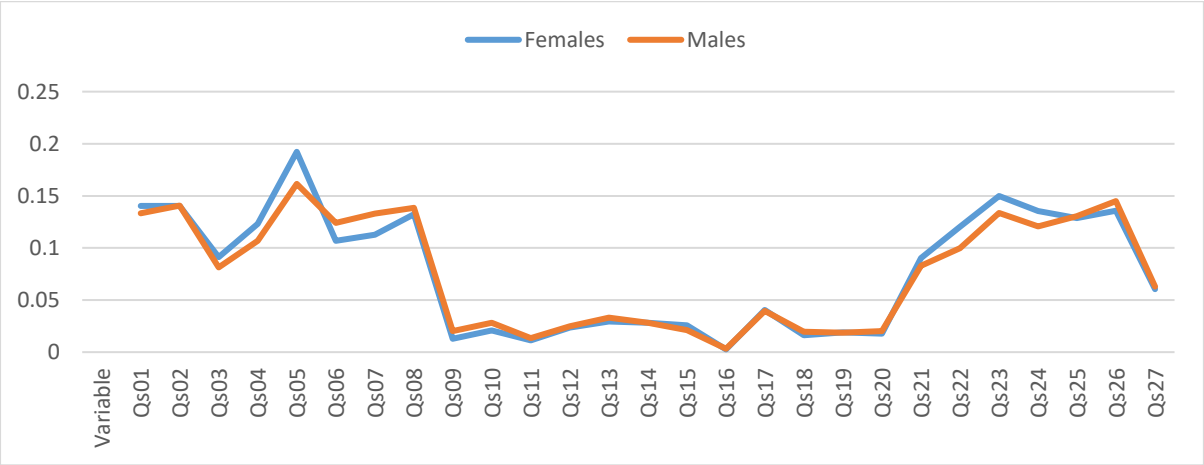

Figure 5-8: Compare Before and After Removing Missing Values for Gender Groups

Females patients gave an average lower score of -0.02 than the overall IPQ mean values while male patients gave an average higher score of 0.06 than the overall IPQ mean values. Comparing the groups of missing and non-missing values revealed that subjects with complete answers gave higher average scores of 0.07 than subjects with missing values. The difference for clinic access and staff information items were considerably higher with almost 0.13 and 0.12 respectively. All but one item (Qs16 “The respect shown to me by doctor”) showed the statistically significant difference between missing and non-missing groups for both females and males patients. Figure 5-8 shows the difference between missing and non-missing answers for females and males patients.

Table 5-8: Mean Values Comparison by Age Groups

Variable	Overall Mean	Non- Missing Cases		
		Young Mean N = 153933	**Middle Mean N = 773476**	**Senior Mean N = 422553**
Qs01	3.61	3.58	3.61	3.83
Qs02	3.40	3.42	3.42	3.60
Qs03	3.66	3.61	3.63	3.89
Qs04	3.48	3.46	3.48	3.68
Qs05	3.27	3.23	3.23	3.61
Qs06	3.31	3.18	3.29	3.54
Qs07	3.60	3.52	3.58	3.85
Qs08	3.20	3.08	3.18	3.49
Qs09	4.18	4.01	4.14	4.38
Qs10	4.23	4.07	4.20	4.42
Qs11	4.25	4.10	4.21	4.43
Qs12	4.19	4.04	4.16	4.35
Qs13	4.14	3.98	4.11	4.32
Qs14	4.27	4.13	4.24	4.45
Qs15	4.17	4.03	4.15	4.34
Qs16	4.33	4.18	4.29	4.51
Qs17	3.91	3.75	3.88	4.11
Qs18	4.09	3.92	4.06	4.28
Qs19	4.13	3.94	4.09	4.33
Qs20	4.22	4.04	4.18	4.41
Qs21	3.99	3.93	3.98	4.20
Qs22	3.97	3.98	3.96	4.18
Qs23	3.83	3.76	3.81	4.11
Qs24	3.58	3.51	3.55	3.83
Qs25	3.73	3.70	3.71	3.98
Qs26	3.64	3.53	3.60	3.93
Qs27	3.64	3.54	3.57	3.87
Scale	73.19	74.97	76.53	81.41

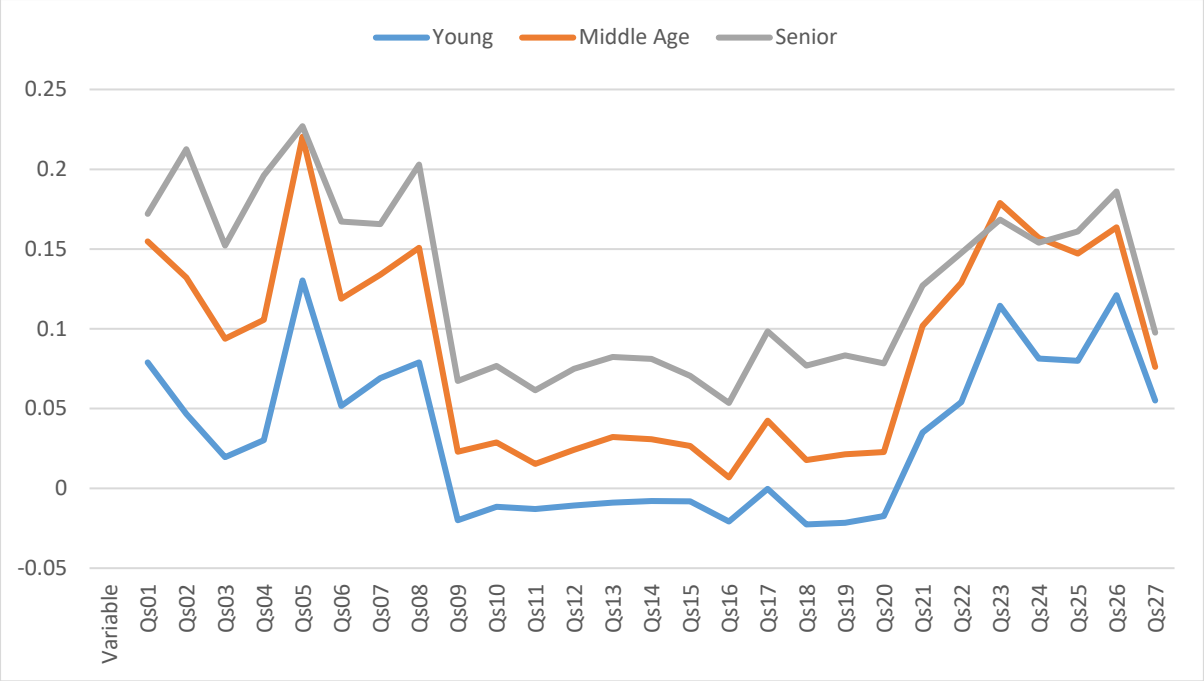

Figure 5-9: Compare Before and After Removing Missing Values Age Groups

Young and middle-aged patients gave average lower scores of -0.11 and -0.06 than the overall IPQ mean values while senior patients gave average higher scores of 0.16 than the overall IPQ mean values. For young patients, the average lower score of doctor communication items was -0.15 which is more than clinic access and staff information items. Comparing the groups of missing and non-missing values revealed that subjects with complete answers gave higher average scores than subjects with missing answers except for young patients whose missing value group were on average lower than non-missing values. The ANOVA test shows statistically significant differences between missing and non-missing groups except items (14, 15, 17) for young patients. This pattern indicates that young patients, at least in some aspects, appear to be more critical about their communication experience with physicians than any other sub-population. Figure 5-9 shows the difference between missing and non-missing answers for young, middle-aged and senior patients.

Table 5-9: Mean Values Comparison by Usual Doctor

		Non Missing Cases	
Variable	**Overall Mean**	**Usual Mean**	**Non-Usual Mean**
Qs01	3.61	3.74	3.52
Qs02	3.40	3.55	3.28
Qs03	3.66	3.78	3.53
Qs04	3.48	3.59	3.39
Qs05	3.27	3.52	2.92
Qs06	3.31	3.44	3.13
Qs07	3.60	3.71	3.53
Qs08	3.20	3.34	3.06
Qs09	4.18	4.29	3.99
Qs10	4.23	4.34	4.05
Qs11	4.25	4.35	4.07
Qs12	4.19	4.28	4.02
Qs13	4.14	4.24	3.96
Qs14	4.27	4.38	4.07
Qs15	4.17	4.28	3.99
Qs16	4.33	4.42	4.15
Qs17	3.91	4.02	3.73
Qs18	4.09	4.20	3.90
Qs19	4.13	4.24	3.92
Qs20	4.22	4.33	4.00
Qs21	3.99	4.09	3.90
Qs22	3.97	4.09	3.88
Qs23	3.83	3.97	3.71
Qs24	3.58	3.70	3.44
Qs25	3.73	3.86	3.61
Qs26	3.64	3.78	3.48
Qs27	3.64	3.74	3.45
Scale	73.19	79.47	73.84

Usual patients gave average higher scores of 0.09 than the overall IPQ mean values while non-usual patients gave average lower scores of -0.17 than the overall IPQ mean values. Comparing the groups of missing and non-missing values revealed that subjects with complete answers gave higher average scores than subjects with missing answers except for non-usual patients who's their missing value group were higher than non-missing values for all doctor communication items. The ANOVA test shows statistically significant differences between missing and non-missing groups in all items for both usual and non-usual groups.

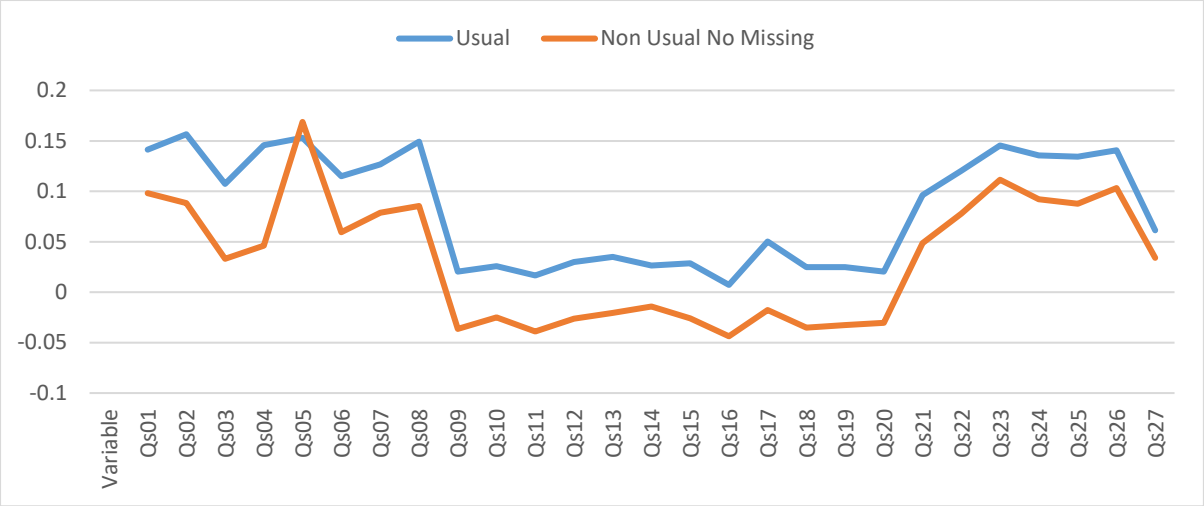

Figure 5-10: Compare Before and After Removing Missing Values for Usual Doctor

Finally, patients who have been seeing their doctor for less than 10 years gave lower scores than the overall IPQ mean values for most of doctor communication and staff information items. However, comparing the groups of missing and non-missing values did not indicate a clear evaluation pattern. Also, the ANOVA test showed a mix of statistically significant and insignificant differences between items within the same component. For patients who have been seeing their doctor for more than 10 years, patients with complete answers gave higher average scores than subjects with missing answers with statistically significant differences across all items.

The results above showed clear patterns about patients behaviour to providing incomplete answers by different sociodemographic sub-groups. The shapes for missing answers percentages were mostly consistent among each sociodemographic sub-groups with few exceptions. Most noticeably, the analysis revealed that young patients usually provide the highest percentage of complete answers despite being the most critical subpopulation. Several data mining algorithms were implemented to extract rules about patients missing values

patterns. A decision tree model using J48 algorithm to predict the type of IPQ items (Clinic Access, Doctor Communication, Staff Information) based on missing values patterns. The result showed that first split of the tree is to check if the level of young missing values is below a certain level. Figure 5-11 shows missing values decision tree model

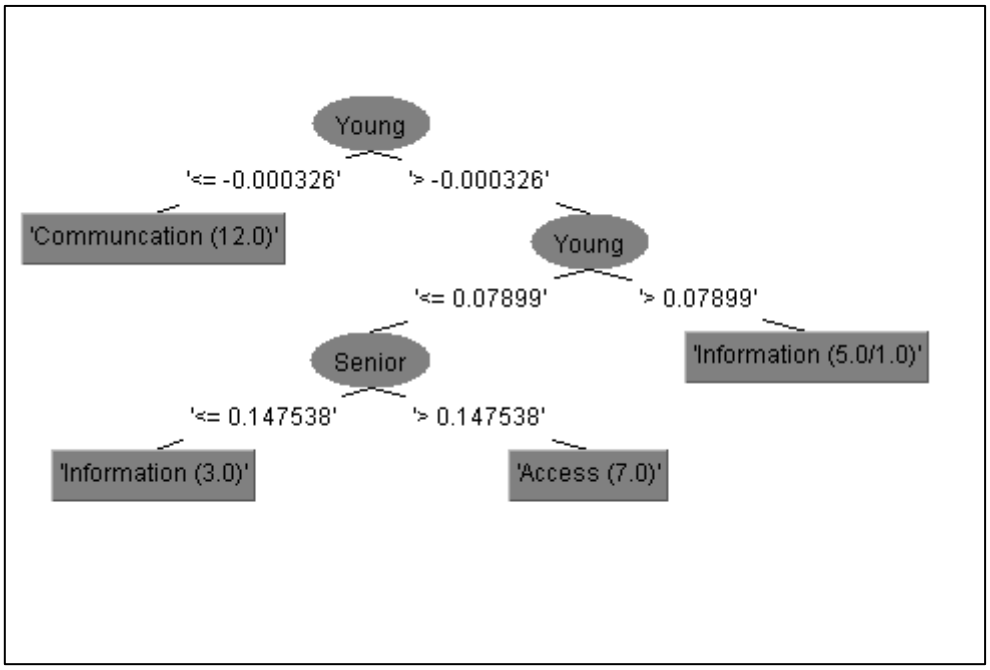

Figure 5-11: Predicting Item Type from Missing Answers Patterns

5.4 Discussion

Providing an uncompleted set of answers is a common challenge in analysing questioners and subjects feedback datasets. Many statistical analysis techniques exclude subjects with any missing variable values from the dataset. Excluding subjects with missing values can lead to a significant reduction in sample size and may ignore the possible systematic difference between the complete cases and incomplete cases. This chapter investigates the effect of incomplete satisfaction feedback on the evaluation profile of different patients sub-population. The analysis investigate if patients prefer to ignore answering some evaluation items as an alternative to provide lower scores.

An exploratory analysis revealed that the average missing values among all 27 items is 5.2% (min = 1.45% item 21, max = 24.8% item 27). Although there is a considerable variation in the rate of missing answers, filtering all unanswered items can lead to 51% loss of the total cases. Adjusting the filter to only include the top 4 missing values items can increase the valid cases to 67% of the original dataset. Analysing the missing values rate by subgroups revealed that young patients provided more complete answers with “4.07%” missing rate in comparison to 4.64% and 5.4% for middle-age and senior patients respectively. Also, patients seeing their usual doctor gave more complete answers that non-usual doctor group. However, a decision about filtering unanswered items or replacing missing data with a generalised value should consider the effect on subpopulation evaluation profiles. Removing all unanswered cases was associated with 0.04 increase in all 27 items mean values. The effect was more clear on clinic access and staff information questions with average mean increase of 0.06 and 0.05 respectively. For doctor communication items, the overall average increase was 0.027. The analysis was then repeated to examine the effect of removing missing answers within different sociodemographic sub-populations.

The analysis revealed several new insights about the feedback behaviour of different sociodemographic sub-populations. Examining missing values rates show that young patients are the most engaged group compare to other sup-populations. In addition to that, removing incomplete answers was generally associated with increase in the overall mean values. This increase was much higher for clinic access and staff information items in comparison to doctor communication items with few exceptions. For both young and non-usual sub-populations, doctor communication items were higher with missing values datasets. This can be interpreted as patients are much more reluctant to criticise their personal doctor in comparison to other service aspects.

Comparing the missing and non-missing datasets showed that most items have statistically significant differences which may imply that removing incomplete cases from the original sample will have a potential bias for statistical procedures that require a full set of scores. Therefore, the missing values analysis can lead to an evidence-based imputation of missing value that take into account sociodemographic profile and would allow to include all raters in the analysis. However, the analysis also revealed that a single item may appear with or without statistical significance based on the sociodemographic factor that is being analysed. Therefore, a stratification analysis approach is recommended to ensure the significance of different items across multiple mutually exclusive sub-groups. The stratification analysis would also identify potential average score values that can be used for an evidence-based imputation method. The next chapter will introduce a top-down stratification strategy to guide researchers in generating mutually exclusive sub-populations.

Chapter 6 Data Stratification

6.1 Introduction

Chapter four of this thesis presented an exploratory analysis of IPQ dataset at zero stratification level where the entire sample population with all sociodemographic factors are included. The analysis showed a high level of overlap in satisfaction distribution among sociodemographic factors at zero stratification level. Experiments using supervised machine learning algorithms highlighted the difficulty of identifying patient's sociodemographic despite having statistically significant differences between subgroups. The exploratory analysis highlighted the need to use a systematic data stratification methodology that can guide researchers in investigating the feedback profiles of smaller subpopulations such as young males vs. middle-age females visiting the usual doctor. Another feature of convenient sampling datasets is patients tend to not provide answers to all questionnaire items. Chapter five presented insights about how certain sociodemographic sub-populations are more likely to provide uncompleted answers. The chapter provided the statistical analysis and justification to use a smaller subset of IPQ dataset with non-missing values for all 27 items and socio-demographic information. The new dataset contains (1,251,357 patients).

This chapter investigates the possibility of implementing a systematic data stratification methodology for sociodemographic factors by controlling the amount of variance change in each sub-group. The methodology is designed to help researchers to create mutually-exclusive groups by minimising or maximising differences between sub-populations. A combination of machine learning and statistical modelling techniques including linear regression and information gain are used to split the data into smaller subgroups. The different techniques are designed to provide researchers with the flexibility to choose numerical or ordinal data types

as data stratification guide. For instance, if a questionnaire has a numerical summative item in the form of numerical scale, researchers can use linear regression modelling to guide the stratification process. In the case no explicit numerical summative variable is available, researchers may use information gain to split the row data into smaller homogeneous subgroups using an ordinal scale. Section 6.2 of this chapter explains the process of entropy-based and regression-based stratification methodology while section 6.3 presents discussions and guidelines for researchers about the advantages and disadvantages of data stratification.

6.2 Stratification Process

Previous analysis in this thesis identified a high level of intercorrelation among all IPQ items with no single question to qualify as a summative item. Therefore, a summative "overall scale" item was constructed as an equal weight average of all 27 items. The constructed item has the same moderate left skewed distribution of its underline items. The overall scale has (mean=77.95, Std=14.5) as a global statistics and irrespective of any patients sub-group. In addition to that, PCA was used to create three uncorrelated components (clinic access, doctor communication, and staff information) that represent the underline structure of IPQ dataset. Further PCA analysis showed that all sociodemographic subgroups confirmed the three components structure of IPQ dataset with "doctor communication" as the most important factor. The overall scale factor was used to guide the stratification process

A new ordinal summative variable was calculated by converting the numerical overall scale into standard deviation classes (SD1 – SD3). The number of cases in each class is similar to a normal distribution with probabilities of (0.62924, 0.340725, and 0.030035) for SD1, SD2 and SD3 respectively. At the zero stratification level, the information gain and linear regression algorithm were run on the ordinal and numerical overall scale items respectively. In addition to that, the ANOVA and Tukey post-hoc tests were calculated for the numerical overall scale

to identify statistically significant differences among sociodemographic subgroups. The results show that “Usual GP” factor achieved the highest information gain and regression coefficient values at the zero stratification level. The ANOVA test identified a statistically significant difference between the two patients subgroups with an average overall score of 79.52 and 73.88 for “usual doctor” and “non-usual doctor” respectively. The ANOVA test also identified a large effect size (0.063) among the different sociodemographic subgroups using the Partial Eta Squared value [144]. The large effect size reflects a significant variance magnitude within IPQ sociodemographic subgroups. Table 6-1 shows the information gain results for the ordinal summative variable at the zero stratification level while Table 6-2 shows the regression coefficient values for the numerical “overall scale” variable, while Table 6-3 shows ANOVA and effect size results at stratification level zero. Table 6-4 and Table 6-5 shows the number of cases and mean values for each of the sociodemographic subgroups.

