Monitoring and Improving User Compliance and Data Quality

For Long and Repetitive Self-Reporting MHealth Surveys

by

Pooja Rallabhandi

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved March 2017 by the
Graduate Supervisory Committee:

Kevin Gary, Chair
Ashraf Gaffar
Ashish Amresh
Srividya Bansal

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

For the past decade, mobile health applications are seeing greater acceptance due to their potential to remotely monitor and increase patient engagement, particularly for chronic disease. Sickle Cell Disease is an inherited chronic disorder of red blood cells requiring careful pain management. A significant number of mHealth applications have been developed in the market to help clinicians collect and monitor information of SCD patients. Surveys are the most common way to self-report patient conditions. These are non-engaging and suffer from poor compliance. The quality of data gathered from survey instruments while using technology can be questioned as patients may be motivated to complete a task but not motivated to do it well. A compromise in quality and quantity of the collected patient data hinders the clinicians' effort to be able to monitor patient's health on a regular basis and derive effective treatment measures. This research study has two goals. The first is to monitor user compliance and data quality in mHealth apps with long and repetitive surveys delivered. The second is to identify possible motivational interventions to help improve compliance and data quality. As a form of intervention, will introduce intrinsic and extrinsic motivational factors within the application and test it on a small target population. I will validate the impact of these motivational factors by performing a comparative analysis on the test results to determine improvements in user performance. This study is relevant, as it will help analyze user behavior in long and repetitive self-reporting tasks and derive measures to improve user performance. The results will assist software engineers working with doctors in designing and developing improved self-reporting mHealth applications for collecting better quality data and enhance user compliance.

To my caring and loving husband.

ACKNOWLEDGMENTS

I would like to specially acknowledge Dr. Kevin Gary for the tremendous amount of guidance and support provided by him in the past two years. His guidance helped me explore various interesting research methods as part of this thesis. I am grateful for getting an opportunity to work with him.

In the past two years, as part of this research I got an opportunity to work with the clinicians, Dr. Zenaide Quezado, M.D., and Dr. Kevin Cleary, PhD, from the Children's National Health System in Washington D.C. They have provided tremendous support in helping me validate my research questions in clinical trials at the hospital. With their help, this entire research process has been an interesting journey for me.

I would like to acknowledge my team members, Christian, Chintan, Deepak and Mohit from the research lab, which helped me in every single phase of this research. This thesis would not have been possible without their help and support. It has been an amazing experience to work with them.

My husband has been my strength in the past two years who encouraged me to pursue research as part of my graduate studies. I am grateful to him for being there for me in all my ups and downs. I would also like to thank my parents for giving me this opportunity to pursue my career in the field of my interest. All this would not have been possible without their love and support.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Mobile health (mHealth) is a relatively new term for providing healthcare solutions via mobile devices, and for collecting and delivering patient related data to the clinicians and researchers on real time [15]. MHealth is gaining popularity due to its advantages of being available anytime and anywhere for the public. Mobile devices have the advantage of being portable, constantly connected via the internet, giving an individual a sense of privacy, and being able to act like sensors for monitoring and collecting an individual's daily activity data. In 2009, World Health Organization (WHO) [24] conducted a survey on the existence of mHealth and its popularity all over the world. The survey report states that the use of mobile and wireless technologies to support the achievement of health objectives (mHealth) has the potential to transform the face of health service delivery across the globe. The report mentions the existence of mHealth and its initiatives in a total of 114 countries including many low and middle income countries all over the world. There are a significant number of mHealth applications currently available in the market for use by the public as well by the various health care providers [7, 8, 23, 34, 40, and 42].

Sickle Cell Disease (SCD) is inherited chronic pain disease largely affecting the African-American population in the U.S. Patients suffering from SCD usually suffer chronic pain in their bones, joints, back, abdomen and the chest. Clinicians treating these patients often rely upon self-reported pain data collected from the patients on regular intervals to provide effective treatments as per each individual patient’s needs. mHealth applications make a significant contribution by providing a platform for collecting self-reported data from the patients.

These applications help in collect the self-reported pain data in fixed intervals which is later analyzed by the clinicians. A thorough analysis of these responses helps the clinicians to address each patient with better treatment procedures. For the clinicians to be able to analyze various patient conditions the quantity and quality of data collected from these self-reported activities plays an important role. The greater the *quantity* of responses collected, the larger the dataset available for analysis. The more the user is engaged in the app, the chances of his participation in the survey activity increases thus increasing the quantity of responses collected. The better the *quality* of responses, the more accurate would be the results of the analysis performed.

Many mHealth applications available in the market today are not able to ensure better quality and quantity of self-reported data collection. There are various factors impacting this behavior. One important factor could be the questionnaire length which is delivered in the survey. The length of a specific task might have adverse effects on the user compliance and the quality of data reported in the task [10]. No research has been found yet which studies the impact of survey length and its repetitiveness on the user compliance and quality of data collected in a self-reported task.

Psychologists suggest the framework of Self Determination Theory (SDT) [61] for improving user participation and data quality in a self-reporting task [33, 60, 61 and 62]. SDT mainly consists of two motivational factors; *intrinsic* and *extrinsic*. Intrinsic motivations are used to make the user aware of the importance of the task and the positive impacts of completing it on time. Extrinsic motivations are used to induce external factors to drive the user towards completion of the task. These factors could be external rewards, such as badges or any monetary benefits provided at the end of the task. Researchers have explored the areas of gamification as a form of external motivation for improving user engagement [16, 31, 68, 70, 71, and 73].

This research is focused on monitoring the impact of long and repetitive nature of surveys on user compliance and data quality; delivered via mHealth applications. We are using an mHealth application designed to monitor SCD patients using Patient-Reported Outcomes Measurement Information System (PROMIS) self-reporting instruments in multiple trials spanning multiple weeks of engagement. These surveys comprise of various question sets from PROMIS short forms, medication adherence as well as adaptive questions and are designed to be lengthy in nature.

My research is based on the following hypotheses:

1. The length and repetitive nature of mHealth surveys impact the user compliance and data quality of the self-reported data collected.
2. The introduction of gamification as an extrinsic motivator can improve user compliance and data quality in long and repetitive self-reporting mHealth tasks.

To validate the first hypothesis, I conducted a pilot study. The details regarding this study are explained in Chapter 6. The results of this pilot study show that the length and repetitiveness of the surveys do impact the user compliance and data quality observed. Further, a second usability study was conducted on actual SCD patients to check whether the existence of issues in user compliance and data quality observed in the mHealth context. The results from the pilot and ongoing clinical trial help validate my first hypothesis and prove that the length and repetitive nature of surveys do impact compliance and quality. To validate the second hypothesis, I further conducted a study by introducing badges and games as rewards for survey completion in the SCD mHealth application. This study helps us validate the second hypothesis and suggests that introducing gamification factors such as badges and game as an external reward helps improve the user compliance rates and data quality for a long and repetitive mHealth survey used for collecting self-reported data.

The rest of this thesis work is presented as follows. Chapter 2 describes literature review. Chapter 3 helps explain the context for this research work by discussing various methods used to assess the user compliance and data quality performance in self-reported tasks. Chapter 4 describes the implementation steps for introducing gamification factors as extrinsic motivational rewards in the mHealth application. Chapter 5 provides an overview of the research methods and algorithms used to validate the experimental studies. Chapter 6 discusses the various experiments conducted in detail to identify the impact of length and repetitive nature of surveys and the results of gamification as extrinsic motivation for long and repetitive task. Chapter 7 concludes with the lessons learnt and the scope for future work in this area.

CHAPTER 2

LITERATURE REVIEW

This chapter discusses literature related to this research. As this research focuses on areas of mHealth, the first section will briefly discuss the evolution of mHealth. The second section provides an overview of the chronic pain diseases and various mHealth applications available today for chronic pain management, focusing mainly on sickle cell disease. The third section talks about the importance of self-reported data and various validated measures available for collecting self-reported data from the patients. The fourth section explains in detail the importance of quantity and quality of data collected in clinical trials. Further the literature reviews the impact of survey length over the quality and quantity of data collected. In section five. The sixth section introduces the various motivational factors suggested for improving user engagement and data quality. There are subsections discussing intrinsic and extrinsic motivations with a focus on gamification and badges. Finally, in the seventh section, existing applications using gamification factors are discussed. As no significant research is found which verifies the impact of survey length and nature in mHealth apps, the focus of this research is discussed in brief.

2.1 Evolution of mHealth

Mobile health (mHealth) is a relatively new term for providing healthcare solutions via mobile devices, and for collecting and delivering patient related data to the clinicians and researchers in real time [46]. Bashshur et al. [12] stated in their research that telemedicine originated in 1969 which further led to the origination of eHealth in late 1990s and mHealth was introduced by Istepanian and colleagues [57] in 2003. According to the WHO Global Observatory for eHealth [1], mhealth is "*medical and public health practice supported by MDs, such as mobile phones, patient monitoring* *devices, personal digital assistants (PDAs), and other wire-less devices*." Mobile devices have the advantage of being portable, constantly connected via the internet, giving an individual a sense of privacy, being able to act like sensors for monitoring and collecting an individual's daily activity data. These factors have made mHealth popular over the period [37, 39, 66, 105, 107, and 110].

2.2 Domain of Chronic Diseases

The authors Michael Ashburn and Peter Staats [4] conducted a research on chronic pain management. According to them, chronic pain is commonly defined as "*pain that persists for longer than expected time frame for healing or pain associated with progressive, nonmalignant disease*". The patients in this condition commonly experience depression, sleep disturbance, fatigue and decreased overall physical and mental functioning. These patients frequently require appropriate medical care to address their pain experience.

One such chronic pain disease attracting clinician's attention these days is Sickle Cell Disease (SCD). SCD is inherited chronic pain disease largely affecting the African-American population in the U.S. Patients suffering from SCD usually suffer chronic pain in their bones, joints, back, abdomen and the chest [104]. This chronic pain is generally referred to as sickle cell disease vaso-occlusive pain which leads to recurrent intermittent pain, tissue damage, strokes and organ failure [104]. These patients require prompt inpatient hospitalization for monitoring their pain intensity levels. Pain intensity is broadly described as the measure of pain strength which a patient goes through due to the disease [108]. The levels of pain intensity are broadly categorized as mild, moderate and severe [108]. Delays in treatment for SCD makes pain more difficult to treat and eventually leads to prolonged hospitalizations. In addition, it can lead to disruptions in the patient's sleep/activity pattern, ability to cope with the pain, and a reduction in quality of life [116]. Therefore, it is critical to effectively monitor patients' current and past status including symptoms, pain level, and all other important information. Mhealth provides a suitable platform for the clinicians to be able to monitor their patients from anyplace at any time [37]. A significant number of applications have been designed and developed for helping the clinicians monitor chronic pain in patients with SCD [24, 43, 102, 103, and 116].