Table 6-1: Information Gain Values

Levels	Sociodemographic Factors	Information Gain
Level Zero	Usual GP	0.054358648
	Age	0.003817677
	Gender	0.000408429
	Years Attending	0.000370313
Level One Usual	Gender	0.000446787
	Age Group	0.003647209
	Years Attending	0.000563325
Level One Non-Usual	Gender	0.000474878
	Age Group	0.002367893
	Years Attending	0.000112913
Level Two Usual Young	Gender	0.000693973
	Years Attending	0.000230908
Level Two Usual Middle-Age	Gender	0.000072006
	Years Attending	0.000191007
Level Two Usual Senior	Gender	0.000333369
	Years Attending	0.000109958
Level Two Non-Usual Young	Gender	0.000591411
	Years Attending	0.000162017
Level Two Non-Usual Middle	Gender	0.000315342
	Years Attending	0.000108294
Level Two Non usual Senior	Gender	0.000107324
	Years Attending	0.000097230

Table 6-2: Regression Coefficient Values

Stratification Level	Sociodemographic	Regression Coefficients	Sig.
Level 0	Usual GP	-4.983	0.000
	Gender	.915	.000
	Age	3.055	0.000
	Attending	-.077	.001
Level One Usual	Gender	.884	.000
	Attending	.083	.002
	Age	3.327	0.000
Level One Non-Usual	Gender	.985	.000
	Attending	-.452	.000
	Age	2.328	0.000
Level Two Usual Young	Gender	1.298	.000
	Attending	-.071	.371
Level Two Usual Middle-Age	Gender	.598	.000
	Attending	.225	.000
Level2 Usual Senior	Gender	1.120	.000
	Attending	-.378	.000
Level Two Non-Usual Young	Gender	1.135	.000
	Attending	-.691	.000
Level Two Non-Usual Middle-Age	Gender	.706	.000
	Attending	-.544	.000
Level Two Non-Usual Senior	Gender	1.290	.000
	Attending	-.456	.000

Table 6-3: Analysis of Variance and Effect Size

Source	Degrees of Freedom	Mean Square	Sig.	Partial Eta Squared
Corrected Model	35	396859.69	0	0.063
Intercept	1	2461749036	0	0.908
stratum	35	396859.69	0	0.063

Table 6-4: Mean Score Values for Stratification Levels Zero and One

Level	Subpopulations		N	Mean
Level Zero	Usual GP	Usual GP	431831	79.52
		Non-Usual GP	168169	73.88
	Age Group	Young	67974	75.05
		Middle	344338	76.6
		Senior	187688	81.45
	Gender	Females	376633	77.34
		Males	223367	78.95
	Years Attending	Less than 5	128572	77.27
		5 to 10 Years	107444	77.03
		More than 10 Years	363984	78.44
Level One Usual	Age Group	Young	40224	76.4914
		Middle	239624	78.1829
		Senior	151983	82.4334
	Gender	Females	268028	78.9456
		Males	163803	80.4633
	Years Attending	Less than 5	85831	78.7401 *
		5 to 10 Years	75812	78.6979 *
		More than 10 Years	270188	80.0005
Level One Non-Usual	Age Group	Young	27750	72.95634 *
		Middle	104714	72.97726 *
		Senior	35705	77.24
	Gender	Females	108605	73.39
		Males	59564	74.77
	Years Attending	Less than 5	42741	74.33
		5 to 10 Years	31632	73.05
		More than 10 Years	93796	73.96

*** homogeneous Subgroup - Tukey hsd post hoc test**

Table 6-5: Mean Score Values for Stratification Level Two

Level	Subpopulations		N	Mean
Level Two Usual Young	Gender	Females	29619	76.15
		Males	10605	77.44
	Years Attending	Less than 5	13800	76.6986 *
		5 to 10 Years	7339	75.92
		More than 10 Years	19085	76.5608 *
Level Two Usual Middle-Age	Gender	Females	158672	77.98
		Males	80952	78.58
	Years Attending	Less than 5	57031	78.07
		5 to 10 Years	50285	77.72
		More than 10 Years	132308	78.41
Level Two Usual Senior	Gender	Females	79737	81.90
		Males	72246	83.02
	Years Attending	Less than 5	15000	83.18
		5 to 10 Years	18188	82.53
		More than 10 Years	118795	82.33
Level Two Non-Usual Young	Gender	Females	20427	72.66
		Males	7323	73.78
	Years Attending	Less than 5	9976	73.98
		5 to 10 Years	4650	71.92
		More than 10 Years	13124	72.54
Level Two Non-Usual Middle Age	Gender	Females	69527	72.73
		Males	35187	73.46
	Years Attending	Less than 5	28743	73.90
		5 to 10 Years	22562	72.48900 *
		More than 10 Years	53409	72.68636 *
Level Two Non-Usual Senior	Gender	Females	18651	76.63
		Males	17054	77.92
	Years Attending	Less than 5	4022	78.25
		5 to 10 Years	4420	77.0702 *
		More than 10 Years	27263	77.1199 *

The “Usual GP” factor was used to make the first data split of two subpopulations at the stratification level one. For each subpopulation, the information gain, linear regression coefficient and ANOVA test was repeated to identify the data split for the next stratification level. The results show the sociodemographic factor “Age Group” was identified to split the data at stratification level two. However, the Tukey HSD post hoc test identified homogeneous subgroups between mean score values (72.96, 72.98) for non-usual young and middle-aged patients respectively. Non-usual and senior patients gave an average score of 77.24. The results indicated that non-usual young and middle age patients can be combined together in a large homogeneous subgroup against non-usual senior patients. Another homogeneous Subgroups identified at stratification level one are patients who have been attending the practice for less than five years and between five to ten years with mean score values of 78.74, 78.70 respectively. Figure 6-1 shows the stratification tree for levels zero to three using information gain as the stratification guide.

The sequence of the stratification processing using based on information gain and linear regression coefficient are in line with exploratory analysis that showed “Usual Doctor” factor to have statistically significant differences even with smaller subgroups. However, the exploratory analysis also showed significant differences start to disappear as samples sizes get smaller. The stratification process implemented in this chapter showed similar effects when smaller subgroups such as “non-usual young and middle-aged patients” start to appear as a homogeneous subpopulation. This can represent a challenge to researchers trying to identify when the stratification process should terminate. Also, different subpopulations may represent homogeneous subgroups even when they belong to different branches of the stratification tree. For instance, Researchers can set a threshold to stop the stratification process after predefined number of levels. Another option is to implement a complete stratification and run Tukey post-hoc analysis to identify possible homogeneous subpopulation.

The stratification process of IPQ dataset was continued until all patients are placed into 36 mutually exclusive subgroups. Then, ANOVA test with Tukey post-hoc option identified 17 possible homogeneous subgroups with mean score values ranges from 71.7 for (Non-usual young females seeing the doctor between 5 to 10 years) to 83.48 for (Usual senior males seeing the doctor for less than 5 years). The results also show that some subpopulations may have more common similarities with subgroups that may not necessarily belong to their own branch. Ranking the patients subgroups in ascending order based on their mean scores can identify homogeneous subgroup that have similarities with subgroups that belong to different branches. For instance, young to middle-aged female patients who are evaluating their non-usual doctor and have been attending the practice for more than five years (rank 1-4) belong to a different homogeneous subgroup from patients with similar characteristics but have been attending the practice for less than 5 years (rank 9-10). The largest homogeneous patients group is in the upper half of the ranking (20-27) and has eight different subgroups that belong to the usual and non-usual branches of stratification tree. The group shows that usual middle-aged females and males patients gave similar average scores to non-usual senior males. The first twelve subgroups in the ranking are dominated by the non-usual branch with the first usual groups appear at rank thirteen and fourteen due to the effect of young and female patients. Similarly, the top six position are dominated by the usual doctor branch with highest non-usual group appears in position (29) due to the effect of senior males subgroups. Table 6-6 shows the non homogeneous patients subgroups ranking and their respective mean score.

Table 6-6: Non - Homogeneous Populations Subgroups

Stratum	N	Homogeneous Subgroups Mean Values																
Non-Usual Young Females 5 to 10 Ys	6805	71.70																
Non-Usual Young Females More than 10	19884	72.15																
Non-Usual Middle Age Females 5 to 10 Ys	30868	72.31	72.31															
Non-Usual Middle Age Females More than 10	74616	72.34	72.34															
Non-Usual Middle Age Males 5 to 10 Ys	16035		72.98	72.98														
Non-Usual Middle Age Males More than 10	36311			73.43	73.43													
Non-Usual Young Males 5 to 10 Ys	2822			73.49	73.49													
Non-Usual Young Males Less than 5	5166			73.65	73.65													
Non-Usual Middle Age Females Less than 5	39154				73.73													
Non-Usual Young Females Less than 5	15612				73.80													
Non-Usual Middle Age Males Less than 5	20768				74.03													
Non-Usual Young Males More than 10	7153				74.12													
Usual Young Females 5 to 10 Ys	10679					75.43												
Usual Young Females More than 10	29090					76.03	76.03											
Non-Usual Senior Males More than 10	30057						76.28	76.28										
Non-Usual Senior Females 5 to 10 Ys	4685						76.43	76.43										
Usual Young Females Less than 5	21778						76.47	76.47										
Usual Young Males 5 to 10 Ys	4486							76.89	76.89									
Usual Young Males Less than 5	6866								77.21	77.21								
Usual Middle Age Females 5 to 10 Ys	69387									77.63	77.63							
Usual Young Males More than 10	10879									77.83	77.83	77.83						
Usual Middle Age Females Less than 5	77709									77.91	77.91	77.91						
Non-Usual Senior Males 5 to 10 Ys	4467									77.92	77.92	77.92						
Non-Usual Senior Males More than 10	27079										77.93	77.93						
Usual Middle Age Males 5 to 10 Ys	35146										77.98	77.98						
Non-Usual Senior Females Less than 5	4413										78.03	78.03						
Usual Middle Age Females More than 10	184685										78.16	78.16						
Usual Middle Age Males Less than 5	41913											78.44	78.44					
Usual Middle Age Males More than 10	91651												78.89	78.89				
Non-Usual Senior Males Less than 5	4151													79.22				
Usual Senior Females More than 10	130447														81.82			
Usual Senior Females 5 to 10 Ys	19786														81.96	81.96		
Usual Senior Females Less than 5	16268															82.65	82.65	
Usual Senior Males More than 10	117139																82.95	82.95
Usual Senior Males 5 to 10 Ys	18411																83.04	83.04
Usual Senior Males Less than 5	14991																	83.48

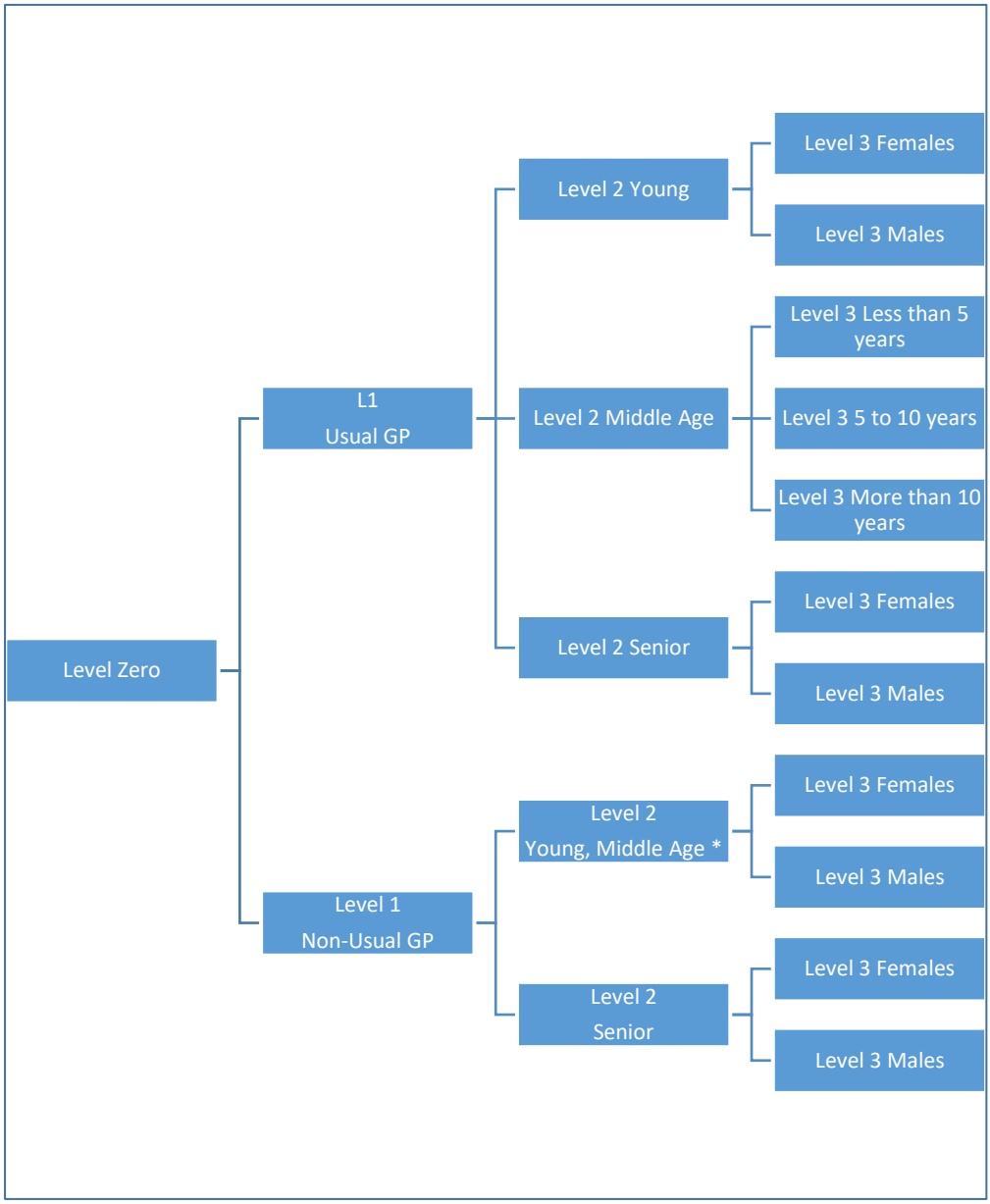

Figure 6-1: Stratification Tree

6.3 Stratification Methodology

The previous section in this chapter highlighted the results and implementation of the stratification process using IPQ dataset. The process started by identifying the “Usual Doctor” factor as having the highest information gain and regression coefficient values. The process resulted in seventeen homogeneous subgroups with sociodemographic factors that belong to different branches of the stratification tree such as 'Usual Young Females' and 'Non-Usual Senior Males'. This section presents formalised steps to implement the stratification process in non-probability and convenience sampling studies.

1- Identify a Summative Scale: The process of developing new instruments for collecting subjects feedback usually involves several steps like identifying the needed constructs and suitable questions in each construct. To guide the stratification process, researchers would need a summative variable that can reflect the overall view of patient experience feedback. However, exploratory analysis using PCA and person correlation highlighted that no single item can work as a summative variable in IPQ dataset. Possible candidates like question nine “My overall satisfaction with this visit to the doctor is …” and question fourteen “My confidence in this doctor’s ability is …” were only highly correlated with doctor communication items. Therefore, a new “Overall Score” item was created as a scale between 0 and 100 to represents patients overall feedback across all IPQ constructs. Exploratory analysis using PCA implemented at different stratification levels revealed that patients have inconsistent view about the importance of IPQ constructs “Clinic Access” and “Staff Information”. Therefore, local summative variables within each construct can be used to stratify subjects based on specific dimension.

2- Scale Configuration: To utilize the summative variables identified in the previous step within a systematic stratification process, the scale values need to be configured to work within the hierarchical modelling environment. The mean and standard deviation of the overall scale was to convert the overall scale into a new ordinal summative variable. The categories of the oridinal scale would have a simi-noraml distribution probabilities for standard deviations between one and three. Both the categorical and numerical scales would allow researchers to apply the stratification process using hierarchical modelling technique such as information gain, or numerical modelling using linear regression.
3- Stratification: After identifying the suitable summative scale and applying the required configuration, researchers can start the data splitting process at the stratification level zero. At this stage, information gain and linear regression would identify the sociodemographic variable with the highest variance. For each stratification level, researchers must re-run the calculation of the summative scale to remove the impact of the stratified factor.
4- Stopping Criteria and Merging Sub-Groups: The statistical ANOVA test can be applied to identify homogeneous subgroups that can be merged into a larger subgroup or to stop stratification process. Researchers can also assign a threshold to limit the stratification process to a predefined number of levels. A homogeneity test such as Tukey HSD can be then applied to identify similarities among subgroups in different branches.

The stratification methodology suggested in this section is derived by the concept of information gain, while standard statistical tests such as ANOVA test and p-value can be used as a general guideline to highlight possible differences among sociodemographic sub groups. P values are used for exploring the possibility of significant differences between subgroups

rather than to test null hypotheses. Therefore, the stratification methodology is unlikely to be subject to type one and type two error as used in in classical, confirmatory hypothesis testing.

6.4 Discussion

Previous chapters in this study have highlighted several challenges when analysing survey dataset collected through non-probability sampling techniques. In specific domains such as healthcare, researchers may not be able to ensure that the sample chosen is representative of the population due to financial and operational constraints. As a result, many healthcare studies have reported inconsistent findings of the satisfaction profile of different sociodemographic subgroups. Other researchers also indicated that patients groups with the highest proportional representation usually skew performance measures. Therefore, there was a need for researchers to adopt a stratified analysis methodology to help reveal patterns about smaller subpopulations that may otherwise remain unknown. However, there was no clear systematic stratification methodology that can help researchers to create smaller subpopulations.

This chapter presented a prototype methodology to split non-probability sampling dataset into smaller subgroups by identifying the effect for each socio-demographic factor. The methodology was implemented using statistical and machine learning techniques that can be used for numerical and categorical variables. the results showed that both information gain and linear regression techniques identified similar sociodemographic factors at different stratification levels. The ANOVA test also identified a large magnitude effect size for sociodemographic factors at stratification level zero. These results are consistent with exploratory analysis that identified the sociodemographic factor “Usual Doctor” to have statistically significant differences even with smaller samples size. Similarly, exploratory analysis identified the sociodemographic factor “Years Attending” to have no statistically significant differences among its subgroups. Both information gain and linear regression techniques identified “Age Group” to split the data at stratification level two followed by “Gender” at stratification level three. However, using the stratification mythology to identify “differences” among subpopulations may also hide “similarities” between groups that belong

to different branches. Also, information gain and linear regression techniques using in this methodology may not help researchers to decide when to stop the stratification process.

The stratification methodology would split the data using the sociodemographic factor that can maximise differences between subpopulations. As the process continued into deeper levels, similar subpopulations may end being at different branches despite having no statistically significant differences among their subjects. Both information gain and linear regression techniques cannot identify if several subgroups should be treated as a large homogeneous group during the stratification process. For instance, splitting the data using "Age Group" factor may need to handle young and middle-aged patients as one larger subgroup against senior patients subgroup. The statistical ANOVA test can be applied to identify homogeneous subgroups that can be merged into a larger subgroup or to stop stratifying a branch. Researchers can also assign a threshold to limit the stratification process to a predefined number of levels. A homogeneity test such as Tukey HSD can be then applied to identify similarities among subgroups in different branches.

The analysis implemented in this study showed that IPQ dataset can be grouped into 17 different homogeneous subpopulations. The stratification analysis results also showed statistically insignificant differences among large population subgroups and statistically significant differences among small patient sub groups. This pattern is consistent with the research objective to examine the validity of applying a systematic stratification analysis on a large scale non-probabilistic survey sample, rather than identifying significant differences between sub-populations. These larger populations can include subgroups that belong to different stratification branches. High scoring subgroups are dominated by senior patients who are evaluating their usual doctors while low scoring subgroups consist of mainly young and middle-aged females evaluating their non-usual doctor. The discovered insights highlighted using stratification analysis can support healthcare organizations to design better healthcare

policies. For example, dedicating more resources to understand and improve the doctor visit experience for the least satisfied subgroups such as females visiting their non-usual doctor.

In conclusion, stratifying survey data into smaller subgroups can help researchers to highlight insights about smaller sociodemographic subpopulations that may otherwise remain unknown. The stratification methodology can guide researchers to find a systematic process to split the data into smaller subgroups. The next chapter will discuss how researchers can use the homogeneous subpopulations to create pseudo-controlled samples for estimating population parameters.

Chapter 7 Data Reliability of Stratification Analysis

7.1 Introduction

Previous chapters in this thesis presented a data stratification methodology to perform drill-down analysis on large-scale survey datasets collected using non-probabilistic sampling techniques. The proposed methodology showed that stratification analysis using machine learning and statistical modelling techniques could identify response differences among raters subpopulations based on sociodemographic characteristics. For instance, the stratification process identified a set of raters subpopulation with a similar response profile despite having a non homogeneous sociodemographic factors such as "Young and middle age patients evaluating their usual doctor; with Senior males patients evaluating their non-usual doctors". However, deriving any insights about population subgroups from research studies that adopt a non probabilistic sampling method (i.e. questionnaires are handed out to raters until the questionnaires are exhausted or time runs out, without any attempt to ensure adequate representation of various sociodemographic groups in the sample), would requires a measure of data reliability as opposed to questionnaire reliability.

Questionnaire reliability can be measured through a number of measures including Cronbach's alpha [145] to provides a coefficient of internal consistency (or reliability) of survey items. However, the provided reliability coefficient requires various sampling assumptions are met (e.g. the data is balanced, crossed or not fully nested). When these assumptions are not met, more complex methods using analysis of variance need to be applied to allow generalizability of results to different populations of raters and rates [108]. On the other hand, measuring data reliability is critical for identifying the reliability of subpopulations at different levels of stratification. If the variance of the subpopulation feedback decreases as the level of

stratification increases, that tells the analyst that increasing amounts of noise are being removed with drill down. Similarly, if variance in the subpopulation feedback increases, stratification is leading to more noise.

So far, data reliability analysis have focused on tackling the nonstandard research design problems usually associated with convenience sampling techniques. This analysis can ensure the availability of adequate minimum numbers of raters and ratees for deeper drill down analysis. However, there has been no attempt to relate reliability with stratification analysis to extract evidence based knowledge for use in questionnaire construction, construct identification and training development programmes. In other words, there is currently no known method for helping analysts identify the reliability of stratification analysis by measuring the feedback consistency among raters subpopulation. This chapter investigates the use of data reliability measures to validate the reliability of stratification analysis.