2.3 Pain Management using PROMIS measures

The authors Dansie and Turk [30] performed an assessment on patients with chronic pain. The authors suggested that the clinicians require information about patient's medical history, symptoms causing chronic pain and any patterns in the patient behavior over the time for making decisions regarding chronic pain treatment. The authors mention that as there is no specific instrument that can provide an objective quantification of the extent or severity of pain experienced by the patient, it can only be assessed based on verbal and non-verbal communication with the patient. This form of data collection is termed as self-reported or patient reported data. There are benefits to using this method for data collection. Information collected in this manner can be considered more accurate as it is directly being reported by the person undergoing pain, thus making it reliable for the clinicians to use it for suggesting various treatments.

To validate the self-reported data collected from the patients, the National Institutes of Health (NIH) initiated a multi-center cooperative group referred to as the Patient-Reported Outcomes Measurement Information System (PROMIS) in late 2004 [23]. PROMIS consists of valid, generalizable item banks used to measure key symptoms and health concepts applicable to a range of chronic conditions, enabling efficient and interpretable clinical trial and clinical practice applications of patient-reported outcomes (PROs). They may also assist individual clinical practitioners in assessing patients' responses to interventions and in modifying treatment plans based on these responses. Surveys are designed using these PROMIS validated question sets. The clinicians make use of these surveys to determine patient's pain intensity levels.

Initially these surveys were delivered in the form of paper-based surveys to be filled in each time the patient visits the clinic or recording the surveys conducted over a telephone conversation in pre-decided time intervals [62, 106]. However, researchers have identified flaws in these data collection measures as the patients must recall recent health experiences and recall is unreliable and rife with inaccuracies and biases [62, 106]. MHealth is playing a vital role to overcome these issues through mobile applications, which are intended to collect and monitor patients' data close to the time and from remote location [106]. As SCD is a specific type of chronic pain disease, there are validated PROMIS measures available for validating and collecting self-reported pain data. Researchers have worked upon developing effective mHealth applications using PROMIS measures for SCD as well. Few examples are: iACT - an interactive mHealth monitoring system to enhance psychotherapy for adolescents with sickle cell disease and SMART - sickle cell disease mobile application to record symptoms via technology; [24, 99].

2.4 Significance of Quantity and Quality of Data Collected in Clinical Trials

Clinical trials are conducted by medical researchers to determine better effective treatment measures by observing a target population of patients for a specific disease. The results of the trial are derived after a thorough analysis of the data collected as part of that trial. To statistically analyze the clinical data, its sample size / quantity and its quality play a vital role.

David et al., [15] conducted a study on the importance of sample size in the planning and interpretation of medical research. In this study, the authors mention that the results observed in clinical studies may differ because of the variability in sample size. A smaller sample size increases the risk of reporting a false-negative finding, thus jeopardizing the motive of the study conducted. James et al., [109] in their research have tried to report various issues in the data collection process for clinical research. The authors state that the ability to determine the better of two treatments is often the outcome of the trial hypothesis, the data elements chosen to evaluate the question and the magnitude of data collected over a specific time. A good sample size is required to have statistically sufficient data points for the clinician to determine various aspects of the study conducted and come up with reliable results.

The Institute of Medicine (IOM) defines quality data as "*data strong enough to support conclusions and interpretations equivalent to those derived from error-free data*" [31]. Binny et al., [68] have defined high-quality data as "*data which should be absolutely accurate and suitable for statistical analysis*". It should meet the protocol specified parameters and comply with the clinical study requirements. The authors' state that the data collected for analyses should possess only an acceptable level of variation that would not affect the conclusion of the study conducted. If the quality of certain data points is compromised then those must be excluded from the sample size thus reducing the total number of data points available for analysis.

We can conclude that both quantity and quality of data play a significant role to derive effective and accurate results from the clinical trials. This case is true for mHealth applications collecting self-reported data from the patients as well. As stated in section 2.4, the clinicians rely upon the self-reported patient data to provide various treatments. A compromise in the quantity or quality of self-reported data may jeopardize the clinical protocol, resulting into loss for medical researchers. Thus, we need to consider various factors impacting the quality and quantity of self-reported data and derive measures to improve them.

2.5 Impact of Survey Length on Quantity and Quality of data

Over 40 years ago, Cannell and Kahn [21] argued that, when the optimal length for a survey is exceeded, respondents become less motivated to respond, put forth less cognitive effort and may skip questions altogether, causing survey data quality to suffer. Empirical studies by Johnson et al. [65] and Kraut et al. [67] suggest that the problem may be especially acute in self-administered surveys where no interviewer is present to maintain engagement. Krosnick [69] coined the term 'satisficing' to describe the tendency for survey respondents to lose interest and become distracted or impatient as they progress through a survey, putting less and less effort into answering questions. The resulting behaviors typically include acquiescent responding, more frequent selection of non-substantive responses such as 'don't know', non-differentiation in rating scales, choosing the first listed response (i.e. primacy) and random responding.