7.2 Data Reliability for Non Probabilistic Sampling Settings

Metrics for measuring the reliability of 'soft' survey data can be based on metrics for measuring 'hard' data reliability, such as accuracy, precision, currency, completeness and consistency. Accuracy is a measure of the closeness of the measurements to their true values, whereas precision is a measure of how repeated measurements show the same results. Currency is usually associated with conformity to standards. Completeness is a measure of the extent to which all data is known. Consistency is a measure of how much of the data does not conflict with other parts of the data. The presence of an objectively measured output variable in hard data allows for data partitioning vertically (to measure individual attribute consistency, and individual and combined attribute completeness, for example) and horizontally (to measure sample accuracy and precision, for example, through test and cross validation techniques). Even if no objectively measured output variable is available, hard data is expected to conform

to statistical properties of real valued random variables to permit parametric modelling for data fitting and testing purposes.

Most of these traditional measures of hard data quality, especially accuracy and precision, do not seem applicable in the context of convenient sampling psychometric data (i.e. data that measures knowledge, abilities and attitudes). Such data is often characterized by unbalanced, not crossed or fully nested problems described previously. Another potential problem for calculating the reliability of psychometric survey data is the varying response rates and possible raters nonresponse bias. A variance based (the squared difference between a raw mean score for an item and the average score for that item), two-level signal to noise ratio (2LSNR) can be used to determine data reliability in such non-probabilistic sampling settings [109]. The variance is calculated from available item scores and ignoring any missing values. This calculation would eliminate the need for the removal or replacement of missing value scores. In its simplest form, one level (1LSNR) is expressed as:

Formula 1

$$R = \frac{Var(\tau)}{Var(\tau) + Var(\varepsilon)}$$

where τ is the true score (signal) and ε is the error (noise). The reliability of a set of observations (the data) is defined as the ratio of the true signal variance to the observed score variance including noise. 1LSNR is applicable when there is only one signal (one ratee, or one item). When many raters are rating many ratees on many items, there are multiple signals at two levels:

a. The raw scores of raters (level 1) irrespective of ratee (r);

b. The aggregated scores received by ratees (level 2) irrespective of number of raters (s); and

c. The aggregated item score received by ratees (level 2) irrespective of the number of items (i).

Noise in an unbalance study consists of two elements:

d. sample effect, or the noise contributed by the average number of raters per ratee at the raw score level to items j; and

e. the interference or crosstalk of the three signals a c above.

Typically, *i=j* unless the raw score items provided by raters are transformed to a different number of component or factor scores for ratees. The larger the sample size effect (caused by relatively smaller average numbers of raters per ratee), the more noise there is in the data. Putting this together using the variance based Formula 1 as a template, and expressing variance as average variances because of multiple ratees, items and raters, gives us an 2LSNR for estimated data reliability *R* as follows:

Formula 2 (2LSNR)

$$R = \frac{avs + avi + vr}{(vi/n) + avs + avi + vr + (avs \times avi) + (avs \times vr) + (avi \times vr) + (avs \times avi \times vr)}$$

where the true signal (numerator) consists of the following:

- *avs*: is the average ratee variance (the variance between practitioners at the average score level for the items of the questionnaire);

- *avi*: is the average aggregated mean item variance (the variance between items at the mean score level, irrespective of practitioner); and

- *vr*: is the average variance of patients providing raw scores (the variance between patients at the raw score level, irrespective of practitioner rated).

The noise (denominator) consists of the following:

- vi/n, the raw score item variance divided by the average number of patients/raters per practitioner contributing to this variance; and
- interactions between the three signals.

The basis of the two-level signal to noise reliability measure is to identify two general sources of variability: variability between raters and variability within the scores for each ratee. The reliability measure can help researchers to ensure whether there are enough raters to give a reliable score for a ratee; and whether there are enough ratees to give a reliable score [111][146]. The next section will investigates the data reliability measure behaviour in the context of raters drill down stratification analysis.

7.3 Data Reliability for Stratification Analysis

The stratification analysis presented in chapter six was repeated with the application of the two-level signal to noise formula to identify mutually exclusive raters subpopulations using the summative "Overall Scale" value described in chapter six. The statistical analysis at zero level (i.e. no stratification of subject scores by rater subgroups) involves the entire sample population with all 4 socio demographic variables. Data Reliability *R* calculated using 2LSNR (Formula 2) is 0.79 at level zero stratification with an average number of 37 patients per doctor. The reliability value can be interpreted as 79% of ratees' real scores can be accredited to ratings from patient raters, and 21% is due to differences among raters. Researchers can use the different variance values (*vi*, *vr*, *avi*, *avs*) as a guideline to select the next stratification level

based on raters sociodemographic variables. Table 7-1 shows the various parameter values for 2LSNR formula.

Table 7-1: Reliability Variables for Signal to Noise Formula

Signal to Noise Reliability	All
average *vi*	0.94
average *vr*	0.54
average *avi*	0.13
average *avs*	0.17
N	37.83
R	0.79

The reliability formula was recalculated at stratification level one using the four possible subpopulations available in IPQ dataset. The results show all possible stratifications options give lower reliability scores at stratification level one (maximum 0.78, minimum 0.60). This is due, as might be expected, to the lower average number of subpopulation raters per ratee in comparison to level zero, leading to higher average vi/n values in the denominator of Formula 2. For instance, the two subpopulations with the lowest number of raters per ratee are Young (Age Group) and Years 2 (Years Attending), with 4.84 and 7.04 raters per ratee, respectively, producing level one stratification *R* ratios of 0.60 and 0.65. Lower average number of raters does not necessarily mean lower *R* ratios, however, and lower numbers of raters can be compensated for by lower aggregated item (*avi*), rater (*vr*) and ratee (*avs*) variances. For instance, if *avi*, *vr* and *avs* had been lower at 0.15 for this Young (Age Group) subpopulation, *R* would be 0.82, which is higher than the *R* ratio for level zero (0.79). Table 7-2 shows reliability values using the possible options at stratification level one.

Table 7-2: Reliability of Possible Stratification at Level One

	Usual Doctor		Age Group			Gender		Years Attending		
	Usual	**Non-Usual**	**Young**	**Middle**	**Senior**	**Females**	**Males**	**Years 1**	**Years 2**	**Years 3**
vi	0.89	0.99	0.99	0.98	0.80	0.96	0.90	0.96	0.96	0.92
vr	0.52	0.59	0.59	0.57	0.46	0.57	0.50	0.52	0.54	0.55
avi	0.14	0.30	0.45	0.17	0.21	0.16	0.19	0.30	0.34	0.16
avs	0.19	0.25	0.33	0.22	0.20	0.19	0.18	0.22	0.27	0.20
n	27.35	11.13	4.84	21.77	12.09	23.79	14.17	8.38	7.04	23.16
R	0.78	0.68	0.60	0.75	0.73	0.76	0.74	0.68	0.65	0.76

The results show several sociodemographic variables (Usual, Male, and Senior) achieved lower items and patients variance (*vi*, *vr*) than stratification level zero. Assume that the age group "Senior" (6th column of Table five) is chosen to make a subpopulation split at level one because it has the lowest average variance of items *vi* (0.802) and variance of patients *vr* (0.46) in comparison to level zero values. The data split will create two subpopulations at stratification level one; "Senior" and "Non Senior" patients. Table 7-3 show reliability values using the possible options at stratification level two, while Table 7-4 and Table 7-5 show the stratification process for non senior patients groups.

Table 7-3: Reliability of Possible Stratification at Level Two (Senior)

	Senior						
	Females	**Males**	**Usual**	**Non-Usual**	**Years 1**	**Years 2**	**Years 3**
vi	0.82	0.78	0.77	0.90	0.79	0.81	0.80
vr	0.48	0.45	0.45	0.52	0.41	0.45	0.47
avi	0.32	0.31	0.23	0.53	0.57	0.56	0.25
avs	0.25	0.24	0.21	0.34	0.32	0.35	0.22
n	6.59	6.15	10.14	3.25	2.19	2.39	9.64
R	0.67	0.68	0.72	0.58	0.57	0.57	0.71

Table 7-4: Reliability at Level One - Non Senior

	Non Senior Branch
vi	0.98
vr	0.58
avi	0.15
avs	0.19
n	26.04
R	0.77

Table 7-5: Reliability of Possible Stratification at Level Two (Non Senior)

	Non Senior								
	Females	**Males**	**Young**	**Middle**	**Usual**	**Non Usual**	**Years 1**	**Years 2**	**Years 3**
vi	0.99	0.95	0.99	0.98	0.94	1.01	0.98	0.99	0.97
vr	0.60	0.53	0.59	0.57	0.56	0.60	0.54	0.57	0.60
avi	0.19	0.28	0.45	0.16	0.19	0.35	0.34	0.41	0.22
avs	0.22	0.23	0.33	0.20	0.22	0.28	0.25	0.31	0.24
n	17.62	8.66	4.83	21.77	17.82	9.01	7.29	5.72	13.97
R	0.73	0.68	0.60	0.75	0.74	0.66	0.66	0.62	0.71

The results show a similar pattern to stratification level one with all possible sociodemographic variable achieved lower reliability scores at stratification level two. This is due to average aggregated item variances "*avi*" and average ratee variances "*avs*" achieved higher variance values than stratification level one for all the possible subpopulations. Also, the smaller average of raters contributed to lowering *R* values. However, other sociodemographic factors like "Seeing Usual Doctor" and "Male" achieved lower values for variance of patients *vr* and variance of items vi than stratification level one.

The stratification process of IPQ dataset was continued until all patients are placed into 36 mutually exclusive subgroups with *R* value in the range of minimum = 0.45 for "Non- usual young females seeing the doctor between 5 to 10 years" and maximum = 0.65 for "Usual senior males seeing the doctor for more than 10 years". The results show that subpopulations that minimized patients and items variance achieved higher reliability scores. Ranking all IPQ subpopulations in descending order based on *R* value shows that subpopulations belong to

"Usual senior or middle-aged" patients have lower items and patients variance in compare to level zero stratification level. On the other hand, subpopulations belong to "Non- usual young and middle-aged" patients have high items and patients variance in compare to level zero stratification level. Table 7-6 shows the different IPQ subpopulations with the various parameter values for 2LSNR formula.

Table 7-6: IPQ Sub Population Reliability Values

ID	Subpopulation Name	*vi*	*vr*	*avi*	*avs*	*N*	Total Patients	*R*
0	All Population	0.94	0.54	0.13	0.17	37.83	1251357	0.79
1	Senior Male More Than 10 Usual GP	0.74	0.44	0.36	0.27	4.49	117139	0.65
2	Senior Female More Than 10 Usual GP	0.78	0.48	0.37	0.29	4.76	130447	0.64
3	Middle Age Female More Than 10 Usual GP	0.94	0.60	0.39	0.35	6.32	184685	0.62
4	Middle Age Male More Than 10 Usual GP	0.90	0.54	0.48	0.35	3.72	91651	0.59
5	Middle Age Female Less Than 5 Usual GP	0.95	0.53	0.54	0.35	3.21	77709	0.57
6	Middle Age Female 5 10 Years Usual GP	0.96	0.57	0.57	0.41	3.01	69387	0.55
7	Middle Age Female More Than 10 Non-Usual GP	1.02	0.64	0.60	0.43	3.42	74616	0.55
8	Middle Age Male Less Than 5 Usual GP	0.92	0.49	0.62	0.36	2.24	41913	0.55
9	Senior Male 5 10 Years Usual GP	0.75	0.42	0.61	0.36	1.61	18411	0.54
10	Senior Male Less Than 5 Usual GP	0.74	0.38	0.63	0.33	1.51	14991	0.54
11	Senior Female Less Than 5 Usual GP	0.78	0.41	0.64	0.36	1.54	16268	0.53
12	Senior Female 5 10 Years Usual GP	0.79	0.45	0.64	0.39	1.65	19786	0.53
13	Senior Male More Than 10 Non-Usual GP	0.88	0.51	0.65	0.40	1.95	27079	0.53
14	Middle Age Male 5 10 Years Usual GP	0.93	0.51	0.65	0.40	2.07	35146	0.53
15	Senior Female More Than 10 Non-Usual GP	0.91	0.54	0.67	0.43	2.03	30057	0.53
16	Middle Age Male More Than 10 Non-Usual GP	0.97	0.59	0.68	0.44	2.24	36311	0.52
17	Middle Age Female Less Than 5 Non-Usual GP	1.02	0.59	0.71	0.43	2.28	39154	0.52
18	Young Female Less Than 5 Usual GP	0.97	0.55	0.75	0.45	1.85	21778	0.51
19	Middle Age Male Less Than 5 Non-Usual GP	0.97	0.53	0.76	0.43	1.73	20768	0.50
20	Middle Age Female 5 10 Years Non-Usual GP	1.02	0.62	0.76	0.49	2.04	30868	0.50
21	Young Female More Than 10 Usual GP	0.98	0.63	0.73	0.51	1.94	29090	0.50
22	Young Female Less Than 5 Non-Usual GP	1.00	0.60	0.84	0.50	1.77	15612	0.49
23	Middle Age Male 5 10 Years Non-Usual GP	0.98	0.56	0.80	0.48	1.59	16035	0.49
24	Young Male Less Than 5 Usual GP	0.93	0.50	0.83	0.45	1.39	6866	0.49
25	Young Male More Than 10 Usual GP	0.90	0.54	0.77	0.48	1.41	10879	0.49
26	Senior Male Less Than 5 Non-Usual GP	0.85	0.45	0.78	0.42	1.18	4151	0.49
27	Senior Male 5 10 Years Non-Usual GP	0.90	0.48	0.82	0.44	1.21	4467	0.48
28	Young Male Less Than 5 Non-Usual GP	0.97	0.53	0.87	0.48	1.43	5166	0.48
29	Young Female More Than 10 Non-Usual GP	1.01	0.66	0.79	0.54	1.77	19884	0.48
30	Senior Female Less Than 5 Non-Usual GP	0.89	0.49	0.83	0.46	1.19	4413	0.48
31	Young Male 5 10 Years Usual GP	0.91	0.50	0.84	0.47	1.22	4486	0.48
32	Young Female 5 10 Years Usual GP	0.98	0.58	0.84	0.53	1.39	10679	0.47
33	Senior Female 5 10 Years Non-Usual GP	0.93	0.53	0.87	0.49	1.21	4685	0.47
34	Young Male More Than 10 Non-Usual GP	0.95	0.58	0.84	0.52	1.31	7153	0.47
35	Young Male 5 10 Years Non-Usual GP	0.92	0.54	0.88	0.52	1.14	2822	0.46
36	Young Female 5 10 Years Non-Usual GP	1.01	0.63	0.91	0.57	1.30	6805	0.45

The stratification process showed a reduction in reliability *R* value that was mainly derived by the decreasing average number of subpopulation raters per ratee after each stratification step. However, the stratification results also highlight the impact of sociodemographic variables to increase or decrease variance variables such as item (*avi*), rater (*vr*) and ratee (*avs*) variances. For example, the impact of sociodemographic factor "Seeing usual doctor" can be demonstrated by comparing the subpopulations "Middle Age Male Less Than 5 Usual GP" and "Middle Age Female 5 to 10 Years Non-Usual GP" with IDs (8, 20) respectively. The first subpopulation achieved lower variance values and thus, a higher *R* value in comparison to the second group. Table 7-7 was generated by simulating different variance values while keeping the average number of patients constant for subpopulation "Middle Age Female More Than 10 Usual GP".

Table 7-7: Impact of Adjusted Variance on Reliability Values

vi	***vr***	***avi***	***avs***	***n***	***R***
0.94	0.60	0.39	0.35	6.32	0.62
0.9	0.9	0.9	0.9	6.32	0.45
0.8	0.8	0.8	0.8	6.32	0.48
0.7	0.7	0.7	0.7	6.32	0.52
0.6	0.6	0.6	0.6	6.32	0.56
0.5	0.5	0.5	0.5	6.32	0.61
0.4	0.4	0.4	0.4	6.32	0.66
0.3	0.3	0.3	0.3	6.32	0.72
0.2	0.2	0.2	0.2	6.32	0.79
0.1	0.1	0.1	0.1	6.32	0.87

The stratification process showed how splitting raters into smaller subpopulations based on sociodemographic factors could lead to a maximizing or minimizing in response variance (noise). Despite having lower *R* values in comparison to stratification level zero, sociodemographic factors like "senior age", "seeing usual doctor" and "male", created subpopulations that reduced the noise among raters while increasing variance among ratees. This pattern confirms results from healthcare domain that highlighted the need to apply a stratified approach when evaluating ratees (Medical professionals). This approach can ensure

identifying the feedback profile of raters subpopulations while maximizing differences among raters.

To identify the validity of the estimated reliability for IPQ dataset following the stratification analysis, a simulation analysis was performed by modelling the multiple variance values in 2LSNR as independent random variables to monitor the impact or *R* values. The different simulated variables were used to reflect the actual values found in IPQ data. The four variance variables were set as random values between 0 (no variance or noise) and 4 (maximum variance in IPQ dataset) for 1000 iteration to generate a simulated variance dataset. The average number of raters per rate was kept constant at 37 (the actual value found in IPQ dataset). The average of the simulated *R* values from 1000 iteration was 0.31 with standard deviation of 0.14 (minimum: 0.1 maximum: 0.9). The reliability of IPQ dataset at zero stratification level is 3.5 standard deviation higher than the random *R* value; while the reliability of the stratified mutually exclusive subpopulations are between 1 and 2.5 standard deviation higher than the random *R* value. The results indicates that while average number of raters available at each stratification level contributed to downgrading *R* values, the reliability of IPQ and all its sociodemographic subpopulations are considerably higher than random variance reliability.

7.4 Discussion

Recent years have witnessed a significant increase in the use of large-scale convenient sampling surveys as organisations seek to uncover new knowledge and information from customer feedback and other users of organisation services. Previous sections in this thesis presented the theoretical and conceptual framework to implement a drill down data stratification analysis using survey dataset collected through non-probabilistic (i.e. convenient sampling) techniques. Analysis of such data usually applies well-established statistical techniques for demonstrating the reliability of questionnaires used for collecting data. However, convenient sampling data have distinctive features including unbalanced and fully nested and multi-level data structure, with no attempt made to ensure adequate representation of various sociodemographic groups in the sample. Therefore, new techniques to ensure data reliability as opposed to questionnaire reliability has been recently developed. The data reliability techniques separate variances at the raters and ratees levels and consider the average number of raters per ratee to derive a reliability coefficient. This chapter investigated the use of signal-to-noise ratio reliability formula to validate the reliability of drill down stratification analysis. The results demonstrated that the reliability formula can be used to guide the stratification analysis to create mutually exclusive raters subpopulation. However, caution must be given to the interpretation of reliability and variance values in order to decide the most appropriate stratification strategy.

The proposed data stratification methodology presented in chapter six utilizes machine learning and statistical modelling techniques to identify response differences among raters subpopulations based on sociodemographic characteristics. The objective of such methodology is to identify smaller raters subpopulations with a similar and high response consistency despite having a non-homogeneous sociodemographic profiles. For instance, the results showed that

senior and young patients might provide a highly similar response (i.e. small variance) depending on whether they are evaluating their usual or non-usual medical professional.

This pattern is reflected in the raters level of signal to noise formula and can be observed using items and patients variance variables (vi, vr). Rater variance can get smaller following stratification, indicating that raters can tend to agree more depending on which subpopulation they belong to. Subpopulations that belong to senior and middle-aged patients who are evaluating their usual medical professional achieved smaller raters variance in compare to raters subpopulations that are dominated by young patients who are evaluating their non-usual medical professional. The results highlight the validity of the stratification methodology that identified "Seeing Usual Doctor" and "Age Group" as the most relevant sociodemographic factors. Nonetheless, the results also showed the overall reliability value (R) is constantly decreasing with stratification due to the smaller average numbers of raters per ratee. This pattern is in line with the design of signal-to-noise formula that intend to ensure the availability of adequate minimum numbers of subjects at the raters and ratees levels. However, simulated values of signal-to-noise formula showed that reliability value could be improved by achieving lower rater variance, ratee variance and item variance. The results highlight the need to look in detail at the variances at each stratification level and compare them with variances at the next level before deciding on the most appropriate stratification strategy.

As the use of large-scale survey data continues to grow in the age of big data, applying statistical analysis techniques such as data reliability and stratification could help to extract new knowledge about different raters subpopulation groups. The reliability analysis presented in this chapter highlighted how data stratification could identify non-homogeneous raters subpopulations that tend to have more agreement than other subpopulations. The results show the validity of using stratification analysis in a domain like, for example, healthcare to generate

multiple performance scores (by minimizing variance within stratum) while maintaining variance between medical professionals.

While the stratification and reliability analysis can be used to identify similarities and biases at the raters and ratees levels, there is still a need to apply this knowledge for estimating population parameters given the inherited biases associated with convenient sampling. The next chapter will discuss how researchers can use the identified raters subpopulations to create pseudo-controlled samples for estimating population parameters. The analysis will validate if patterns found at the raters subpopulations levels would be reflected at the individual ratee level.

Chapter 8 Estimating Population Parameters using Probability Sampling

8.1 Introduction

The data stratification methodology and reliability analysis presented in chapters six and seven can guide researchers to extract new insights about the feedback profiles of smaller raters subpopulations. For the objective of estimating the population parameters from a non-probability sampling survey dataset, researchers would need a reliable stratification methodology to create mutually exclusive sub raters populations. The stratification process would help to quantify the amount of bias for different sociodemographic factors so that a pseudo controlled samples can be created to provide an accurate and precise representation of population parameters.

To estimate the population parameters, researchers can choose to stratify raters data using different sociodemographic factors. The stratification process can start using sociodemographic factors that can maximize or minimize variance among subpopulations. Calculating statistics at multiple stratification levels can reveal patterns about how the feedback of smaller subpopulations differs from the overall population. However, convenient sampling datasets would usually have inherited biases due to the uncontrolled proportionality of different sociodemographic factors. Therefore, using a conveniently sampled dataset as a source to create pseudo controlled samples will require "controlling" the proportionality of demographic factors to estimate the real population parameters.

Chapter two of this thesis reviews the different theoretical and conceptual frameworks that are intended to define and describe the construct of "Patients Satisfaction". The review highlighted that the majority of patients satisfaction theories were adapted from a business research

perspective. Several theories presented patients satisfaction as a value measured within the context of service expectancy as well as prior expectations about care. However, evidence presented in this thesis so far suggest that a significant amount of patients satisfaction variance can be attributed to raters sociodemographic factors. Therefore, estimating population parameters using pseudo controlled samples can support an evidence-based theory of patients' satisfaction.

The advantage of using large-scale survey datasets is the ability to repeat the analysis by taking average of multiple samples datasets. For instance, the smallest stratified subgroup in IPQ dataset "Non-Usual Young Males visiting their doctor between five to ten tears" contains more than 2800 patients. For the analysis implemented in this chapter, random samples are selected from each subgroup and their mean values are calculated. The analysis is repeated for 1000 times and the reported value represent the mean of means. **Error! Reference source not found.** display the calculation code template.

```
vMean = []
for i in range (0,1000):
    frames = [GROUP1.Overall_Scale.sample(1000), GROUP2.Overall_Scale.sample(1000)]
    result = pd.concat(frames)
    vMean.append(result.mean())
    sMean = pd.Series(vMean)
print (sMean.mean())
```

Figure 8-1. Calculation Code Template

This chapter investigates the process of estimating population parameters by using random sampling to create pseudo controlled samples.

8.2 Maximum to Minimum Variance Stratification

The maximum to minimum variance stratification approach utilizes raters sociodemographic factors that can introduces the highest amount of variance between patients subpopulations. Results from chapters six and seven highlighted that the sociodemographic factor "Evaluating Usual Doctor" introduces the highest amount of variance in comparison to other sociodemographic factors. The overall average value of the summative scale at zero stratification level is (77.9). Patients who answered they are visiting the usual doctor represented two third of the sample population and gave an average overall scores of 79.5 in comparison to 73.89 for patients who reported they are visiting their Non-Usual doctor. Removing the proportionality bias by taking multiple "equal size" sample from usual and non-usual patients groups revealed an average score of (76.4). Table 8-1 shows the results of estimating the population parameters at the zero aggregation level.

Table 8-1: Estimating parameters at Level Zero Aggregation

Level 0 Stratification	Based on Level 1	N	Original Mean	New Mean
L0_IPQ_No_Missing			77.94	76.70
	L1_usual	901311	79.52	
	L1_non_usual	350046	73.89	

The analysis at the stratification level one showed a relatively large and statistically significant difference in response values between both usual and non-usual groups. However, each one of the two subpopulation groups may also have a sub proportionality bias within its sociodemographic factors. The analysis at the stratification level two shows the effect of sub proportionality bias for the "Age Group" factor within "Usual and Non-Usual" patients subpopulations. Table 8-2 shows the effect of removing the proportionality bias at the aggregation level one. While tables Table 8-3 and Table 8-4 show the analysis results at the stratification levels three and four.

Table 8-2: Estimating parameters at Level 1 Aggregation

Level 1 Stratification	Based on Level 2	N	Original Mean	New Mean
L1_usual			79.52	79.03
	L2_usual_young	83778	76.44	
	L2_usual_middle	500491	78.19	
	L2_usual_senior	317042	82.44	
L1_non_usual			73.89	74.39
	L2_non_usual_young	57442	72.99	
	L2_non_usual_middle	217752	72.97	
	L2_non_usual_senior	74852	77.24	

Table 8-3: Estimating parameters at Level 2 Aggregation

Level 2 Stratification	Based on level 3	N	Original Mean	New Mean
L2_usual_young			76.44324551	76.77361296
	L3_usual_young_female	61547	76.07988836	
	L3_usual_young_male	22231	77.44920756	
L2_usual_middle			78.19125915	78.08382074
	L3_usual_middle_attending1	119622	78.09327778	
	L3_usual_middle_attending2	104533	77.7449262	
	L3_usual_middle_attending3	276336	78.40251374	
L2_usual_senior			82.4403917	82.64091037
	L3_usual_senior_attending1	31259	83.04827173	
	L3_usual_senior_attending2	38197	82.48309204	
	L3_usual_senior_attending3	247586	82.35705603	
L2_non_usual_young			72.99012079	73.23333333
	L3_non_usual_young_female	42301	72.68561202	
	L3_non_usual_young_male	15141	73.84085889	
L2_non_usual_middle			72.97491336	73.11594185
	L3_non_usual_middle_female	144638	72.70949748	
	L3_non_usual_middle_male	73114	73.49997315	
L2_non_usual_senior			77.24719991	77.28100296
	L3_non_usual_senior_female	39155	76.49158851	
	L3_non_usual_senior_male	35697	78.07600805	

Table 8-4: Estimating parameters at Level 3 Aggregation

Level 3 Stratification	Based on Level 4	N	Original Mean	New Mean
L3_usual_young_female			76.080	75.984
	L4_usual_young_female_attending1	21778	76.469	
	L4_usual_young_female_attending2	10679	75.431	
	L4_usual_young_female_attending3	29090	76.027	
L3_usual_young_male			77.449	76.554
	L4_usual_young_male_attending1	6866	77.211	
	L4_usual_young_male_attending2	4486	76.890	
	L4_usual_young_male_attending3	10879	77.830	
L3_usual_middle_attending1			78.093	78.183
	L4_usual_middle_attending1_female	77709	77.905	
	L4_usual_middle_attending1_male	41913	78.442	
L3_usual_middle_attending2			77.745	77.988
	L4_usual_middle_attending2_female	69387	77.626	
	L4_usual_middle_attending2_male	35146	77.980	
L3_usual_middle_attending3			78.403	78.523
	L4_usual_middle_attending3_female	184685	78.158	
	L4_usual_middle_attending3_male	91651	78.895	
L3_usual_senior_attending1			83.048	83.070
	L4_usual_senior_attending1_female	16268	82.648	
	L4_usual_senior_attending1_male	14991	83.483	
L3_usual_senior_attending2			82.483	82.509
	L4_usual_senior_attending2_female	19786	81.962	
	L4_usual_senior_attending2_male	18411	83.044	
L3_usual_senior_attending3			82.357	82.388
	L4_usual_senior_attending3_female	130447	81.823	
	L4_usual_senior_attending3_male	117139	82.952	
L3_non_usual_young_female			72.686	72.550
	L4_non_usual_young_female_attending1	15612	73.798	
	L4_non_usual_young_female_attending2	6805	71.699	
	L4_non_usual_young_female_attending3	19884	72.150	
L3_non_usual_young_male			73.841	73.747
	L4_non_usual_young_male_attending1	5166	73.647	
	L4_non_usual_young_male_attending2	2822	73.493	
	L4_non_usual_young_male_attending3	7153	74.118	
L3_non_usual_middle_female			72.709	72.783
	L4_non_usual_middle_female_attending1	39154	73.733	
	L4_non_usual_middle_female_attending2	30868	72.306	
	L4_non_usual_middle_female_attending3	74616	72.339	
L3_non_usual_middle_male			73.500	73.485
	L4_non_usual_middle_male_attending1	20768	74.029	
	L4_non_usual_middle_male_attending2	16035	72.983	
	L4_non_usual_middle_male_attending3	36311	73.426	
L3_non_usual_senior_female			76.492	76.918

Level 3 Stratification	Based on Level 4	N	Original Mean	New Mean
	L4_non_usual_senior_female_attending1	4413	78.029	
	L4_non_usual_senior_female_attending2	4685	76.432	
	L4_non_usual_senior_female_attending3	30057	76.275	
L3_non_usual_senior_male			78.076	78.348
	L4_non_usual_senior_male_attending1	4151	79.219	
	L4_non_usual_senior_male_attending2	4467	77.916	
	L4_non_usual_senior_male_attending3	27079	77.927	

The results from the tables above show that following the maximum to minimum variance stratification process would create mutually exclusive raters subpopulations that have small variance (i.e. high agreement) between raters within the same group. Comparing the mean score ranking values of the created subpopulations shows that aggregating level four sociodemographic factor "Years Attending" would have a minimal effect on the ranking of level three subgroups with no other changes in ranking found within other stratification levels. Table 8-5 shows the changes in mean score ranking at stratification level three.

Table 8-5: Level 3 Ranking – Top Down

Level 3 Stratification	Old Mean	New Mean	Old Rank	New Rank	R. Change
L3_usual_senior_attending1	83.05	83.07	1	1	0
L3_usual_senior_attending2	82.48	82.51	2	2	0
L3_usual_senior_attending3	82.36	82.39	3	3	0
L3_usual_middle_attending3	78.40	78.52	4	4	0
L3_non_usual_senior_male	78.08	78.35	6	5	1
L3_usual_middle_attending1	78.09	78.18	5	6	1
L3_usual_middle_attending2	77.74	77.99	7	7	0
L3_non_usual_senior_female	76.49	76.92	9	8	1
L3_usual_young_male	77.45	76.55	8	9	1
L3_usual_young_female	76.08	75.98	10	10	0
L3_non_usual_young_male	73.84	73.75	11	11	0
L3_non_usual_middle_male	73.50	73.48	12	12	0
L3_non_usual_middle_female	72.71	72.78	13	13	0
L3_non_usual_young_female	72.69	72.55	14	14	0

*Green = Higher New Ranking, Red = Lower New Ranking

The stratification process was repeated starting with sociodemographic factors that can introduces the lowest amount of variance between raters subpopulations. The stratification and reliability analysis implemented in chapters six and seven showed that "Years Attending" factor has the lowest amount of variance between raters subpopulations in comparison to other factors. The analysis results showed that minimum to maximum stratification process created subpopulations with a large amount of variance at the lower stratification levels. Table 8-6 to Table 8-9 shows the stratification process using the sociodemographic factors "Years Attending", "Gender", "Age" and "Usual Doctor" respectively. Examining the old and new mean scores ranking at stratification level three after removing proportionality bias for "Usual Doctor" factor shows a much higher change in the ranking order in comparison to similar process following maximum to minimum stratification process. Table 8-10 shows changes in ranking order for three stratification analysis.

Table 8-6: Estimating parameters at Level 0 Aggregation

Level 0 Stratification	Based on Level 1	N	Original Mean	New Mean
L0		1251357	77.94785	77.58
	L1_year1	268789	77.24	
	L1_year2	223577	77.07	
	L1_year3	758991	78.45	

Table 8-7: Estimating parameters at Level 1 Aggregation

Level 1 Stratification	Based on Level 2	N	Original Mean	New Mean
L1_year1		268789	77.24765798	77.4
	L2_year1_female	174934	76.87	
	L2_year1_male	93855	77.95	
L1_year2		223577	77.07	77.25
	L2_year2_female	142210	76.5865	
	L2_year2_male	81367	77.9	
L1_year3		758991	78.45	78.67
	L2_year3_female	468779	77.7	
	L2_year3_male	290212	79.6	

Table 8-8: Estimating parameters at Level 2 Aggregation

Level 2 Stratification	Based on level 3	N	Old Mean	New Mean
L2_year1_female		174934	76.87	77.84
	L3_year1_female_young	37390	75.35	
	L3_year1_female_middle	116863	76.507	
	L3_year1_female_senior	20681	81.66	
L2_year1_male		93855	77.95	78.4
	L3_year1_male_young	12032	75.68	
	L3_year1_male_middle	62681	76.98	
	L3_year1_male_senior	19142	82.558	
L2_year2_female		142210	76.5865	76.95
	L3_year2_female_young	17484	73.978	
	L3_year2_female_middle	100255	75.98777	
	L3_year2_female_senior	24471	80.9	
L2_year2_male		81367	77.9	78.01
	L3_year2_male_young	7308	75.578	
	L3_year2_male_middle	51181	76.4	
	L3_year2_male_senior	22878	82.04	
L2_year3_female		468779	77.7	77.238
	L3_year3_female_young	48974	74.45	
	L3_year3_female_middle	259301	76.48	
	L3_year3_female_senior	160504	80.78	
L2_year3_male		290212	79.6	78.56
	L3_year3_male_young	18032	76.3575865	
	L3_year3_male_middle	127962	77.34	
	L3_year3_male_senior	144218	82.008677	

Table 8-9: Estimating parameters at Level 3 Aggregation

Level 3 Stratification	Based on Level 4	N	Old Mean	New Mean
L3_year1_female_young		37390	75.353639	75.1
	L4_year1_female_young_usual	21778	76.46884	
	L4_year1_female_young_non_usual	15612	73.79798	
L3_year1_female_middle		116863	76.507	75.8
	L4_year1_female_middle_usual	77709	77.905	
	L4_year1_female_middle_non_usual	39154	73.733	
L3_year1_female_senior		20681	81.662	80.3
	L4_year1_female_senior_usual	16268	82.647	
	L4_year1_female_senior_non_usual	4413	78.029	
L3_year1_male_young		12032	75.68	75.4
	L4_year1_male_young_usual	6866	77.21	
	L4_year1_male_young_non_usual	5166	73.646	
L3_year1_male_middle		62681	76.98	76.23
	L4_year1_male_middle_usual	41913	78.44	
	L4_year1_male_middle_non_usual	20768	74.029	
L3_year1_male_senior		19142	82.558	81.36
	L4_year1_male_senior_usual	14991	83.483	
	L4_year1_male_senior_non_usual	4151	79.219	
L3_year2_female_young		17484	73.978	73.56
	L4_year2_female_young_usual	10679	75.431	
	L4_year2_female_young_non_usual	6805	71.699	
L3_year2_female_middle		100255	75.987	74.96
	L4_year2_female_middle_usual	69387	77.625	
	L4_year2_female_middle_non_usual	30868	72.305	
L3_year2_female_senior		24471	80.9	79.197
	L4_year2_female_senior_usual	19786	81.96	
	L4_year2_female_senior_non_usual	4685	76.43	
L3_year2_male_young		7308	75.578	75.19
	L4_year2_male_young_usual	4486	76.8897	
	L4_year2_male_young_non_usual	2822	73.493	
L3_year2_male_middle		51181	76.41	75.48
	L4_year2_male_middle_usual	35146	77.97998	
	L4_year2_male_middle_non_usual	16035	72.982	
L3_year2_male_senior		22878	82.04	80.477
	L4_year2_male_senior_usual	18411	83.04	
	L4_year2_male_senior_non_usual	4467	77.915	
L3_year3_female_young		48974	74.45	74.08
	L4_year3_female_young_usual	29090	76.02	
	L4_year3_female_young_non_usual	19884	72.149988	
L3_year3_female_middle		259301	76.48	75.249
	L4_year3_female_middle_usual	184685	78.158	
	L4_year3_female_middle_non_usual	74616	72.339	
L3_year3_female_senior		160504	80.78	79.05
	L4_year3_female_senior_usual	130447	81.82	
	L4_year3_female_senior_non_usual	30057	76.275	
L3_year3_male_young		18032	76.3575865	75.97
	L4_year3_male_young_usual	10879	77.8	
	L4_year3_male_young_non_usual	7153	74.1	
L3_year3_male_middle		127962	77.34	76.159
	L4_year3_male_middle_usual	91651	78.89	
	L4_year3_male_middle_non_usual	36311	73.4	
L3_year3_male_senior		144218	82.008677	80.44
	L4_year3_male_senior_usual	117139	82.95	
	L4_year3_male_senior_non_usual	27079	77.9	

Table 8-10: Level 3 Ranking – Bottom Up

Stratification Level	Old Mean	New Mean	Old Rank	New Rank	R. Change
L3_year1_male_senior	82.558	81.36	1	1	0
L3_year2_male_senior	82.04	80.477	2	2	0
L3_year3_male_senior	82.008677	80.44	3	3	0
L3_year1_female_senior	81.662	80.3	4	4	0
L3_year2_female_senior	80.9	79.197	5	5	0
L3_year3_female_senior	80.78	79.05	6	6	0
L3_year1_male_middle	76.98	76.23	8	7	1
L3_year3_male_middle	77.34	76.159	7	8	1
L3_year3_male_young	76.3575865	75.97	12	9	3
L3_year1_female_middle	76.507	75.8	9	10	1
L3_year2_male_middle	76.41	75.48	11	11	0
L3_year1_male_young	75.68	75.4	14	12	2
L3_year3_female_middle	76.48	75.249	10	13	3
L3_year2_male_young	75.578	75.19	15	14	1
L3_year1_female_young	75.353639	75.1	16	15	1
L3_year2_female_middle	75.987	74.96	13	16	3
L3_year3_female_young	74.45	74.08	17	17	0
L3_year2_female_young	73.978	73.56	18	18	0

*Green = Higher New Ranking, Red = Lower New Ranking

8.3 Doctor Level Analysis

The previous section of this chapter focused on estimating population parameters by taking random samples from a stratified dataset. At each stratification level, researchers can generate random or probability samples by controlling the effect of sociodemographic factor below that level. For example, a researcher may want to estimate the evaluation average of all patients who are seeing their non-usual doctors by taking equal sample size of different age groups. These findings would essentially support researchers who are trying to understand the general patients population by analysing smaller sociodemographic subgroups.

One of the features for large scale healthcare satisfaction survey data is its usually fully nested where patients or raters feedback is collected through the doctor or ratee they are visiting. Patients satisfaction data collected through convenient sampling methodology would have inherited biases caused by the disproportionate sociodemographic distribution that can lead to distort the real evaluation scores. Several statistical techniques such as risk adjustment were developed to provide correction for differences in patient characteristics not under the control of the healthcare provider. These techniques would help improving physician acceptance of patients' feedbacks and decrease the chance that physicians will seek to exclude patients likely to lower their measured scores. However, correcting the disproportionate biases in patients' feedback would first require identifying the most influential and effective sociodemographic characteristics.

In recent years, healthcare studies have called to adopt a stratification approach for analyzing patients feedback as an alternative to risk adjustment. In this approach, the performance of healthcare providers such as clinics and doctors can be measured using multiple performance values (one for each stratum) rather than one overall performance score. For example, researchers can create multiple strata to represent different age groups where patients

evaluation is computed for each stratum. Although researchers acknowledge the advantages of stratification methodology to highlight disparities in healthcare services, there is currently no clear methodology in the literature about how to create systematic stratification methodology or how to consider stratification as a type of risk adjustment.

Applying a stratification approach to identify influential sociodemographic factors at the doctor level can be challenging due to the smaller number of cases available for each doctor. A stratification process would divide each healthcare unit patients into smaller groups of homogeneous subpopulations. If the stratification process involves more than one sociodemographic factor such as gender and age, then stratification becomes more complex and the sample size available for analysis get smaller. Therefore, the stratification methodology presented in chapter six can be used to identify a sociodemographic factor that introduces the highest amount of variance into raters samples.

In this analysis, a small dataset of 150 doctors with each doctor have at least 200 (min = 200, max = 425) patients were extracted from IPQ dataset. The doctors' average values were calculated with scores range from 61.7 (Very Good) to 88.3 (Excellent). Appendix one shows the doctors average score and patients counts. For each doctor, a new average score was calculated by removing the proportionality bias of sociodemographic factors and the differences between the old and new mean scores are calculated. Probability or random sampling techniques can be used to calculate new average score values for each doctor using different demographic variables. Appendices 2 – 5 shows the values of old and new mean scores after correcting for the four sociodemographic factors in IPQ dataset.

Investigative the patients sociodemographic distribution reviles that the majority of doctors have similar distribution to the overall population. Most patients represent the subgroups of middle age females visiting their usual doctor for more than ten years. Examining the mean

different values (New Mean – Old Mean) shows that majority of doctors achieved lower scores after controlling for "Usual Doctor" factor. This is because the patients who are evaluating their usual doctor represent the majority of cases and provide the highest feedback scores. Similarly, the results shows most doctors received higher score values after controlling for the "Gender" factor. This is because female patients who usually provide lower scores represent the majority of cases. Table 8-11 shows the range of mean difference values for the four sociodemographic variables of IPQ dataset. Figure 8-2 shows distribution of mean difference values for usual and gender factors

Table 8-11: Mean Difference Range

Sociodemographic Factor	Smallest value	Highest Value
Usual	3.35	0.9
Age	4.6	3.2
Gender	0.57	1.42
Years Attending	3.2	3.13

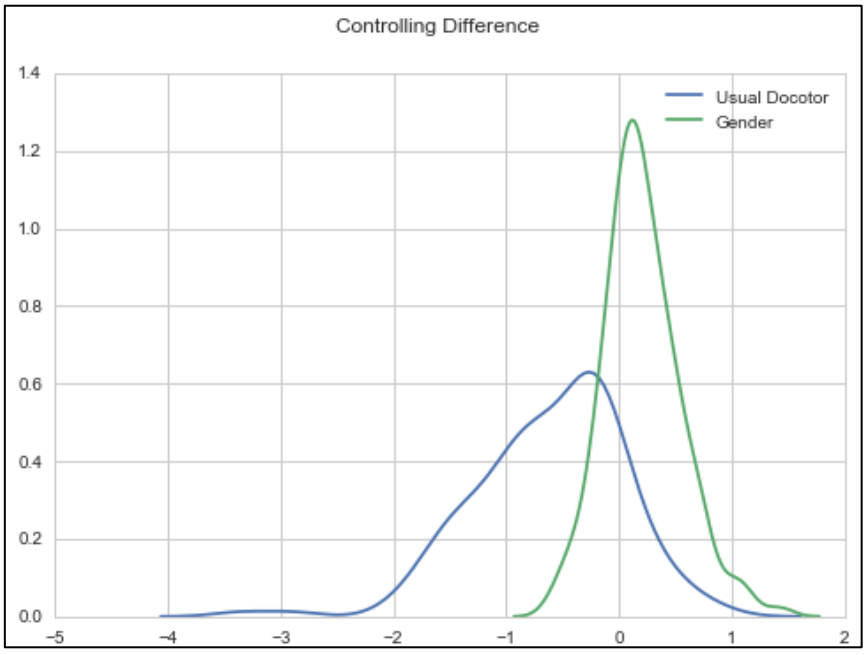

Figure 8-2: Mean Difference Distribution

The results above shows that the majority of doctors would receive similar effect after controlling for different sociodemographic variables. For example, the mean score value for one doctor was lowered (from 75.25 to 71.8). This doctor got the highest percentage of usual patients (83%) which is higher than the general population percentage. However, exploratory analysis on patients demographic distribution also revealed that some doctors may have different demographic distributions than general population demographic distribution. For example, the mean score value for one doctor was increased (from 69.7 to 70.6). This doctor was ranked by majority of non-usual patients, therefore, the same controlling procedure resulted in a higher mean score.