We expected similar problems in web-based surveys as well. Indeed, these effects have been widely documented [52, 78 and 82]. In addition, experimental studies by Galesic and Bosnjak [40] and Lugtigheid and Rathod [76] have shown that, as questionnaires become longer, engagement declines, resulting in classic satisficing behaviors and even survey abandonment. They observed that lesser time was spent by the user to answer questions, which were delivered to the end of the survey as compared to the ones at the beginning. The authors concluded that as fatigue and boredom accumulate throughout the survey, the respondents might be less willing to invest the effort needed for good quality answers. Crawford et al. [27] conducted a research to study the perception of burden observed by the users in answering web-surveys. They have stated in their research that the non-response rates are lower when the user is delivered a shorter length survey as compared to the user who is delivered longer length web-based survey.

The general concept underlying all of this is generally referred to as ‘respondent burden’. Bradburn [17] described respondent burden as the combination of four factors: the length of the interview; the amount of effort (cognitive and otherwise) required of the respondent; the amount of emotional stress a respondent might feel during the interview; and the frequency with which the respondent is asked to participate in a survey. His central argument is that *‘respondents seem to be willing to accept higher levels of burden if they are convinced that the data are important*’. Bradburn also mentioned that making the interview ‘an enjoyable social event in its own right’ might lower a respondent’s perception of the survey’s burden and encourage engagement throughout a long survey.

2.6 Factors to Improve User Engagement and Quality of Responses

Psychologists have suggested the framework of Self Determination Theory (SDT) to improve user engagement and quality of task performed. Although the initial work leading to SDT dates to the 1970s and the first relatively comprehensive statement of SDT appeared in the mid-1980s [32], it has been during the past decade that research on SDT has truly flourished. The authors Ryan and Deci have performed a significant amount of research on SDT [32, 33, 98 and 99]. Through their research, the authors have stated SDT as “*approach to human motivation and personality that uses traditional empirical methods while employing an organismic meta-theory that highlights the importance of humans' evolved inner resources for personality development and behavioral self-regulation*”. In 2008, the authors performed another study on SDT [33]. In this study, they state that SDT addresses such basic issues as personality development, self-regulation, universal psychological needs, life goals and aspirations, energy and vitality, non-conscious processes, the relations of culture to motivation, and the impact of social environments on motivation, behavior, and wellbeing. The authors distinguish between different types of motivation based on the different reasons or goals that give rise to an action. These can be broadly categorized as: *autonomous*, *controlled* and *amotivation*.

Autonomous is the type of motivation in which people identify value associated to a specific activity and integrate that value into their sense of self. When people are autonomously motivated, they experience volition, or a self-endorsement of their actions. This can be further sub-divided into two types: *intrinsic* and *extrinsic* motivations which will be discussed later in section 2.6.1. Controlled motivation, in contrast, consists of external regulations in which one's behavior is a function of external contingencies such as rewards, recognition or punishment. When people are controlled, they experience pressure to think, feel, or behave ways. Both autonomous and controlled motivation energizes and directs behavior, and they stand in contrast to amotivation, which refers to a lack of intention and motivation. When people are amotivated, they either do not act at all or act without any specific intention. This type results from not valuing an activity [97], not feeling competent to do it [11], or not expecting it to yield a desired outcome [100]. Figure 1 provides an overview of the taxonomy of human motivation provided by Ryan and Deci [95].