In addition to demographic proportionality, results revealed that some doctors may have different demographic evaluation behaviour than what was highlighted at the patients' level analysis. For instance, a healthcare provider may receive a higher feedback scores from young patients because it could be more equipped to cater for young aged patients than senior patients. Therefore, the demographic distribution and behaviour for individual doctors must be considered when applying probability or random sampling techniques at the doctor level. For example, the sociodemographic factors "Usual Doctor" and "Age Groups" were used to examine the stratification impact at doctor level. It identified a doctor that have "Young and Usual" patients as having the highest mean scores values. Figure 8-3 shows the impact of stratification analysis for all doctors, while figure 8-4 shows stratification impact for a single doctor.

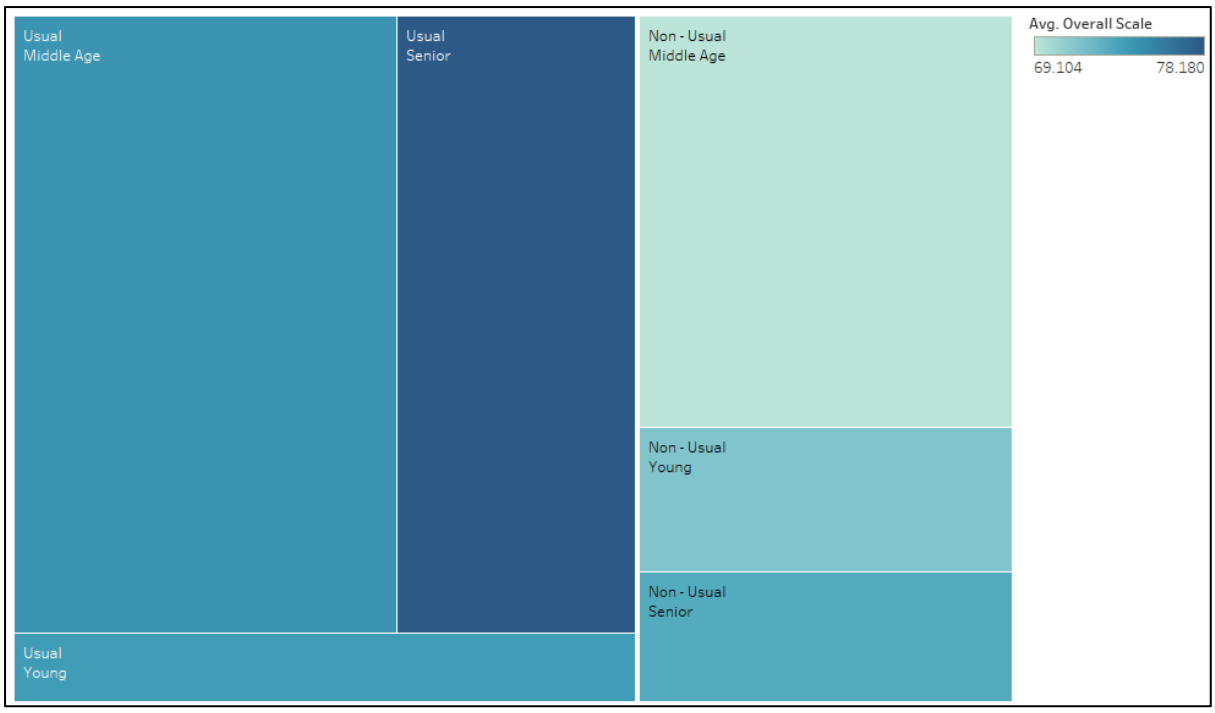

Figure 8-3: Mean Scores with Stratification - All Doctors

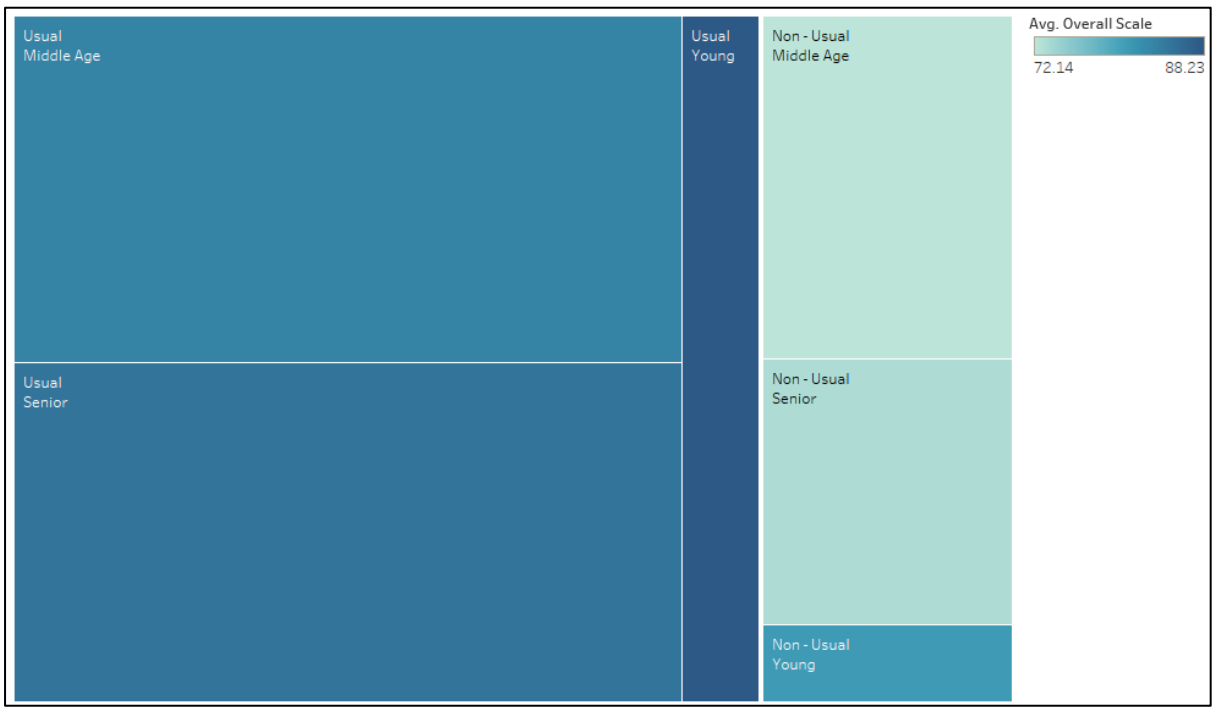

Figure 8-4: Mean Score with Stratification - A Single Doctor

8.4 Discussion

This chapter investigated the possibility of generating pseudo-controlled samples from a pool of mutually exclusive stratified datasets. Researchers can data stratification analysis to estimate population parameters be applying random or probability sampling at different stratification levels. For example, a researcher can use random probability to estimate the population average score by controlling for "Evaluating Usual Doctor" factor. However, the generated score could be biased by the mis proportionality of other demographic factors. Therefore, researchers can select the stratification level they need to estimate a population parameter. Furthermore, a researcher may be interested in estimating a population parameter by controlling for all possible demographic factors. This can be done by applying probability sampling at the leaf stratification level to generate a "Feature Less" patients or raters level score. The stratification and probability sampling techniques can also be used to estimate scores at the aggregated ratees level.

Results showed that controlling for specific sociodemographic factors can have a more significant effect on scores ranking in comparison to other factors. Controlling for the sociodemographic factor "Usual Doctor" would increase the ranking of "young males who have been visiting their doctor for more than ten years" subgroup by three positions; on the other hand, controlling the sociodemographic factor "Years Attending" only introduced minimum change to subgroups ranking. These findings would guide researchers who try to study the effect patients' sociodemographic factors on doctors evaluation results. For researchers trying to generate subgroups with smaller differences in variance, the stratification process should start with an independent variable that maximises the difference such as "Usual / Non-Usual" and ends with a variable that minimises the difference such as "Years Attending" and vice versa. The results showed that aggregating the final level leaf groups would affect how the ranking would change at higher level. If the leaf group is based on variable that

minimizes the difference such as “Years Attending”, then the changes would be much smaller at higher level in comparison to using variable that maximizes the difference such as “Usual / Non-Usual”. Once researchers have a pool of stratified subpopulations, probability or random sampling could be used to create pseudo-controlled samples. The new sampling methodology can be used to for estimating parameters at the general raters or ratees levels. The results and findings presented in this chapter provide enough evidence to support an evidence-based theory that patient satisfaction construct can be inferred by applying rigorous statistical analysis techniques on a large-scale convenience sampling survey dataset. Researchers can apply data stratification and reliability techniques to identify and correct potential sampling biases in patients satisfaction surveys.

The knowledge and stratification rules generated at the patients’ level can be used to estimate doctors’ parameters. However, an exploratory analysis perfumed at the doctor level revealed that the population demographic distribution may not be reflected for individual doctors. For example, some doctors may have a higher proportion of young patients compared to other age groups. An analysis performed to calculate the change in doctor ranking after controlling for different demographic variables highlighted how ranking can change based on demographic distribution attending the specific clinic. In other words, doctors who have a demographic distribution that is different from the population distribution could have an advantage or disadvantage when controlling for sociodemographic factors.

Chapter 9 Conclusion and Future Work

9.1 Research Discussion

This thesis investigated the analysis of large-scale survey datasets collected through the convenient sampling technique. In recent years, this sampling methodology has been embraced by a widening spectrum of domains, including the healthcare and education domains, owing to its cost-effectiveness and the potential operational constraints blocking access to raters' personal data. The analysis utilized a large-scale survey dataset called the 'Improving Practice Questionnaire' (IPQ), which quantifies patients' satisfaction and obtains feedback on their doctor's visit. By examining the satisfaction feedback profiles provided by sociodemographic subpopulations, researchers can assess patients' true perceptions of the healthcare services provided. Such developments have largely reduced the waiting times for treatment and improved the access, support, and information provided to patients. However, despite the increasing popularity of non-probability and convenient sampling techniques, the effect of sociodemographic factors (such as gender, age and race) on patients' satisfaction level remains unknown, because the results of different healthcare studies are inconsistent. In most analyses of patient characteristics, the effect of each sociodemographic attribute (gender, age-group, and other factors) is sequentially analysed on the entire sample population. This methodology provides multiple analyses of individual patients, which may confuse the inferences. Many examples of inconsistent and even contradictory findings on the effects of sociodemographic factors were found in the literature. Today, patient satisfaction feedback is a common performance criterion in programmes involving financial incentives and re-licencing of physicians. To ensure a fair assessment, healthcare researchers have advocated a stratified analysis approach. However, how to construct the strata and perform a drill-down data analysis given the large overlap in sociodemographic distribution is lacking in the literature. As one of

its research questions, this thesis investigates whether machine learning and statistical analysis techniques can be integrated into a systematic stratification methodology.

Theories on patient satisfaction have been largely influenced by empirical studies of satisfaction in job and business environments. Motivated by the importance of customer satisfaction in the business domain, service providers have sought to understand the needs of their customer sub-groups, in particular, how individuals from different sociodemographic groups hold particular expectations. With this knowledge, service providers can focus on specific performance dimensions. Many healthcare providers have adopted patient-centred care policies, in which patients are treated as customers while the physicians and hospital staff are regarded as service providers. Therefore, this thesis also investigates whether stratification analysis on a large-scale patient-satisfaction dataset can generate an evidence-based theory that highlight the differences and similarities in satisfaction patterns between the healthcare and business environments.

Most statisticians consider that convenient sampling data contain inherited biases that degrade the precision of the estimated population parameters. This view is supported by the many examples of inconsistent and contradictory findings on the effects of sociodemographic factors in healthcare satisfaction studies. However, surveys based on probability sampling techniques can also wrongly estimate the population parameters. Probability surveys typically adopt small size, cross-sectional sampling approaches that are non-generalisable to the population at large. Probability sampling also limits the number of sociodemographic factors available for analysis and the amount of rater feedback from each group (which may be insufficient). Therefore, the data of small-scale probability surveys may not reveal the overlapping and inter-correlation feedback among small sub-populations. This was recently demonstrated by the wrong generalisation of many cross-sectional sampling surveys, leading to false predictions of the outcomes of Brexit in the UK and the US presidential elections.

Because convenience sampling is cheaper than other sampling methods, it can acquire large sample datasets. In the healthcare sector, this sampling methodology has become the preferred method for obtaining direct and immediate in-context feedback on patients' experience. In this thesis, we demonstrated that machine learning and standard statistical techniques can improve the data quality over traditional probabilistic techniques. Using drill-down data stratification and reliability techniques, the data were divided into mutually exclusive, non-homogeneous sociodemographic sub-populations. Large-scale big data provide opportunities for quantifying the bias amounts in different strata by novel methods. Therefore, respondents in individual strata can be selected through probability or random sampling, creating pseudo-controlled samples.

The next section discusses the results and outcomes of each research question highlighted in Chapter Three of this thesis.

9.2 Research Contribution

The previous section highlighted several knowledge gaps in current healthcare satisfaction studies. This section discusses the results and findings of the three research questions highlighted in Chapter Three of this thesis.

9.2.1 Research Question One

Can traditional statistical methods combined with machine learning techniques create a systematic data stratification methodology?

Yes, as demonstrated in the data stratification and reliability results (see Chapters Six and Seven), combined machine learning and statistical techniques can create mutually exclusive sub- populations with minimum or maximum variance among the stratum members. The literature review and thesis problem statements highlighted the need for a systematic stratification analysis by which researchers can identify the satisfaction ratings of different sociodemographic subgroups. The big survey data accumulated by non-probability sampling methods provides opportunities for investigating the conflicting and inconsistent findings on how gender, race and other sociodemographic variables affect the satisfaction level of health care. Supervised machine learning identifies the set of independent variables yielding the best prediction model. Guided by entropy and information gain, hierarchical modelling algorithms such as ID3 and C4.5 can divide the search space into smaller subgroups containing similar cases. Meanwhile, numerical supervised learning such as linear regression models the relationships between the output dependent variable and several independent variables. The regression model identifies the contributions of the independent variables by calculating the change in the predicted dependent value per unit change in the independent value.

Survey analysts identify the categorical and numerical target variables that summarise the subjects' feedback in the allocated survey. If no clear summative variable is available, the analyst can engineer a variable that represents some or all of the constructs in the survey. In the exploratory analysis (Chapter Four), mean scores differences for certain sociodemographic factors such as 'Usual Doctor' were statistically significant across all 27 IPQ items in different sample sizes, whereas mean differences with other factors such as 'Years Attending' were statistically insignificant even for larger sample sizes. Guided by linear regression and information gain techniques, the stratification process implemented on both numerical and categorical values divided the patients into the same sub-groups.

The machine learning technique stratifies the survey data into non-homogeneous subpopulations. This technique can generate many strata belonging to different branches, with no statistically significant differences among the subjects in each stratum. Therefore, combining the stratification process with standard statistical techniques can identify the subgroups with minimum variance among the raters, which can be merged into a larger subgroup or signify termination of the branch stratification. Statistical techniques such as ANOVA and subsequent post-hoc analysis can highlight homogeneous subgroups belonging to different branches of the stratification tree. This approach ensures that all raters belong to a mutually exclusive sub-groups, in which each rater is analysed exactly once.

9.2.2 Research Question Two

Can a proposed data stratification methodology create pseudo-controlled samples for estimating population parameters?

Yes, Chapters Six to Eight of this thesis demonstrated that real population parameters can be estimated by creating pseudo-controlled samples from a pool of mutually exclusive sub-populations of raters. Stratification and data reliability analysis can guide the creation of samples using sociodemographic factors that maximise or minimise the variance between the subpopulations. A main criticism of convenient sampling is the inheritance of biases from the uncontrolled proportionality of different sociodemographic factors. As shown in Chapter Eight, controlling for the sub proportionality at each stratification level provides a new mean value for that strata. For example, controlling the proportionality of 'Usual Doctor' at the zeroth stratification level decreased the overall mean by a greater extent than controlling the proportionality of 'Years Attending'. This is consistent with the results, in which the sociodemographic factor 'Usual Doctor' accounted for most of the variance, followed by age, gender and number of years attending.

Using stratification analysis, researchers can also create samples from subgroups with sociodemographic factors in different branches of the stratification tree. For example, in Chapter Six, the responses of patients who were 'Non-usual', 'Senior', 'Female' and 'Over 5 years' were not statistically different from those of patients identifying as 'Usual', 'Young', 'Female' and 'Less than 10 years'. Therefore, these sub-groups can be combined into a single subpopulation despite having large variance factors (such as 'Usual Doctor' and 'Age Group').

Analysing the sociodemographic factors in non-probability sampling surveys by a stratification approach can help researchers to explain and avoid inconsistent findings when reporting the effects of variables such as 'Usual Doctor', 'Age Group' and 'Gender'. When gender factor is

deemed statistically insignificant at the zeroth stratification level, the researcher can infer that gender has less variance than the other factors and emerges as a significant factor at a deeper stratification level. Such drill-down data analysis is possible only when machine learning is combined with standard statistical techniques.

9.2.3 Research Question Three

Can a proposed data stratification methodology create missing values imputation strategy sensitive to sociodemographic sub groups?

Yes, chapters Four and Five of this thesis demonstrated that removing all response cases containing missing answers increased the average satisfaction score by 0.04 points across all survey items, with a higher increase for specific questionnaire components such as staff information and clinic access. Therefore, statistical analysis techniques requiring completely answered cases could be working on biased survey datasets. Moreover, senior patients evaluating their non-usual doctors were more likely to return incomplete questionnaires than other sociodemographic subgroups, such as young and middle-aged patients. When the missing-values patterns were analysed by sociodemographic factors, the general pattern of increased satisfaction scores was inconsistent across the sociodemographic subgroups. The 'Young' and 'Non-usual' subpopulations yielded higher evaluation scores for doctor communication items before removing the missing values responses than afterwards. This indicates that patients are much more reluctant to criticise their personal doctor than other service aspects like clinic access or staff communication. Therefore, by identifying the similarities and differences in the missing-values patterns among sociodemographic sub populations, we can establish an evidence-based imputation strategy of missing values that considers the sociodemographic profiles while including all raters in the analysis.

9.2.4 Novel Survey Design Techniques and Evidence-Based Patient Satisfaction Theory

Chapter Two of this thesis hypothesised a novel evidence-based satisfaction theory from the findings of large-scale, non-probabilistic patient-satisfaction survey data. This hypothesis was proposed to acquire new knowledge on the differences and similarities in healthcare experiences and satisfaction among different sub-populations of raters.

After analysing the IPQ dataset, the factor 'Evaluating Usual Doctor' was associated with the highest information gain and the minimal variance among the rater sub-populations. In addition, principal component analysis on different stratification levels revealed 'Doctor Communication' as the most significant factor in all IPQ sub-populations. These results highlight the importance of building a strong doctor–patient relationship through active communication. The results also contradict earlier findings, in which young patients were associated with lower satisfaction levels. Like patients in any age group, the satisfaction level of young patients was strongly influenced by the communication experience. As shown in Chapter Five, young patients were the least likely sub-population to provide unfinished survey answers, and were keen to establish a communication channel. This knowledge can be integrated into healthcare and medical training programs, encouraging medical professionals to invest more time in researching their patients' background, and thereby improve the communication experience. This evidence-based theory of patient satisfaction outperforms service expectancy theories borrowed from the business domain, which are not entirely applicable to healthcare and medical services.

The research work presented in this thesis highlighted some similarities and differences with earlier research work in the domain of healthcare and patients' satisfaction. The literature review explains that the roots of patient satisfaction studies were influenced by 'customer

satisfaction' studies in business environments. However, the literature also shows that after several decades of research in healthcare services satisfaction, a clear theory that explains the influences of sociodemographic variables on patient satisfaction is still lacking. Therefore, the novel evidence-based satisfaction theory presented in this thesis would help researchers to design suitable data stratification and segmentation strategies that account for the inherited sociodemographic sampling biases in patient satisfaction data. Such a theory would accept the statistical limitations of non-probability sampling in a rigorous drill-down data stratification and reliability analysis. Evidence-based theory aims to reveal the similarities and differences in satisfaction experience among different sub-populations of raters. This evidence-based theory of patient satisfaction outperforms service expectancy theories borrowed from the business domain, which are not entirely applicable to healthcare and medical services. The application of such theory can extend the healthcare domain into other research areas such as education, business and so on.

There are also several lessons and recommendations can be derived from this study. Correlation analysis implemented in chapter four showed that no single item qualifies as a summative variable for IPQ survey dataset. Survey designers can consider adding overall summative items as well as within each survey section. This research presented some practical suggestions to feature engineer categorical and numerical target variables that represents some or all of survey constructs. In addition to that, the analysis results and the suggested stratification methodology presented in this thesis will help survey analysis studies that depend on non-probabilistic sampling techniques to investigating the conflicting and inconsistent findings on how gender, race and other sociodemographic variables affect raters feedback and satisfaction levels. The knowledge generated from this study can potentially be used to improve patient experiences of practice and healthcare services. At the policy level, understanding the determinants of healthcare satisfaction and obtaining feedback on patient experience can help decision makers

to remove potential disparities among different patients or subgroups of survey raters. For example, the findings presented in this research integrated into healthcare and medical training programs, encouraging medical professionals to invest more time in researching their patients' background, and thereby improve the communication experience.

9.3 Limitations

This thesis has proposed the first step towards generalising knowledge of population parameters estimated from non-probability sampling data. Combining statistical and machine learning techniques was useful for highlighting the hidden patterns introduced by sociodemographic factors when estimating patient's satisfaction with their primary healthcare providers, which otherwise remain unknown. A potential limitation of this research is how the stratification methodology can be applied to other survey studies in other domains. Non-probability convenient sampling techniques are currently the preferred method for collecting satisfaction feedback in healthcare, business and education.

Another potential limitation is the guidance of the stratification process by engineered summative variables. After examining both the inter-correlations and the wording of the survey items, it was found that no single item qualifies as a summative variable for the survey dataset. Therefore, all survey items were summarized into a single-valued overall summative scale item, which assigns equal weight or importance to all survey items. The summative item exhibited a negative skewness, like the original distribution shape of the 27 items. However, other research studies inserted a summative item in the original survey design.

Another limitation of the present results is the small number of categories in each sociodemographic factor. All factors contained two or three categories, whereas other sociodemographic factors such as ethnicity and language can contain many categories.

Different numbers of categories can potentially impact the calculated entropy and information gain that guide the stratification process.

In summary, the proposed drill-down data stratification analysis assumes several prerequisites related to data presentation and processing. Identifying the impacts of different sociodemographic factors and their mutually exclusive sub-populations requires at least two sociodemographic factors per rater. For compatibility with standard statistical calculations, the analysis also assumes that rater feedback is provided in an ordinal or Likert-scale variable format. Many of the abovementioned research limitations could be tackled in further development of the stratification methodology for analysing non-probability sampling data.

9.4 Future Work

In future research, the proposed stratification methodology could be improved in several ways. The consistency of identifying mutually exclusive and nonhomogeneous subgroups by machine learning and statistical techniques should be tested on other large-scale non-probability sampling data in the healthcare and business domains. The number of categories in diverse sociodemographic factors could be reduced by data segmentation as a pre-processing step. Other possible expansions are adjustment to the multi-level reliability calculations presented in Chapter Eight, which constantly reduce the average number of raters per ratee during the drill down stratification analysis, and minimising the impact of large multi-valued attributes using the information gain ratio. The latter improvement considers the number and size of the stratification branches when choosing an attribute. The stratification methodology presented in this thesis can support the healthcare service in New Zealand through the use of a systematic methodology to identify the satisfaction level of small population sub groups that otherwise may remain unknown. Also, the research can support New Zealand healthcare service by adding sub population feedback and satisfaction levels as an additional priority

factor. However, further research is needed to examine the validity of this approach within NZ context. The stratification analysis presented in this thesis can be also adapted to the research area of recommendation systems by considering users sociodemographic characteristics. The recommendations can depend on the preferences of smaller populations subgroups in addition the selections similarities factor. However, data privacy issues must be considered in this type of applications.

9.5 Summary

This thesis investigated the opportunities and challenges associated with the increasing popularity of the large-scale non-probability sampling methodology. The effects of sociodemographic factors reported in business and healthcare studies are inconsistent and sometimes contradictory. The consensus belief is that the results of non-probability sampling cannot precisely estimate the differences in statistical properties of the sample and the population, because of inherent biases in the non-probability sample. Recent studies have highlighted the need for a stratified approach when analysing large scale survey feedback. However, a methodology for developing a systematic stratification process has not been clarified in the literature. This thesis presented a stratification methodology that identifies the feedback profiles of small subpopulations by combined machine learning and standard statistical techniques. The combined methods detect and adjust the inherited biases in convenience sampling.

The research work presented in this thesis investigated the feasibility of combining standard statistical and machine learning techniques into a systematic stratification methodology to analysis survey data collected through non-probability sampling. The objective of the new methodology is to divides the raters (‘population’) into mutually exclusive and collectively exhaustive subpopulations for individual and comparative analysis. A quantifiable

stratification analysis can reveal whether the performance outcomes depend on one or more specific sociodemographic factors. It also identifies and facilitates the reduction of sociodemographic disparities. The literature review presented in chapter two highlighted there is currently no clear guidelines for implementing a systematic stratification approach. The results from this analysis also presented a novel and evidence-based satisfaction theory that accounts for the inherited sociodemographic sampling biases in patients' satisfaction data.

Reference

[1] R. A. Carr-Hill, "The measurement of patient satisfaction," *J. Public Health (Bangkok).*, vol. 14, no. 3, pp. 236–249, 1992.

[2] R. Baker and M. Whitfield, "Measuring patient satisfaction: a test of construct validity.," *Qual. Saf. Heal. Care*, vol. 1, no. 2, pp. 104–109, 1992.

[3] H. Crow *et al.*, "Measurement of satisfaction with health care: Implications for practice from a systematic review of the literature," *Health Technol. Assess. (Rockv).*, 2002.

[4] R. Bilberg, B. Nørgaard, S. Overgaard, and K. K. Roessler, "Patient anxiety and concern as predictors for the perceived quality of treatment and patient reported outcome (PRO) in orthopaedic surgery," *BMC Health Serv. Res.*, vol. 12, no. 1, p. 244, 2012.

[5] M. R. Hasan, "Productivity and performance in NHS hospitals," *BMJ*, vol. 340, p. c1776, 2010.

[6] C. Ham, "Improving the performance of the English NHS." British Medical Journal Publishing Group, 2010.

[7] N. H. S. Employers, G. P. Committee, and others, "Quality and Outcomes Framework guidance for GMS contract 2009/10," *Deliv. Invest. Gen. Pract. (The NHS Confed. Co. Ltd, London)*, 2009.

[8] M. Greco, R. Powell, and K. Sweeney, "The Improving Practice Questionnaire (IPQ): a practical tool for general practices seeking patient views," *Educ. Prim. Care*, vol. 14, no. 4, pp. 440–448, 2003.

[9] C. Paddison, G. A. Abel, M. Roland, M. N. Elliott, G. Lyratzopoulos, and J. L. Campbell, "Drivers of Overall Satisfaction with Primary Care," 2013.

[10] C. A. M. Paddison, G. A. Abel, M. O. Roland, M. N. Elliott, G. Lyratzopoulos, and J. L. Campbell, "Drivers of overall satisfaction with primary care: evidence from the English General Practice Patient Survey," *Heal. Expect.*, vol. 18, no. 5, pp. 1081–1092, 2015.

[11] M. Roland, M. Roberts, V. Rhenius, and J. L. Campbell, "GPAQ-R: development and psychometric properties of a version of the General Practice Assessment Questionnaire for use for revalidation by general practitioners in the UK," *BMC Fam. Pract.*, vol. 14, no. 1, p. 160, 2013.

[12] S. Reimann and D. Strech, "The representation of patient experience and satisfaction in physician rating sites. A criteria-based analysis of English-and German-language sites," *BMC Health Serv. Res.*, vol. 10, no. 1, p. 332, 2010.

[13] C. Salisbury, M. Wallace, and A. A. Montgomery, "Patients experience and satisfaction in primary care: secondary analysis using multilevel modelling," *BmJ*, vol. 341, p. c5004, 2010.

[14] G. Lyratzopoulos *et al.*, "Understanding ethnic and other socio-demographic differences in patient experience of primary care: evidence from the English General Practice Patient Survey," *BMJ Qual Saf*, vol. 21, no. 1, pp. 21–29, 2012.

[15] J. E. Croker, D. R. Swancutt, M. J. Roberts, G. A. Abel, M. Roland, and J. L. Campbell, "Factors affecting patients trust and confidence in GPs: evidence from the English national GP patient survey," *BMJ Open*, vol. 3, no. 5, p. e002762, 2013.

[16] O. A. Bjertnaes, I. S. Sjetne, and H. H. Iversen, "Overall patient satisfaction with hospitals: effects of patient-reported experiences and fulfilment of expectations," *BMJ Qual Saf*, vol. 21, no. 1, pp. 39–46, 2012.

[17] J. R. A. Santos, "Cronbach's alpha: A tool for assessing the reliability of scales," *J. Ext.*, vol. 37, no. 2, pp. 1–5, 1999.

[18] V. Vehovar, V. Toepoel, and S. Steinmetz, "Non-probability Sampling," *SAGE Handb. Surv. Methodol.*, p. 329, 2016.

[19] S. O. Becker, T. Fetzer, and D. Novy, "Who voted for Brexit? A comprehensive district-level analysis," *Econ. Policy*, vol. 32, no. 92, pp. 601–650, 2017.

[20] G. Terhanian, "What Survey Researchers Can Learn From the 2016 US Pre-Election Polls: How to Fine-Tune Survey Methods And Restore Credibility," *J. Advert. Res.*, vol. 57, no. 2, pp. 182–189, 2017.

[21] K. French, "Methodological considerations in hospital patient opinion surveys," *Int. J. Nurs. Stud.*, vol. 18, no. 1, pp. 7–32, 1981.

[22] National Quality Forum, *Risk adjustment for sociodemographic factors*. 2014.

[23] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[24] "Improving quality and safety through partnerships with patients and consumers."

[25] S. L. Lavela, "Evaluation and measurement of patient experience," vol. 1, no. 1, 2014.

[26] J. K. HJ Sixma, "Quality of care from the patients' perspective: from theoretical concept to a new measuring instrument." .

[27] V. Espeland and O. Indrehus, "Evaluation of students' satisfaction with nursing education in Norway," *J. Adv. Nurs.*, vol. 42, no. 3, pp. 226–236, 2003.

[28] S. T. Wong and J. Haggerty, "Measuring patient experiences in primary health care," *Vancouver UBC Cent. Heal. Serv. Policy Res.*, 2013.

[29] A. V. S. J. F. de R. A. Eguskiza P Arrate A and J. A., "User satisfaction with primary care teams: relationship of satisfaction to the doctor's training in the field of doctors patient relations," *Aten. Primaria*, vol. 16, no. 1, pp. 45–50, 1995.

[30] J. L. Campbell, J. Ramsay, and J. Green, "Age, gender, socioeconomic, and ethnic differences in patients9 assessments of primary health care," *Qual. Saf. Heal. Care*, vol. 10, no. 2, pp. 90–95, 2001.

[31] S. Linder-Pelz, "Toward a theory of patient satisfaction," *Soc. Sci. Med.*, vol. 16, no. 5, pp. 577–582, 1982.

[32] M. Negi and D. P. Kaur, "A Study Of Customer Satisfaction With Life Insurance In Chandigarh Tricity," *Paradigm*, vol. 14, no. 2, pp. 29–44, 2010.

[33] G. A. Churchill Jr and C. Surprenant, "An investigation into the determinants of customer satisfaction," *J. Mark. Res.*, pp. 491–504, 1982.

[34] S. Abramowitz *et al.*, "Consumer satisfaction, dissatisfaction, and complaining behavior," *Soc. Sci. Med.*, vol. 14, no. 1, pp. 1005–1026, 2013.

[35] R. L. Day, "Extending the concept of consumer satisfaction," *ACR North Am. Adv.*, 1977.

[36] R. L. Day, "Research perspectives on consumer complaining behavior," *Theor. Dev. Mark.*, pp. 211–215, 1980.

[37] R. L. Day, "Modeling choices among alternative responses to dissatisfaction," *ACR North Am. Adv.*, 1984.

[38] Y. Yi, "A critical review of consumer satisfaction," *Rev. Mark.*, vol. 4, no. 1, pp. 68–123, 1990.

[39] J. A. Miller, "Studying satisfaction, modifying models, eliciting expectations, posing problems, and making meaningful measurements," *Conceptualization Meas. Consum. Satisf. dissatisfaction*, pp. 72–91, 1977.

[40] R. L. Day, "Toward a process model of consumer satisfaction," *Conceptualization Meas. Consum. Satisf. dissatisfaction*, pp. 153–183, 1977.

[41] R. W. Olshavsky and J. A. Miller, "Consumer expectations, product performance, and perceived product quality," *J. Mark. Res.*, pp. 19–21, 1972.

[42] R. L. Oliver, "A theoretical reinterpretation of expectation and disconfirmation effects on posterior product evaluation: Experiences in the field," *Consum. Satisf. dissatisfaction Complain. Behav.*, pp. 2–9, 1977.

[43] A. Yüksel and F. Yüksel, "The expectancy-disconfirmation paradigm: a critique," *J. Hosp. Tour. Res.*, vol. 25, no. 2, pp. 107–131, 2001.

[44] R. L. Oliver, "A cognitive model of the antecedents and consequences of satisfaction decisions," *J. Mark. Res.*, pp. 460–469, 1980.

[45] J. Kandampully and D. Suhartanto, "Customer loyalty in the hotel industry: the role of customer satisfaction and image," *Int. J. Contemp. Hosp. Manag.*, vol. 12, no. 6, pp. 346–351, 2000.

[46] Z. S. Dimitriades, "Customer satisfaction, loyalty and commitment in service organizations: Some evidence from Greece," *Manag. Res. News*, vol. 29, no. 12, pp. 782–800, 2006.

[47] P. Ramseook-Munhurrun, V. N. Seebaluck, and P. Naidoo, "Examining the structural relationships of destination image, perceived value, tourist satisfaction and loyalty: case of Mauritius," *Procedia-Social Behav. Sci.*, vol. 175, pp. 252–259, 2015.

[48] M. Söderlund, "Customer satisfaction and its consequences on customer behaviour revisited: The impact of different levels of satisfaction on word-of-mouth, feedback to the supplier and loyalty," *Int. J. Serv. Ind. Manag.*, vol. 9, no. 2, pp. 169–188, 1998.

[49] A. S. Mattila, A. A. Grandey, and G. M. Fisk, "The interplay of gender and affective tone in service encounter satisfaction," *J. Serv. Res.*, vol. 6, no. 2, pp. 136–143, 2003.

[50] C. Homburg and A. Giering, "Personal characteristics as moderators of the relationship between customer satisfaction and loyalty: an empirical analysis," *Psychol. Mark.*, vol. 18, no. 1, pp. 43–66, 2001.

[51] V. Mittal and W. A. Kamakura, “Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics,” *J. Mark. Res.*, vol. 38, no. 1, pp. 131–142, 2001.

[52] H. Evanschitzky and M. Wunderlich, “An examination of moderator effects in the four-stage loyalty model,” *J. Serv. Res.*, vol. 8, no. 4, pp. 330–345, 2006.

[53] P. G. Patterson, “Demographic correlates of loyalty in a service context,” *J. Serv. Mark.*, vol. 21, no. 2, pp. 112–121, 2007.

[54] M. K. Brady, G. A. Knight, J. J. Cronin, G. Tomas, M. Hult, and B. D. Keillor, “Removing the contextual lens: A multinational, multi-setting comparison of service evaluation models,” *J. Retail.*, vol. 81, no. 3, pp. 215–230, 2005.

[55] B.-L. Cheng, “Service quality and the mediating effect of corporate image on the relationship between customer satisfaction and customer loyalty in the Malaysian hotel industry,” *Gadjah Mada Int. J. Bus.*, vol. 15, no. 2, pp. 99–112, 2013.

[56] P. Sharma, I. S. N. Chen, and S. T. K. Luk, “Gender and age as moderators in the service evaluation process,” *J. Serv. Mark.*, vol. 26, no. 2, pp. 102–114, 2012.

[57] J. Sitzia and N. Wood, “Patient satisfaction with cancer chemotherapy nursing: a review of the literature,” *Int. J. Nurs. Stud.*, vol. 35, no. 1–2, pp. 1–12, 1998.

[58] L. Strasen, “Incorporating patient satisfaction standards into quality of care measures.,” *J. Nurs. Adm.*, vol. 19, no. 11, pp. 5–6, 1989.

[59] G. C. Pascoe, “Patient satisfaction in primary health care: a literature review and analysis,” *Eval. Program Plann.*, vol. 6, no. 3–4, pp. 185–210, 1983.

[60] R. N. Axon and M. V Williams, “Hospital Readmission as an Accountability Measure,” *JAMA*, vol. 305, no. 5, p. 504, Feb. 2011.

[61] D. Locker and D. Dunt, “Theoretical and methodological issues in sociological studies of consumer satisfaction with medical care,” *Soc. Sci. Med. Part A Med. Psychol. Med. Sociol.*, vol. 12, pp. 283–292, 1978.

[62] J. Sitzia and N. Wood, “Patient satisfaction: a review of issues and concepts,” *Soc. Sci. Med.*, vol. 45, no. 12, pp. 1829–1843, 1997.

[63] K. Collins and P. Nicolson, “The meaning of satisfaction for people with dermatological problems: Reassessing approaches to qualitative health psychology research,” *J. Health Psychol.*, vol. 7, no. 5, pp. 615–629, 2002.

[64] R. Fitzpatrick and A. Hopkins, “Problems in the conceptual framework of patient satisfaction research: an empirical exploration,” *Sociol. Health Illn.*, vol. 5, no. 3, pp. 297–311, 1983.

[65] L. Gill and L. White, “A critical review of patient satisfaction,” *Leadersh. Heal. Serv.*, vol. 22, no. 1, pp. 8–19, 2009.

[66] J. G. Fox and D. M. Storms, “A different approach to sociodemographic predictors of satisfaction with health care,” *Soc. Sci. Med. Part A Med. Psychol. Med. Sociol.*, vol. 15, no. 5, pp. 557–564, 1981.

[67] J. E. Ware, M. K. Snyder, W. R. Wright, and A. R. Davies, “Defining and measuring patient satisfaction with medical care,” *Eval. Program Plann.*, vol. 6, no. 3–4, pp. 247–

263, 1983.

[68] A. Donabedian, “The definition of quality and approaches to its assessment,” *Explor. Qual. Assess. Monit.*, 1980.

[69] S. Abramowitz, A. A. Coté, and E. Berry, “Analyzing patient satisfaction: a multianalytic approach,” *QRB-Quality Rev. Bull.*, vol. 13, no. 4, pp. 122–130, 1987.

[70] B. Williams, “Patient satisfaction: a valid concept?,” *Soc. Sci. Med.*, vol. 38, no. 4, pp. 509–516, 1994.

[71] M. E. Reid, “Going to See the Doctor: The Consultation Process in General Practice,” *Sociology*, vol. 10, no. 1, pp. 192–193, 1976.

[72] D. E. Larsen and I. Rootman, “Physician role performance and patient satisfaction,” *Soc. Sci. Med.*, vol. 10, no. 1, pp. 29–32, 1976.

[73] A. Donabedian, “The Lichfield Lecture. Quality assurance in health care: consumers’ role.,” *Qual. Heal. care*, vol. 1, no. 4, p. 247, 1992.

[74] R. L. Kravitz, “Patients’ expectations for medical care: an expanded formulation based on review of the literature,” *Med. Care Res. Rev.*, vol. 53, no. 1, pp. 3–27, 1996.

[75] D. S. Brody, S. M. Miller, C. E. Lerman, D. G. Smith, C. G. Lazaro, and M. J. Blum, “The relationship between patients’ satisfaction with their physicians and perceptions about interventions they desired and received,” *Med. Care*, pp. 1027–1035, 1989.

[76] M. Haas, “The relationship between expectations and satisfaction: a qualitative study of patients’ experiences of surgery for gynaecological cancer,” *Heal. Expect.*, vol. 2, no. 1, pp. 51–60, 1999.

[77] J. E. Ware Jr and R. D. Hays, “Methods for measuring patient satisfaction with specific medical encounters.,” *Med. Care*, vol. 26, no. 4, pp. 393–402, 1988.

[78] A. R. Davies and J. E. Ware, *GHAA’s consumer satisfaction survey and user’s manual*. GHAA, 1991.

[79] A. Narayanan, M. Greco, P. Reeves, A. Matthews, and J. Bergin, “Community pharmacy performance evaluation: Reliability and validity of the Pharmacy Patient Questionnaire,” *Int. J. Healthc. Manag.*, vol. 7, no. 2, pp. 103–119, 2014.

[80] C. Jenkinson, A. Coulter, S. Bruster, N. Richards, and T. Chandola, “Patients’ experiences and satisfaction with health care: results of a questionnaire study of specific aspects of care,” *Qual Saf Heal. Care*, vol. 11, no. 4, pp. 335–339, 2002.

[81] S. J. Williams and M. Calnan, “Key determinants of consumer satisfaction with general practice,” *Fam. Pract.*, vol. 8, no. 3, pp. 237–242, 1991.

[82] N. Mead and M. Roland, “Understanding why some ethnic minority patients evaluate medical care more negatively than white patients: a cross sectional analysis of a routine patient survey in English general practices,” *BMJ*, vol. 339, p. b3450, 2009.

[83] S. N. Bleich, E. Özaltin, and C. J. L. Murray, “How does satisfaction with the health-care system relate to patient experience?,” *Bull. World Health Organ.*, vol. 87, no. 4, pp. 271–278, 2009.

[84] K. A. Collins, “The meaning of patient satisfaction: re-assessing a qualitative

psychological research methodology," University of Sheffield, 2002.

[85] a Narayanan and M. Greco, "The Dental Practice Questionnaire: a patient feedback tool for improving the quality of dental practices.," *Aust. Dent. J.*, vol. 59, no. 3, pp. 334–48, Sep. 2014.

[86] J. A. Hall and M. C. Dornan, "Patient sociodemographic characteristics as predictors of satisfaction with medical care: a meta-analysis," *Soc. Sci. Med.*, vol. 30, no. 7, pp. 811–818, 1990.

[87] R. Fitzpatrick, "Surveys of patients satisfaction: I--Important general considerations.," *BMJ Br. Med. J.*, vol. 302, no. 6781, p. 887, 1991.