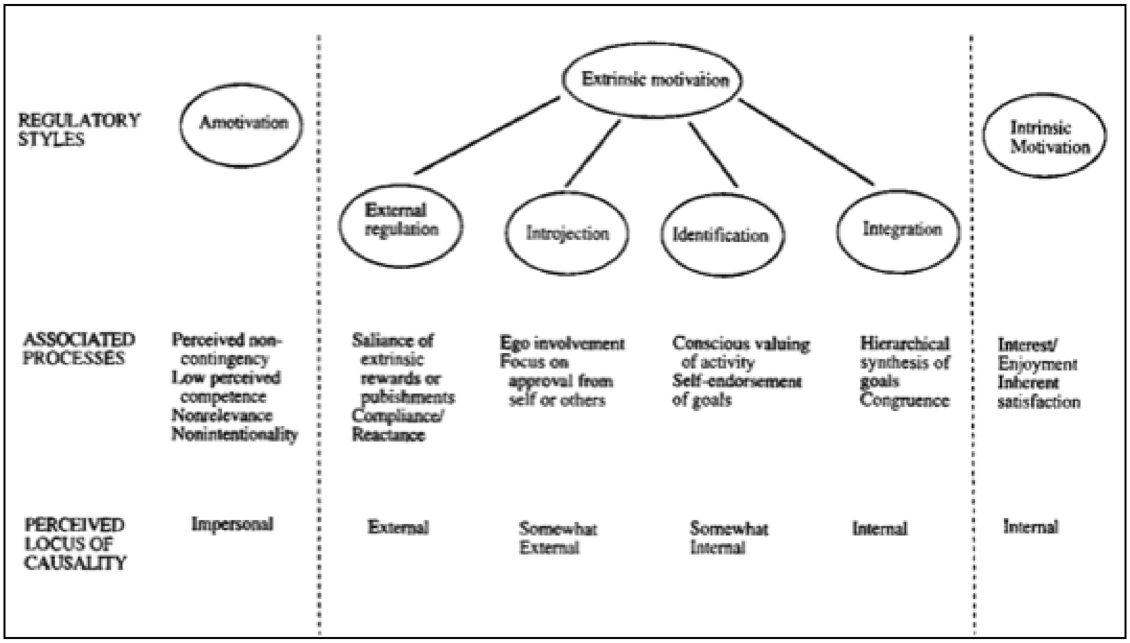


FIG.1. A Taxonomy of Human Motivation [95]

2.6.1 Intrinsic and Extrinsic Motivation

Ryan and Deci [98] have provided definitions for intrinsic and extrinsic motivation. Intrinsic motivation is defined as the doing of an activity for its inherent satisfactions than for some external consequences. When a person is intrinsically motivated, he is moved to act for the fun or challenges entailed in the task rather than because of external prods, pressures or rewards. Humans generally have the tendency to be active, inquisitive, curious and playful, displaying a ubiquitous readiness to learn and explore and they do not require any kind of external motivation to do so. Thus, one can say that humans, by default, are intrinsically motivated towards certain tasks as per their interest areas.

Extrinsic motivation is a construct that pertains whenever an activity is done to attain some external separable outcome. When a person is extrinsically motivated, he performs a specific task to receive either recognition or a reward or price for successful completion of a task. Ryan and Deci in their study [95], state that SDT proposes extrinsic motivation varies greatly in the degree to which it is autonomous for everyone. They have explained this concept using an example. A student who does his homework because he fears to get punished by his parents is extrinsically motivated because he is doing a specific task to escape the negative outcomes. Like this case if a student who does the work because he believes it valuable for the career is also extrinsically motivated because the student is doing it for an instrumental gain / reward as an outcome of the task completion. These examples give a clear idea as to how extrinsic motivations can differ in their relative autonomy from one individual to another.

Researchers have carried out various studies to find out the factors facilitating or undermining each of the intrinsic and extrinsic motivations for improving user engagement [49, 51, 93 and 98]. As this research is limited towards exploring extrinsic motivations, we will further consider the various factors that tend to assist this motivational type.

2.6.2 Gamification & Badges

As stated in the earlier section, a person is extrinsically motivated to do a task only with an expectation to achieve some reward or separable entity in return. These rewards can be of monetary or non-monetary in nature and can be used to recognize users' achievements. This research study will explore the concepts of non-monetary rewards in the form of gamification and badges.

During the last couple of years, gamification [34, 35, 36, 51, 63, 81 and 93] has been a trending topic and a as a means of supporting user engagement and enhancing positive patterns in service use, such as increasing user activity, social interaction, or quality and productivity of actions. Deterding et al. [35] have defined gamification as "*the use of game-play mechanics for non-game applications*". The authors state that gamification's main goal is to increase the user engagement by using game-like techniques to make people feel more ownership and purpose towards the tasks engaged. Gamification desires to combine intrinsic motivation with extrinsic one to raise motivation and engagement. Figure 2 gives a brief idea of how does the gamification concept work as motivation to change user behavior.

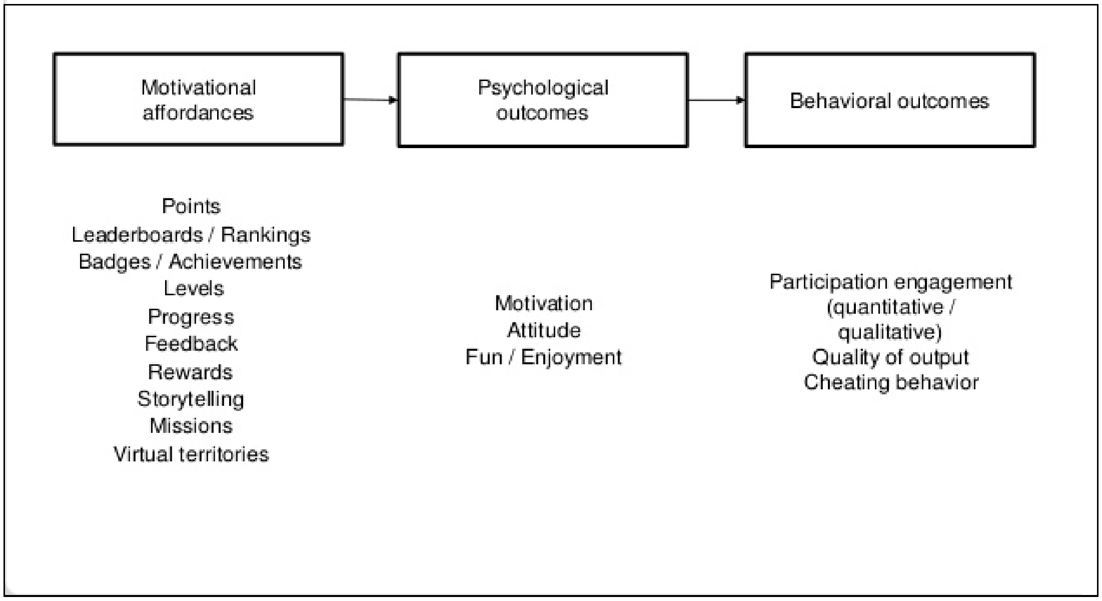


FIG.2. How Does Gamification Work? [35]