[88] W. E. Saris, J. A. Krosnick, and E. M. Shaeffer, "Comparing questions with agree/disagree response options to questions with construct-specific response options," *Unpubl. manuscript, Polit. Soc. Cult. Sci. Univ. Amsterdam*, 2005.

[89] D. J. Owens and C. Batchelor, "Patient satisfaction and the elderly," *Soc. Sci. Med.*, vol. 42, no. 11, pp. 1483–1491, 1996.

[90] D. A. Revicki, "Patient assessment of treatment satisfaction: methods and practical issues," *Gut*, vol. 53, no. suppl 4, p. iv40--iv44, 2004.

[91] S. J. Gillam, A. N. Siriwardena, and N. Steel, "Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework: a systematic review," *Ann. Fam. Med.*, vol. 10, no. 5, pp. 461–468, 2012.

[92] C. Wright *et al.*, "Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council patient and colleague questionnaires," *Acad. Med.*, vol. 87, no. 12, pp. 1668–1678, 2012.

[93] M. A. Bunge, "Treatise on Basic Philosophy: Ontology II: A World of Systems. Dordrecht, Holland: D." Reidel Publishing Company, 1979.

[94] P. J. Lavrakas, *Encyclopedia of survey research methods*. Sage Publications, 2008.

[95] S. M. Campbell, M. O. Roland, and S. A. Buetow, "Defining quality of care," *Soc. Sci. Med.*, vol. 51, no. 11, pp. 1611–1625, 2000.

[96] W. Hogg, M. Rowan, G. Russell, R. Geneau, and L. Muldoon, "Framework for primary care organizations: the importance of a structural domain," *Int. J. Qual. Heal. Care*, vol. 20, no. 5, pp. 308–313, 2007.

[97] J. Haggerty *et al.*, "Operational definitions of attributes of primary health care: consensus among Canadian experts," *Ann. Fam. Med.*, vol. 5, no. 4, pp. 336–344, 2007.

[98] A. L. Stewart, A. Nápoles-Springer, and E. J. Pérez-Stable, "Interpersonal processes of care in diverse populations," *Milbank Q.*, vol. 77, no. 3, pp. 305–339, 1999.

[99] D. McMurchy, *What are the critical attributes and benefits of a high-quality primary healthcare system?* Canadian Health Services Research Foundation= Fondation canadienne de la recherche sur les Services de sant{é}, 2009.

[100] D. H. Thom *et al.*, "Physician trust in the patient: development and validation of a new measure," *Ann. Fam. Med.*, vol. 9, no. 2, pp. 148–154, 2011.

[101] J. H. Hibbard, J. Stockard, E. R. Mahoney, and M. Tusler, "Development of the Patient

Activation Measure (PAM): conceptualizing and measuring activation in patients and consumers," *Health Serv. Res.*, vol. 39, no. 4p1, pp. 1005–1026, 2004.

[102] S. T. Wong, S. Peterson, and C. Black, "Patient activation in primary healthcare: a comparison between healthier individuals and those with a chronic illness," *Med. Care*, pp. 469–479, 2011.

[103] K. Checkland, M. Marshall, and S. Harrison, "Re-thinking accountability: trust versus confidence in medical practice," *Qual. Saf. Heal. Care*, vol. 13, no. 2, pp. 130–135, 2004.

[104] N. Dalkey and O. Helmer, "An experimental application of the Delphi method to the use of experts," *Manage. Sci.*, vol. 9, no. 3, pp. 458–467, 1963.

[105] B. S. Hulka, S. J. Zyzanski, J. C. Cassel, and S. J. Thompson, "Scale for the measurement of attitudes toward physicians and primary medical care," *Med. Care*, vol. 8, no. 5, pp. 429–436, 1970.

[106] J. L. Campbell, P. Smith, S. Nissen, P. Bower, M. Elliott, and M. Roland, "The GP Patient Survey for use in primary care in the National Health Service in the UK--development and psychometric characteristics," *BMC Fam. Pract.*, vol. 10, no. 1, p. 57, 2009.

[107] J. L. Campbell, S. H. Richards, A. Dickens, M. Greco, A. Narayanan, and S. Brearley, "Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires," *Qual. Saf. Heal. Care*, vol. 17, no. 3, pp. 187–193, 2008.

[108] R. J. Shavelson and N. M. Webb, *Generalizability theory: A primer*. Sage, 1991.

[109] A. Narayanan, M. Greco, H. Powell, and L. Coleman, "The Reliability of Big 'Patient Satisfaction' Data," *Big Data*, vol. 1, no. 3, pp. 141–151, Sep. 2013.

[110] A. Narayanan, M. Greco, H. Powell, and L. Coleman, "The Reliability of Big Patient Satisfaction Data," *Big data*, vol. 1, no. 3, pp. 141–151, 2013.

[111] A. Narayanan, M. Greco, and J. L. Campbell, "Generalisability in unbalanced , uncrossed and fully nested studies," pp. 367–378, 2010.

[112] J. Brick and G. Kalton, "Handling missing data in survey research," *Stat. Methods Med. Res.*, vol. 5, no. 3, pp. 215–238, Sep. 1996.

[113] R. j. A. Little and D. b. Rubin, "The Analysis of Social Science Data with Missing Values," *Sociol. Methods Res.*, vol. 18, no. 2–3, pp. 292–326, Nov. 1989.

[114] G. Kalton and D. Kasprzyk, "Imputing for missing survey responses," in *Proceedings of the section on survey research methods, American Statistical Association*, 1982, vol. 22, p. 31.

[115] O. Harel, R. Zimmerman, and O. Dekhtyar, *Approaches to the handling of missing data in communication research*. University of Connecticut, Department of Statistics, 2007.

[116] T. A. Myers, "Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data," *Commun. Methods Meas.*, vol. 5, no. 4, pp. 297–310, 2011.

[117] C. C. Thiedke, "What do we really know about patient satisfaction?," *Fam. Pract. Manag.*, vol. 14, no. 1, p. 33, 2007.

[118] H. J. Sixma, P. M. M. Spreeuwenberg, and M. A. A. van der Pasch, "Patient satisfaction with the general practitioner: a two-level analysis," *Med. Care*, vol. 36, no. 2, pp. 212–229, 1998.

[119] S. J. Notaro *et al.*, "Analysis of the demographic characteristics and medical conditions of the uninsured utilizing a free clinic," *J. Community Health*, vol. 37, no. 2, pp. 501–506, 2012.

[120] L. Moret, J.-M. Nguyen, C. Volteau, B. Falissard, P. Lombrail, and I. Gasquet, "Evidence of a non-linear influence of patient age on satisfaction with hospital care," *Int. J. Qual. Heal. Care*, vol. 19, no. 6, pp. 382–389, 2007.

[121] M. N. Elliott, W. G. Lehrman, M. K. Beckett, E. Goldstein, K. Hambarsoomian, and L. A. Giordano, "Gender differences in patients' perceptions of inpatient care," *Health Serv. Res.*, vol. 47, no. 4, pp. 1482–1501, 2012.

[122] D. C. Mcfarland, K. A. Ornstein, and R. F. Holcombe, "Demographic factors and hospital size predict patient satisfaction variance-implications for hospital value-based purchasing," *J. Hosp. Med.*, vol. 10, no. 8, pp. 503–509, 2015.

[123] S. A. Robert, A. Trentham-Dietz, J. M. Hampton, J. A. McElroy, P. A. Newcomb, and P. L. Remington, "Socioeconomic risk factors for breast cancer: distinguishing individual-and community-level effects," *Epidemiology*, vol. 15, no. 4, pp. 442–450, 2004.

[124] K. M. Bennett, J. E. Scarborough, T. N. Pappas, and T. B. Kepler, "Patient socioeconomic status is an independent predictor of operative mortality," *Ann. Surg.*, vol. 252, no. 3, pp. 552–558, 2010.

[125] G. J. J. W. Bours, R. J. G. Halfens, M. P. F. Berger, H. H. Abu-Saad, and R. T. P. M. Grol, "Development of a model for case-mix adjustment of pressure ulcer prevalence rates," *Med. Care*, vol. 41, no. 1, pp. 45–55, 2003.

[126] J. Dimick, J. Ruhter, M. V. Sarrazin, and J. D. Birkmeyer, "Black patients more likely than whites to undergo surgery at low-quality hospitals in segregated regions.," *Health Aff. (Millwood).*, vol. 32, no. 6, pp. 1046–1053, Jun. 2013.

[127] E. C. Schneider, A. M. Zaslavsky, and A. M. Epstein, "Racial disparities in the quality of care for enrollees in Medicare managed care," *Jama*, vol. 287, no. 10, pp. 1288–1294, 2002.

[128] A. M. Zaslavsky and A. M. Epstein, "How patients sociodemographic characteristics affect comparisons of competing health plans in California on HEDIS®quality measures," *Int. J. Qual. Heal. Care*, vol. 17, no. 1, pp. 67–74, 2005.

[129] A. M. Zaslavsky *et al.*, "Impact of sociodemographic case mix on the HEDIS measures of health plan quality," *Med. Care*, pp. 981–992, 2000.

[130] K. E. Joynt, R. Zuckerman, and A. M. Epstein, "Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs," *Circ. Cardiovasc. Qual. Outcomes*, vol. 10, no. 5, p. e003587, 2017.

[131] V. Parsonnet, D. Dean, and A. D. Bernstein, "A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease.," *Circulation*, vol. 79, no. 6 Pt 2, pp. I3--12, 1989.

[132] J. S. House, J. M. Lepkowski, A. M. Kinney, R. P. Mero, R. C. Kessler, and A. R. Herzog, "The social stratification of aging and health," *J. Health Soc. Behav.*, pp. 213–234, 1994.

[133] J. Collinson *et al.*, "Clinical outcomes, risk stratification and practice patterns of unstable angina and myocardial infarction without ST elevation: Prospective Registry of Acute Ischaemic Syndromes in the UK (PRAIS-UK)," *Eur. Heart J.*, vol. 21, no. 17, pp. 1450–1457, 2000.

[134] N. M. Nasrabadi, "Pattern recognition and machine learning," *J. Electron. Imaging*, vol. 16, no. 4, p. 49901, 2007.

[135] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161–168.

[136] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, 2015.

[137] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[138] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[139] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[140] J. Fox, *Applied regression analysis and generalized linear models*. Sage Publications, 2015.

[141] F. De Battisti, G. Nicolini, and S. Salini, "The Rasch model to measure service quality," *ICFAI J. Serv. …*, no. September, pp. 58–80, 2005.

[142] R. L. Brennan, "Generalizability theory," *Educ. Meas. Issues Pract.*, vol. 11, no. 4, pp. 27–34, 1992.

[143] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[144] J. Cohen, "Statistical power analysis for the behavioral sciences 2nd edn." Erlbaum Associates, Hillsdale, 1988.

[145] L. Crocker and J. Algina, *Introduction to classical and modern test theory*. ERIC, 1986.

[146] J. Campbell, A. Narayanan, B. Burford, and M. Greco, "Validation of a multi-source feedback tool for use in general practice," vol. 9879, no. December, 2010.

Appendix A. Raters Aggregate Counts

Doctors with more than 200 patients

PSID	Avg. Overall Scale	No. of Patients	PSID	Avg. Overall Scale	No. of Patients
50080	70.93	200	82407	61.68	210
53441	76.26	200	76485	76.35	211
60895	75.27	201	51632	77.83	211
50005	75.99	201	86575	78.59	211
59503	79.73	201	73386	78.60	211
82502	72.56	202	59912	68.37	212
78754	75.70	202	91958	76.77	212
49845	77.48	202	59633	72.27	213
68173	78.21	202	47534	64.68	214
100501	78.49	202	84612	69.30	214
80940	70.58	203	83196	74.22	214
84388	72.51	203	95007	77.38	214
81699	73.32	203	95301	79.41	214
67103	75.55	203	60521	80.87	214
73947	72.35	204	59640	68.13	215
75385	73.05	204	60565	72.84	215
99836	75.99	204	83183	75.47	215
47282	77.64	204	60643	78.09	215
53033	77.85	204	51556	79.36	215
91136	78.73	204	53337	79.84	215
90765	63.51	205	44464	80.59	215
59398	66.27	205	85234	72.99	216
80868	69.70	205	46922	77.75	217
84881	71.18	205	81267	81.22	217
72197	71.96	205	70008	66.27	218
89440	72.42	205	63598	71.17	218
48760	73.23	205	102905	73.26	218
61319	76.35	205	61149	78.29	218
47283	77.41	205	46448	63.20	219
87941	71.86	206	83093	64.99	219
52465	74.72	206	60107	70.88	220
80076	76.15	206	87274	73.00	220
79476	73.66	207	63366	73.82	220
48761	73.95	207	85943	74.22	220
80095	73.87	208	81315	80.88	220
60585	74.97	208	86599	72.32	221
103573	75.26	208	86501	74.51	221
64180	77.35	208	93192	74.94	221
64306	62.87	209	48359	65.96	222

PSID	Avg. Overall Scale	No. of Patients	PSID	Avg. Overall Scale	No. of Patients
50433	68.91	223	96373	69.50	250
93880	73.91	223	53510	72.87	250
89986	64.98	224	78503	67.34	252
101943	72.37	224	91128	71.87	255
89360	73.50	224	105220	79.00	255
50007	77.75	224	73412	66.60	256
86486	80.03	224	52499	76.11	257
96632	88.26	224	51191	71.88	258
74045	75.36	225	51266	69.68	260
52247	78.92	225	70140	65.32	264
50304	78.33	226	83481	75.52	264
47445	80.38	226	79867	77.27	265
51401	78.74	227	69224	78.57	265
62237	69.84	228	62199	69.11	267
71493	73.01	228	79378	75.83	268
86633	69.23	229	79318	76.91	269
50628	74.78	229	50034	78.29	269
107174	74.97	229	80724	66.07	270
65970	76.45	230	84585	69.74	271
69835	76.90	233	72807	75.59	276
66034	64.91	234	84581	63.88	278
76085	68.16	234	81016	65.38	284
48471	76.71	234	45269	68.06	285
71609	71.90	236	75148	67.61	288
81005	80.43	236	77661	74.23	302
78274	75.19	238	96375	74.51	308
53444	75.89	238	101497	68.88	309
67013	76.55	238	62238	77.40	309
89682	66.29	239	66972	71.06	310
70419	75.82	239	86627	65.55	312
47517	77.55	239	71791	72.36	321
82233	75.25	240	90758	69.08	346
50319	72.55	241	69119	76.59	353
50230	66.00	243	82321	70.31	396
76804	75.96	243	86756	79.14	400
69354	81.37	243	83037	75.18	425
100710	81.90	245	51374	75.52	425
92374	76.25	247	68342	78.34	248
86489	74.42	248			

Appendix B. Doctors Mean Scores after Controlling "Usual Doctor" Factor

PSID	No of patients	Old Mean	New Mean	Mean Difference
44464	215	80.59	79.48	-1.11
45269	285	68.06	67.72	-0.34
46448	219	63.20	63.90	0.70
46922	217	77.75	76.63	-1.12
47282	204	77.64	77.40	-0.24
47283	205	77.41	76.43	-0.98
47445	226	80.38	79.57	-0.81
47517	239	77.55	76.09	-1.46
47534	214	64.68	63.26	-1.43
48359	222	65.96	65.17	-0.79
48471	234	76.71	76.34	-0.37
48760	205	73.23	72.51	-0.71
48761	207	73.95	72.65	-1.29
49845	202	77.48	75.98	-1.50
50005	201	75.99	75.27	-0.72
50007	224	77.75	77.61	-0.14
50034	269	78.29	78.03	-0.25
50080	200	70.93	69.95	-0.98
50230	243	66.00	65.92	-0.07
50304	226	78.33	77.05	-1.29
50319	241	72.55	71.47	-1.08
50433	223	68.91	68.21	-0.70
50628	229	74.78	73.12	-1.66
51191	258	71.88	70.45	-1.43
51266	260	69.68	69.34	-0.33
51374	425	75.52	75.25	-0.27
51401	227	78.74	78.36	-0.37
51556	215	79.36	78.78	-0.58
51632	211	77.83	76.90	-0.93
52247	225	78.92	77.67	-1.25
52465	206	74.72	74.56	-0.16
52499	257	76.11	76.09	-0.02
53033	204	77.85	78.29	0.44
53337	215	79.84	79.41	-0.43
53441	200	76.26	75.10	-1.15
53444	238	75.89	75.70	-0.19
53510	250	72.87	72.73	-0.14
59398	205	66.27	65.92	-0.35
59503	201	79.73	78.89	-0.84

PSID	No of patients	Old Mean	New Mean	Mean Difference
59633	213	72.27	72.37	0.10
59640	215	68.13	67.41	-0.72
59912	212	68.37	67.52	-0.85
60107	220	70.88	70.29	-0.60
60521	214	80.87	79.60	-1.28
60565	215	72.84	72.66	-0.18
60585	208	74.97	73.95	-1.02
60643	215	78.09	76.11	-1.98
60895	201	75.27	73.85	-1.42
61149	218	78.29	77.31	-0.99
61319	205	76.35	76.12	-0.23
62199	267	69.11	68.88	-0.24
62237	228	69.84	69.77	-0.07
63366	220	73.82	72.33	-1.49
63598	218	71.17	69.65	-1.52
64180	208	77.35	75.45	-1.90
64306	209	62.87	63.15	0.28
65970	230	76.45	77.10	0.65
66034	234	64.91	64.73	-0.17
66972	310	71.06	70.88	-0.19
67013	238	76.55	76.00	-0.55
67103	203	75.55	74.88	-0.67
68173	202	78.21	77.18	-1.02
68342	248	78.34	76.54	-1.80
69224	265	78.57	78.03	-0.54
69354	243	81.37	79.90	-1.47
69835	233	76.90	76.42	-0.49
70008	218	66.27	66.14	-0.12
70140	264	65.32	65.32	0.00
70419	239	75.82	75.40	-0.41
71493	228	73.01	72.38	-0.64
71609	236	71.90	71.46	-0.44
71791	321	72.36	72.35	-0.01
72197	205	71.96	71.64	-0.31
72807	276	75.59	74.75	-0.85
73386	211	78.60	77.92	-0.68
73412	256	66.60	66.68	0.08
73947	204	72.35	71.24	-1.11
74045	225	75.36	75.33	-0.03
75148	288	67.61	67.54	-0.06
75385	204	73.05	72.76	-0.29
76085	234	68.16	66.50	-1.66
76485	211	76.35	75.58	-0.77
76804	243	75.96	75.88	-0.09
77661	302	74.23	72.48	-1.75
78274	238	75.19	75.19	0.00

PSID	No of patients	Old Mean	New Mean	Mean Difference
78503	252	67.34	67.62	0.28
78754	202	75.70	75.32	-0.38
79318	269	76.91	76.70	-0.21
79378	268	75.83	76.27	0.45
79476	207	73.66	71.97	-1.69
79867	265	77.27	76.52	-0.74
80076	206	76.15	75.47	-0.68
80095	208	73.87	72.98	-0.89
80724	270	66.07	66.39	0.32
80868	205	69.70	70.61	0.91
80940	203	70.58	71.17	0.59
81005	236	80.43	79.49	-0.94
81016	284	65.38	65.35	-0.03
81267	217	81.22	81.13	-0.09
81315	220	80.88	79.63	-1.25
81699	203	73.32	73.12	-0.19
82233	240	75.25	71.89	-3.36
82321	396	70.31	70.60	0.29
82407	210	61.68	61.43	-0.25
82502	202	72.56	72.01	-0.55
83093	219	64.99	65.13	0.14
83183	215	75.47	74.50	-0.96
83196	214	74.22	73.91	-0.31
83481	264	75.52	75.35	-0.16
84388	203	72.51	71.81	-0.70
84581	278	63.88	63.49	-0.39
84585	271	69.74	69.45	-0.29
84612	214	69.30	68.71	-0.59
84881	205	71.18	70.29	-0.89
85234	216	72.99	72.14	-0.85
85943	220	74.22	73.08	-1.14
86486	224	80.03	78.80	-1.23
86489	248	74.42	74.14	-0.27
86501	221	74.51	74.70	0.19
86575	211	78.59	77.22	-1.37
86599	221	72.32	70.88	-1.44
86627	312	65.55	65.46	-0.09
86633	229	69.23	68.09	-1.14
86756	400	79.14	78.30	-0.84
87274	220	73.00	72.93	-0.06
87941	206	71.86	71.33	-0.53
89360	224	73.50	71.89	-1.60
89440	205	72.42	72.22	-0.21
89682	239	66.29	65.34	-0.95
89986	224	64.98	64.15	-0.83
90758	346	69.08	68.46	-0.63

PSID	No of patients	Old Mean	New Mean	Mean Difference
90765	205	63.51	63.76	0.25
91128	255	71.87	71.67	-0.20
91136	204	78.73	77.11	-1.62
91958	212	76.77	76.67	-0.10
92374	247	76.25	75.75	-0.50
93192	221	74.94	74.66	-0.27
93880	223	73.91	73.62	-0.30
95007	214	77.38	75.95	-1.43
95301	214	79.41	78.80	-0.61
96373	250	69.50	69.48	-0.02
96632	224	88.26	88.58	0.32
99836	204	75.99	75.96	-0.03
100501	202	78.49	77.60	-0.89
100710	245	81.90	79.01	-2.90
101497	309	68.88	69.04	0.16
101943	224	72.37	72.09	-0.28
102905	218	73.26	72.75	-0.51
103573	208	75.26	74.50	-0.76
105220	255	79.00	78.76	-0.24
107174	229	74.97	74.04	-0.93