Chrons and Sundell [25] have stated that mundane activities especially for a longer period are not appealing but by combining these activities with simple games we can create a more effective way to motivate people. Deterding in his study on gamification as motivation [35], has stated that the entity being gamified needs to have some intrinsic value already, a reason for users to engage with it. But if the offered activity has core intrinsic value that user's desire, then weaving gamification into it can deepen their engagement and desire to participate.

Badges are one form of gamification factors introduced as extrinsic motivators to improve user engagement and quality of tasks performed. Gibson et al. [45] have defined a badge to be "*a representation of an accomplishment, interest or affiliation* *that is visual and available for the user to display*". The authors state that badges motivate continued user engagement in a task, which increases time spent on the task and supports skill acquisition through performance. Badges are said to fulfill five social and psychological functions: setting goals for users, instructing about possible further activities, visualizing past activity, providing status symbols, and supporting group identification [2]. It has been argued that badges function as a guidance mechanic in a service, providing the user with an idea of how the service is meant to be used and what is expected of the user, thus increasing the amount and quality of those actions within a service [50, 63, 83]. According to Bandura [10], set goals (such as those in badges) increase performance in three ways: (1) people anchor their expectations higher, which in turn increases their performance; (2) assigned goals enhance self-efficacy; (3) the completion of goals leads to increased satisfaction, which in turn leads to increased future performance within the same activities. These effects are further strengthened if the goals are context-related, immediate, and the users are provided with (immediate) feedback.

Considering this literature one can conclude that gamification factors & badges are being used as forms of extrinsic motivation, to encourage people to adopt usage of gamified applications in increased frequency. Let us further consider the applications where gamified factors are being used to improve user engagement and data quality.

2.7 Usage of Gamification in Online and Web Applications

As mentioned in the previous section, gamification has a great potential to improve user behavior and increase user engagement in various tasks. It has been implemented in various areas including, market research, education and health industry to name a few [22, 30, 42, 49, 51].

A significant amount of online and web applications have been developed in the market which make use of gamification factors to improve user engagement and data quality. Jennie et al. [71] conducted a pilot study using an iOS app which collected self-reported data from the end users. In their study, they made use of game mechanics such as badges, points and levels in the app to make the task more engaging. As it was a pilot study it had few limitations however the results of the study did show that introduction of gamification factors showed an improvement in user engagement levels in the app. Miller et al. [81] have performed a research on various applications making use of gamification factors for chronic disease management. Few of the applications mentioned as part of their research are: bant which uses points, levels, and social engagement, mySugr which uses challenges and quests, RunKeeper which uses leaderboards as well as social engagement loops and onboarding, Fitocracy which uses badges, and Mango Health, which uses points and levels.

As per section 2.5, literature shows us that survey length might have a negative impact on user engagement and quality of responses collected both in online and web applications. However, no significant research has been found to verify the impact of survey length in mHealth applications collecting self-reported data. This research mainly focuses on finding out whether does a long and repetitive survey via mHealth applications has an impact on the quantity and quality of self-reported responses collected. It further proceeds towards finding out whether an introduction of gamified measures helps improve the user performance in the mHealth apps delivering long and repetitive surveys. This leads us further to the next chapter, research context.

CHAPTER 3

RESEARCH CONTEXT

This research studies the impact of long and repetitive nature of surveys on user compliance and data quality in mHealth applications followed by deriving a method to improve these factors in the mHealth context. The main contribution of this thesis is to derive a method which will help me monitor the compliance and quality measures of self-reported survey responses and help in improving these factors over the period in a clinical trial. This chapter sets the context of my research by presenting the research questions, followed by a description of the research methodology consisting of a combination of response time and response pattern approaches resulting from literature survey and an overview of the case study conducted for validation.

3.1 Research Questions

This thesis' contributions are a case study in participatory design of an mHealth app for Sickle Cell Disease (SCD) patients, and an innovative method of compliance and data quality validation (response time and pattern methods) that attempts to assign good or bad labels to the self-reported responses collected as part of clinical studies. My research questions are:

> RQ1: Does the long, repetitive, and intermittent nature of surveys in a self-reporting task affect the user compliance and data quality in mHealth applications?
>
> RQ2: Can an intervention based of intrinsic and extrinsic motivational factors implemented in mHealth technology help improve compliance and data quality?

I would like to explain below two terms in detail before proceeding with discussing my research questions.

Compliance: As per Farlex Partner Medical Dictionary [38], compliance is defined as the consistency and accuracy with which someone follows the regimen prescribed by a physician or other health professional. This can be related to the number of surveys being taken by the user against the number of actual surveys delivered to him / her. It also measures the time used by the user to fill in a survey from the moment it was due. If a user is skipping certain number of surveys then he is less compliant than a user who is regularly filling in all the surveys delivered. If a user is filling up the surveys just before the deadline would be considered less compliant as compared to the user who is taking the surveys as soon as they are available.

Data Quality: As per Farlex Medical Dictionary [38], *data quality* is defined as a dimension of data contributing to its trustworthiness and pertaining to accuracy, sensitivity, validity and fitness to purpose. In the context of surveys, research has shown that questionnaire length has an impact on the data quality of collected responses [40]. With relation to this research one can measure how much time a user is spending on each individual question in the survey and can make assumptions regarding data quality. If a user spends relatively less time on a question, it impacts his perception about the question and his ability to provide an appropriate response. Then one can state that the data quality is compromised.

The first research question addresses whether the long and repetitive nature of surveys impacts the user compliance and data quality of self-reported survey responses collected from the patients in a mHealth app. In a clinical protocol targeted towards patients suffering from SCD, the patients must complete a set of surveys delivered to them on weekly or daily basis via a mHealth application. The patients are said to be compliant with the protocol if they respond to all the set of surveys delivered to them as part of the clinical trial. These self-reported responses are said to be of good quality if they are strong enough to derive interpretations and conclusions supporting the clinical protocol. My hypothesis is that the long and repetitive nature of mHealth survey does impact the compliance and data quality of the self-reported data.

The first research question simply asks if the compliance and quality are impacted due to the design of lengthy and repetitive surveys. Answering this question positively validates that the quantity and quality of data collected is low in the clinical practices using such mHealth applications. However, this alone is not sufficient to resolve the issues in mHealth applications with lengthy and repetitive surveys. The second research question dives deeper into the details of the measures to monitor user compliance and data quality in self-reported surveys that may help in deriving methods to improve these factors over a period. The methods for measuring response time and observing the response patterns of the self-reported data may help us derive conclusions whether the data collected is good or bad. Furthermore, the introduction of extrinsic motivational factors such as badges and gamification as rewards for providing compliant and good quality data may affect the clinical protocol in a positive way. This combination of multiple methods (response time and pattern) to determine a label of good or bad quality for the survey responses and accordingly provide extrinsic rewards as motivations to improve the user participation in the protocol is a novel contribution of this work. The next section will discuss various methods approached to assess the self-reported data collected from the mHealth app.

3.2 Research Methodology

This thesis focuses mainly on monitoring and evaluating user compliance and data quality measures in a self-reported survey task. The research community has emphasized on various methods to identify low quality responses from the data set [see Chapter 3 section 3.2.2]. These methods primarily rely on the responses collected as well as the time spent on each question in the entire survey. The mobile platform provides us a convenience to log each user interaction within the application and save it for further data analysis in the database. This methodology can be termed as clickstream analysis. This thesis makes use of data collected using clickstream analysis and passes it as inputs to the combination of methods to determine the data quality of the survey responses collected. The two main methods used in this thesis are of response time and response pattern and a combination of these two to derive more accurate results. I have explored both the methods individually and have used them to classify the survey responses as good or bad. However, I did find many survey responses that do fall into the intersection of these individual methods and are ambiguously classified if done using individual methods. Usage of these methods in a combination gives a better classification of the entire data set into good and bad responses which is better classification derived as compared to the individual methods alone. The next subsections will discuss these methods in detail. The discussion will start with clickstream analysis, followed by response time and response pattern methods discussed briefly, and lastly the combination of these methods.

3.2.1 Clickstream Analysis

The process of collecting, analyzing and reporting aggregate data about web pages visited by the users and the order in which they were visited is termed as clickstream analysis. Researchers perform this type of analysis by making use of the succession of mouse clicks made by the user and log all this data for further processing. This idea is very popular for web applications but limited literature is available for its use in mobile applications. The use of clickstream analysis in this research deals with logging all the user interactions within the mobile application including monitoring the survey activity as well as various mobile app life cycle activities.

The survey is designed to be consisting of multiple-choice questions with the answer options listed out in the order of Likert scale in the range of “Never” to “Almost Always”. The entire survey is designed to display one question on a page with Next and Previous buttons available once the appropriate answer option is selected for the question on the current page. These buttons help in navigating within the survey as well as to proceed and submit it on completion. The button clicks and answer option selections are logged for each question along with the timestamps. This type of data is referred to as log data and all this data is made use later to run the derived research methods to flag the survey response data as good or bad. Table 1 provides an insight of the log data captured for a question for a specific user within a survey activity. Similar logs are captured for the survey activity start button as well as submit button clicks as well to capture the total time spent by the user in completing a specific activity. This clickstream data is further submitted to the server and saved in the database for further use.

No	**Log Attributes**	**Description**
1	PIN	This is a 4-digit unique identifier provided to each user to receive and complete survey activities
2	ActvitiyInstanceId	This is unique identifier associated with the survey activity delivered to the user at any instance of time
3	QuestionId	This is a unique identifier from the question set saved in the database to identify a question delivered in the survey activity.
4	AnswerOptionId	This is a unique identifier from the answer option set saved in the database to identify an answer option selected for that question in the survey activity.
5	PrevButtonTimestamp	The timestamp captured when the previous button was clicked by the user for this question in the survey activity.
6	NextButtonTimestamp	The timestamp captured when the next button is clicked by the user for this question in the survey activity.