Appendix C. Doctors Mean Scores after Controlling "Age Group" Factor

PSID	No of patients	Old Mean	New Mean	Mean Difference
44464	215	80.589	80.873	0.284
45269	285	68.062	67.392	-0.670
46448	219	63.200	62.886	-0.314
46922	217	77.750	77.054	-0.696
47282	204	77.636	76.980	-0.656
47283	205	77.409	78.320	0.911
47445	226	80.380	79.954	-0.427
47517	239	77.552	76.856	-0.695
47534	214	64.683	65.237	0.554
48359	222	65.959	65.556	-0.404
48471	234	76.711	76.221	-0.490
48760	205	73.229	73.885	0.656
48761	207	73.945	74.546	0.601
49845	202	77.477	78.264	0.786
50005	201	75.990	75.968	-0.023
50007	224	77.748	77.326	-0.422
50034	269	78.287	77.671	-0.616
50080	200	70.930	71.788	0.859
50230	243	65.996	67.498	1.502
50304	226	78.332	74.956	-3.376
50319	241	72.550	72.327	-0.223
50433	223	68.912	68.644	-0.268
50628	229	74.776	73.256	-1.520
51191	258	71.881	70.404	-1.476
51266	260	69.678	70.107	0.429
51374	425	75.522	76.085	0.563
51401	227	78.737	78.896	0.159
51556	215	79.359	79.133	-0.227
51632	211	77.830	77.166	-0.665
52247	225	78.917	78.600	-0.317
52465	206	74.725	75.118	0.394
52499	257	76.112	76.020	-0.092
53033	204	77.847	78.941	1.094
53337	215	79.842	78.521	-1.320
53441	200	76.256	75.851	-0.404
53444	238	75.892	76.859	0.968
53510	250	72.868	73.549	0.681
59398	205	66.269	65.770	-0.499

PSID	No of patients	Old Mean	New Mean	Mean Difference
59503	201	79.735	78.634	-1.101
59633	213	72.273	73.219	0.946
59640	215	68.127	67.003	-1.125
59912	212	68.368	67.987	-0.381
60107	220	70.882	71.276	0.394
60521	214	80.872	79.531	-1.342
60565	215	72.837	72.885	0.048
60585	208	74.968	73.957	-1.011
60643	215	78.091	77.328	-0.763
60895	201	75.272	76.496	1.224
61149	218	78.294	77.712	-0.582
61319	205	76.347	73.402	-2.945
62199	267	69.114	70.017	0.903
62237	228	69.841	69.848	0.007
63366	220	73.822	72.900	-0.922
63598	218	71.172	70.250	-0.923
64180	208	77.350	72.752	-4.598
64306	209	62.867	63.364	0.497
65970	230	76.448	76.427	-0.020
66034	234	64.907	66.369	1.462
66972	310	71.063	70.640	-0.423
67013	238	76.548	77.193	0.645
67103	203	75.548	75.941	0.393
68173	202	78.207	77.619	-0.588
68342	248	78.342	76.888	-1.454
69224	265	78.569	79.118	0.549
69354	243	81.366	81.031	-0.334
69835	233	76.904	76.624	-0.280
70008	218	66.266	66.307	0.041
70140	264	65.323	63.480	-1.843
70419	239	75.816	76.611	0.795
71493	228	73.015	71.755	-1.260
71609	236	71.902	72.059	0.157
71791	321	72.357	72.756	0.399
72197	205	71.957	75.245	3.289
72807	276	75.593	75.672	0.079
73386	211	78.603	78.565	-0.038
73412	256	66.600	67.764	1.163
73947	204	72.349	71.318	-1.031
74045	225	75.358	75.515	0.157
75148	288	67.605	68.545	0.940
75385	204	73.046	73.272	0.225
76085	234	68.158	67.723	-0.435
76485	211	76.345	75.326	-1.020
76804	243	75.964	77.186	1.222

PSID	No of patients	Old Mean	New Mean	Mean Difference
77661	302	74.231	73.103	-1.128
78274	238	75.191	74.934	-0.258
78503	252	67.343	68.524	1.181
78754	202	75.702	74.947	-0.755
79318	269	76.913	77.029	0.116
79378	268	75.826	76.246	0.419
79476	207	73.663	72.971	-0.692
79867	265	77.266	78.048	0.782
80076	206	76.145	77.602	1.457
80095	208	73.868	73.991	0.124
80724	270	66.069	67.173	1.105
80868	205	69.702	69.824	0.122
80940	203	70.582	71.457	0.875
81005	236	80.433	79.904	-0.529
81016	284	65.378	64.340	-1.038
81267	217	81.222	80.924	-0.298
81315	220	80.879	80.888	0.010
81699	203	73.315	73.167	-0.148
82233	240	75.250	73.141	-2.109
82321	396	70.314	69.703	-0.611
82407	210	61.675	60.792	-0.883
82502	202	72.563	74.220	1.656
83093	219	64.986	66.499	1.513
83183	215	75.466	75.300	-0.166
83196	214	74.216	73.894	-0.322
83481	264	75.519	76.422	0.903
84388	203	72.512	72.010	-0.502
84581	278	63.880	63.625	-0.254
84585	271	69.742	69.534	-0.207
84612	214	69.301	70.039	0.738
84881	205	71.183	70.686	-0.497
85234	216	72.987	73.664	0.677
85943	220	74.222	72.805	-1.418
86486	224	80.030	78.398	-1.632
86489	248	74.418	74.372	-0.046
86501	221	74.513	75.509	0.996
86575	211	78.589	76.927	-1.662
86599	221	72.318	74.518	2.200
86627	312	65.551	63.945	-1.606
86633	229	69.229	68.111	-1.117
86756	400	79.139	79.653	0.514
87274	220	72.997	71.822	-1.175
87941	206	71.863	71.739	-0.124
89360	224	73.495	73.800	0.305
89440	205	72.423	71.825	-0.597

PSID	No of patients	Old Mean	New Mean	Mean Difference
89682	239	66.292	67.617	1.326
89986	224	64.977	64.741	-0.236
90758	346	69.084	68.540	-0.544
90765	205	63.509	64.067	0.558
91128	255	71.866	70.236	-1.630
91136	204	78.729	79.360	0.631
91958	212	76.771	76.972	0.201
92374	247	76.248	76.251	0.003
93192	221	74.935	75.426	0.490
93880	223	73.915	75.965	2.050
95007	214	77.380	78.086	0.706
95301	214	79.408	79.123	-0.285
96373	250	69.502	70.668	1.165
96632	224	88.264	88.322	0.058
99836	204	75.991	75.366	-0.625
100501	202	78.489	77.501	-0.988
100710	245	81.902	83.057	1.155
101497	309	68.879	69.561	0.681
101943	224	72.371	71.149	-1.222
102905	218	73.255	73.367	0.111
103573	208	75.260	74.829	-0.431
105220	255	79.001	78.804	-0.197
107174	229	74.970	75.735	0.765

Appendix D. Doctors Mean Scores after Controlling "Gender" Factor

PSID	No of Patients	Old Mean	New Mean	Mean Difference
44464	215	80.589147	81.051578	0.4624305
45269	285	68.062378	68.229696	0.1673181
46448	219	63.199729	63.955304	0.7555743
46922	217	77.750469	77.817822	0.0673529
47282	204	77.636166	77.877163	0.2409974
47283	205	77.409214	77.257059	-0.1521548
47445	226	80.380203	80.340267	-0.0399365
47517	239	77.551526	78.111311	0.5597847
47534	214	64.683281	65.231993	0.5487112
48359	222	65.959293	66.713111	0.7538185
48471	234	76.710984	77.097741	0.3867563
48760	205	73.228546	73.29943	0.070884
48761	207	73.94525	73.369007	-0.5762422
49845	202	77.477081	77.559526	0.0824449
50005	201	75.990418	75.968133	-0.0222849
50007	224	77.748016	78.285222	0.5372063
50034	269	78.287209	78.239244	-0.0479647
50080	200	70.92963	71.327837	0.3982074
50230	243	65.996037	66.286556	0.2905184
50304	226	78.331695	78.565215	0.2335203
50319	241	72.549562	72.887578	0.3380158
50433	223	68.912141	69.365904	0.4537629
50628	229	74.775999	74.856348	0.0803494
51191	258	71.880563	71.899681	0.0191187
51266	260	69.678063	70.424719	0.7466558
51374	425	75.52244	75.985704	0.4632636
51401	227	78.737151	78.868178	0.1310265
51556	215	79.359173	79.549163	0.1899898
51632	211	77.830437	78.438548	0.6081111
52247	225	78.916872	79.438526	0.5216535
52465	206	74.724919	74.35263	-0.3722895
52499	257	76.111832	76.272378	0.1605461
53033	204	77.846768	77.861919	0.0151502
53337	215	79.841516	80.622296	0.7807804

PSID	No of Patients	Old Mean	New Mean	Mean Difference
53441	200	76.255556	76.671356	0.4158
53444	238	75.89169	76.159289	0.2675989
53510	250	72.868148	72.474452	-0.3936963
59398	205	66.269196	67.008659	0.7394632
59503	201	79.73466	80.003674	0.269014
59633	213	72.272648	72.194119	-0.0785297
59640	215	68.127476	69.144089	1.0166126
59912	212	68.368274	67.836059	-0.5322147
60107	220	70.882155	70.931652	0.049497
60521	214	80.872274	81.254978	0.3827036
60565	215	72.837209	72.938459	0.10125
60585	208	74.967949	74.848444	-0.1195043
60643	215	78.091301	78.709556	0.618255
60895	201	75.271789	75.972896	0.7011071
61149	218	78.294258	79.029822	0.7355647
61319	205	76.346883	75.925244	-0.421639
62199	267	69.113608	69.420904	0.3072957
62237	228	69.840806	70.06977	0.2289647
62238	220	77.401414	77.34777	-0.053644
63366	218	73.821549	74.210296	0.3887475
63598	208	71.172273	71.458622	0.286349
64180	209	77.350427	77.602933	0.252506
64306	230	62.867269	63.259793	0.3925234
65970	234	76.447665	76.571289	0.1236238
66034	310	64.906616	64.914111	0.0074951
66972	238	71.063321	71.229496	0.1661749
67013	203	76.548397	76.117067	-0.4313305
67103	202	75.548258	75.48463	-0.063628
68173	248	78.206821	79.30897	1.1021497
68342	265	78.342294	78.60357	0.2612765
69119	243	76.592173	76.841504	0.2493308
69224	233	78.568833	78.959119	0.3902855
69354	218	81.365645	81.881422	0.5157767
69835	264	76.903513	77.045644	0.1421315
70008	239	66.265715	66.852919	0.5872033
70140	228	65.322671	65.326074	0.0034029
70419	236	75.8159	76.050874	0.2349745
71493	321	73.014945	72.987467	-0.0274781
71609	205	71.902072	71.898326	-0.0037456
71791	276	72.357217	72.278874	-0.078343

PSID	No of Patients	Old Mean	New Mean	Mean Difference
72197	211	71.95664	72.290133	0.3334938
72807	256	75.593129	75.858933	0.265804
73386	204	78.602773	78.59323	-0.0095438
73412	225	66.600116	68.026993	1.4268769
73947	288	72.34931	72.268919	-0.0803916
74045	204	75.358025	75.175452	-0.1825728
75148	234	67.605453	67.834052	0.2285992
75385	211	73.046478	73.496452	0.449974
76085	243	68.157645	68.646563	0.4889181
76485	302	76.345445	76.289985	-0.0554598
76804	238	75.96403	75.690644	-0.2733854
77661	252	74.231052	74.026963	-0.2040893
78274	202	75.19141	75.217993	0.0265827
78503	269	67.34274	67.447563	0.1048234
78754	268	75.702237	75.500644	-0.2015924
79318	207	76.913121	77.379385	0.4662639
79378	265	75.826423	75.620496	-0.2059271
79476	206	73.662551	73.681133	0.0185819
79867	208	77.266247	77.068896	-0.1973511
80076	270	76.145271	76.58697	0.4416989
80095	205	73.867521	73.442496	-0.4250251
80724	203	66.068587	66.466281	0.3976944
80868	236	69.701897	69.974074	0.2721771
80940	284	70.582011	70.783089	0.2010783
81005	217	80.433145	80.521859	0.0887143
81016	220	65.378195	66.086096	0.7079012
81267	203	81.222052	81.344978	0.1229262
81315	240	80.878788	81.02083	0.1420418
81699	396	73.315088	73.976319	0.66123
82233	210	75.25	75.423215	0.1732148
82321	202	70.314254	70.398807	0.0845538
82407	219	61.675485	61.432896	-0.2425887
82502	215	72.563256	72.864104	0.3008474
83037	214	75.180828	75.480007	0.2991795
83093	264	64.985625	66.012178	1.0265529
83183	203	75.465978	75.631193	0.165215
83196	278	74.215992	74.860333	0.6443416
83481	271	75.51908	75.659778	0.1406981
84388	214	72.512315	72.258215	-0.2541005
84581	205	63.879563	63.850844	-0.0287186

PSID	No of Patients	Old Mean	New Mean	Mean Difference
84585	216	69.741697	69.690111	-0.0515863
84612	220	69.300796	69.484044	0.1832483
84881	224	71.183379	72.095578	0.9121993
85234	248	72.986968	73.542993	0.5560241
85943	221	74.222222	74.378874	0.1566519
86486	211	80.029762	80.309319	0.2795566
86489	221	74.417563	74.5344	0.1168373
86501	312	74.513156	75.01677	0.5036147
86575	229	78.588731	78.338437	-0.2502939
86599	400	72.317748	72.630111	0.3123635
86627	220	65.550807	65.789393	0.2385854
86633	206	69.22853	69.405837	0.1773072
86756	224	79.138889	79.203637	0.0647481
87274	205	72.996633	73.417859	0.4212263
87941	239	71.862639	71.643044	-0.2195949
89360	224	73.49537	73.788081	0.2927111
89440	346	72.422764	72.256615	-0.1661494
89682	205	66.291647	66.499881	0.2082342
89986	255	64.976852	66.113778	1.1369259
90758	204	69.083708	69.222037	0.1383291
90765	212	63.508582	63.806444	0.2978627
91128	247	71.866376	71.857215	-0.0091614
91136	221	78.729121	78.838985	0.1098639
91958	223	76.771488	76.64163	-0.1298588
92374	214	76.248313	76.640785	0.3924721
93192	214	74.935478	75.005089	0.0696104
93880	250	73.914632	74.021933	0.1073012
95007	224	77.379716	77.338941	-0.0407754
95301	204	79.4081	79.350593	-0.0575071
96373	202	69.502222	70.120933	0.6187111
96375	245	74.511785	75.27783	0.7660451
96632	309	88.263889	88.408822	0.1449333
99836	224	75.991285	76.095644	0.104359
100501	218	78.489182	78.460326	-0.0288563
100710	208	81.901738	81.906185	0.0044467
101497	255	68.8793	69.28023	0.4009296
101943	229	72.371032	72.233837	-0.1371947
102905		73.255182	73.104111	-0.1510707
103573		75.259972	75.38317	0.1231989
105220		79.000726	79.087237	0.0865108

PSID	No of Patients	Old Mean	New Mean	Mean Difference
107174		74.970079	75.020681	0.0506022

Appendix E. Doctors Mean Scores after Controlling "Years Attending" Factor

PSID	No of patients	Old Mean	New Mean	Mean Difference
44464	215	80.59	79.56	-1.03
45269	285	68.06	68.48	0.42
46448	219	63.20	63.52	0.32
46922	217	77.75	77.69	-0.06
47282	204	77.64	77.30	-0.33
47283	205	77.41	77.47	0.06
47445	226	80.38	79.94	-0.44
47517	239	77.55	77.18	-0.37
47534	214	64.68	63.98	-0.70
48359	222	65.96	64.98	-0.98
48471	234	76.71	75.77	-0.94
48760	205	73.23	72.12	-1.10
48761	207	73.95	74.98	1.04
49845	202	77.48	76.58	-0.89
50005	201	75.99	75.58	-0.41
50007	224	77.75	78.13	0.38
50034	269	78.29	77.21	-1.07
50080	200	70.93	74.07	3.14
50230	243	66.00	66.18	0.19
50304	226	78.33	77.65	-0.68
50319	241	72.55	72.26	-0.28
50433	223	68.91	69.38	0.46
50628	229	74.78	73.85	-0.93
51191	258	71.88	73.00	1.12
51266	260	69.68	70.64	0.96
51374	425	75.52	74.79	-0.73
51401	227	78.74	80.98	2.24
51556	215	79.36	77.74	-1.62
51632	211	77.83	78.95	1.12
52247	225	78.92	77.88	-1.04
52465	206	74.72	72.58	-2.15
52499	257	76.11	74.22	-1.89
53033	204	77.85	77.94	0.10
53337	215	79.84	79.86	0.02
53441	200	76.26	76.54	0.28

PSID	No of patients	Old Mean	New Mean	Mean Difference
53444	238	75.89	75.89	-0.01
53510	250	72.87	72.80	-0.07
59398	205	66.27	66.21	-0.06
59503	201	79.73	79.46	-0.28
59633	213	72.27	72.40	0.13
59640	215	68.13	67.10	-1.03
59912	212	68.37	68.51	0.15
60107	220	70.88	70.62	-0.26
60521	214	80.87	79.42	-1.45
60565	215	72.84	73.13	0.29
60585	208	74.97	73.30	-1.67
60643	215	78.09	78.22	0.13
60895	201	75.27	75.07	-0.20
61149	218	78.29	78.19	-0.10
61319	205	76.35	76.78	0.43
62199	267	69.11	69.29	0.17
62237	228	69.84	71.04	1.20
63366	220	73.82	71.99	-1.83
63598	218	71.17	71.40	0.23
64180	208	77.35	74.14	-3.21
64306	209	62.87	62.25	-0.62
65970	230	76.45	76.73	0.28
66034	234	64.91	65.00	0.10
66972	310	71.06	71.33	0.27
67013	238	76.55	78.42	1.87
67103	203	75.55	75.61	0.06
68173	202	78.21	78.51	0.30
68342	248	78.34	77.99	-0.35
69224	265	78.57	78.62	0.05
69354	243	81.37	81.05	-0.32
69835	233	76.90	78.78	1.88
70008	218	66.27	66.35	0.09
70140	264	65.32	65.18	-0.15
70419	239	75.82	75.04	-0.77
71493	228	73.01	73.51	0.49
71609	236	71.90	71.83	-0.07
71791	321	72.36	72.65	0.29
72197	205	71.96	71.83	-0.13
72807	276	75.59	75.50	-0.09
73386	211	78.60	77.97	-0.63
73412	256	66.60	65.56	-1.04

PSID	No of patients	Old Mean	New Mean	Mean Difference
73947	204	72.35	72.88	0.53
74045	225	75.36	74.91	-0.45
75148	288	67.61	67.53	-0.07
75385	204	73.05	72.43	-0.62
76085	234	68.16	66.14	-2.02
76485	211	76.35	75.30	-1.05
76804	243	75.96	74.60	-1.36
77661	302	74.23	73.58	-0.65
78274	238	75.19	74.94	-0.25
78503	252	67.34	67.79	0.45
78754	202	75.70	75.66	-0.05
79318	269	76.91	76.89	-0.02
79378	268	75.83	75.82	-0.01
79476	207	73.66	72.87	-0.80
79867	265	77.27	77.79	0.53
80076	206	76.15	77.07	0.93
80095	208	73.87	74.25	0.38
80724	270	66.07	65.64	-0.43
80868	205	69.70	69.75	0.04
80940	203	70.58	71.25	0.66
81005	236	80.43	79.21	-1.22
81016	284	65.38	67.68	2.30
81267	217	81.22	81.89	0.66
81315	220	80.88	80.91	0.03
81699	203	73.32	72.27	-1.05
82233	240	75.25	75.44	0.19
82321	396	70.31	70.09	-0.22
82407	210	61.68	61.60	-0.07
82502	202	72.56	71.32	-1.24
83093	219	64.99	65.07	0.09
83183	215	75.47	76.53	1.07
83196	214	74.22	74.93	0.72
83481	264	75.52	76.37	0.85
84388	203	72.51	71.86	-0.65
84581	278	63.88	63.43	-0.45
84585	271	69.74	69.84	0.10
84612	214	69.30	67.61	-1.69
84881	205	71.18	71.27	0.08
85234	216	72.99	72.85	-0.14
85943	220	74.22	73.10	-1.12
86486	224	80.03	79.74	-0.29

PSID	No of patients	Old Mean	New Mean	Mean Difference
86489	248	74.42	75.22	0.80
86501	221	74.51	74.48	-0.03
86575	211	78.59	78.17	-0.42
86599	221	72.32	72.16	-0.16
86627	312	65.55	65.98	0.43
86633	229	69.23	67.78	-1.45
86756	400	79.14	79.07	-0.07
87274	220	73.00	71.65	-1.35
87941	206	71.86	71.56	-0.30
89360	224	73.50	74.27	0.77
89440	205	72.42	71.92	-0.50
89682	239	66.29	64.84	-1.46
89986	224	64.98	64.48	-0.49
90758	346	69.08	68.54	-0.54
90765	205	63.51	63.54	0.03
91128	255	71.87	70.84	-1.03
91136	204	78.73	78.18	-0.55
91958	212	76.77	76.08	-0.69
92374	247	76.25	77.11	0.86
93192	221	74.94	75.55	0.61
93880	223	73.91	74.18	0.27
95007	214	77.38	78.98	1.60
95301	214	79.41	79.29	-0.12
96373	250	69.50	68.75	-0.75
96632	224	88.26	88.52	0.25
99836	204	75.99	76.04	0.04
100501	202	78.49	78.44	-0.05
100710	245	81.90	82.54	0.64
101497	309	68.88	69.60	0.72
101943	224	72.37	70.84	-1.53
102905	218	73.26	73.49	0.23
103573	208	75.26	75.46	0.20
105220	255	79.00	78.64	-0.36
107174	229	74.97	74.83	-0.14