Tb1: Clickstream Log Structure

Along with the user interactions captured for in survey activity, all the app life cycle activities are also captured and sent back to the server using the same logging technique. The application used in this research is based on mobile platform and delivered via Android and iOS platforms. Thus, the mobile app has its own life cycle activities with respect to Android or an iOS platform that helps us know the exact timestamps as to when the app was started, paused (put into background), resumed and stopped during the period of clinical trial. The app life cycle activity log data is useful to check whether a user is fully compliant in the clinical protocol or not. This log data helps us determine the exact timestamps when the user interacted with the mobile app and for how long. Table 2 provides the list of life cycle activities; both Android and iOS which are captured in the user interaction logs and sent back to the server.

No	**Activity Name**	**Description**
1	Created	Called when the app context gets created for the first time in the mobile device
2	Started	Called when the app is started on the mobile device
3	Paused	Called when the app is paused due to any other activity on the mobile device such as receiving a phone call or a text message
4	Resumed	Called when the app is resumed back and is in focus in the mobile device
5	Stopped	Called when the application is stopped due to any outside activity or manually stopped or put in background by the user
6	Destroyed	Called when the app is killed or forcefully stopped on the mobile device

Tb2: Mobile Application Life Cycle Activities

This method will help in determining the user compliance levels of a participant in the specified protocol. The app life cycle activity log data will help us know whether the user is opening the app when an activity has been delivered to him with a mentioned due date. It will also help us determine whether the user is motivated enough to resume back a survey activity if intercepted by an external phone call or text message on the mobile device.

This section provided information about all the user interaction logging which is done within the mobile application; both in survey and in app activity logging which is required to be provided as input to the research methods used for identifying the survey responses as good or bad. The next section provides details regarding the research methods derived after going through existing literature for identifying bad quality responses.

3.2.2 Methods for Detecting Careless Responses

Respondent-administered web based surveys are known as efficient and cost minimizing method for collecting data. One of the advantages in web-based surveys is automated response encoding for closed-ended questions. Every click is automatically interpreted and stored in the data set, no matter if the respondent is earnestly completing the survey, is just having a look at the survey, or is filling in arbitrary responses to receive a reward. The arbitrary filling of survey responses forms the source of invalid data or meaningless data. Leiner [74] has defined meaningful response as the respondent's intention and ability to give a qualified answer to a question. This answer may be biased or purposefully faked; it is meaningful if it is an expression of the respondent's considerations on the question. Meaningful responses can depend upon the understanding of the questions by the user as well as the amount of effort put in by the user to understand and respond to the survey questions. The degree of understanding and cognitive effort varies within a broad continuum. The term *meaningless data* describes such data that was collected near the continuum's lower end.

There are several reasons why research participants may provide responses that are in some way invalid; that is, data that do not represent actual 'true' values. Johnson [66] identified three main classes of this invalid data: (1) linguistic incompetence or misunderstanding, (2) misrepresentation, and (3) careless or inattentive response. Linguistic incompetence deals with the construction of items and the process by which a survey is aimed (properly or improperly) at the population of interest. Misrepresentation deals with the issue of cheating, or faking behaviors that are most likely on high-stake surveys or tests. Carelessness or inattentive responding deals with participants who will simply do not put in the effort required to respond accurately or thoughtfully to all questions asked of them. In his paper [64] he states that hurrying on web-based inventories, combined with a sense of reduced accountability, increases the probability of response styles associated with too little attention. These styles could be reading items carelessly or not at all, random responding, skipping items, marking answers next to the wrong item, using the response scale in the wrong direction (marking "agree" when "disagree" was intended), and/or using the same response category (e.g., "3" on a 5-point scale) repeatedly to get through the inventory as quickly as possible to see the results.

To provide a comprehensive depiction of the phenomenon of interest, Huang et al. [54], proposed the label of *insufficient effort responding* (IER). IER is defined as a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses. IER underscores the cause of the response behavior without presupposing specific patterns or outcomes. Thus, IER includes random occurrence of response options as well as nonrandom repeated occurrence of the same response option. IER may vary in its intentionality — ranging, for example, from inadvertent misinterpretation of negatively keyed items to intentional disregard for item content, concealing one's true opinion.

The literature on careless response detection suggests six ways: (i) infrequency method, (ii) inconsistency method, (iii) response time, (iv) response pattern, (v) odd-even index calculation and (vi) Mahalanobi's distance calculation. Each method is described in detail below.

3.2.2.1 Infrequency Method

Beach [13] and Green et al. [48], state that the infrequency approach uses items in which all, or virtually all, attentive participants should provide the same response. The questions are designed to have one highly probable response and a deviation from this response provides an indicator for IER. For example, Green and Stutzman [48] embedded task statements that were clearly unrelated to a focal job in a job analysis inventory administered to incumbents. Beach [13] asked participant's questions such as "I was born on February 30th". Each of these questions has one clear answer and the presence of IER is easily detected when a participant selects improbable response options. A participant who incorrectly answers many infrequency items is presumed to display high levels of IER and the data can be considered as low quality data. Despite some evidence which shows that infrequency approach can detect instructed random responses [7, 8], few researchers [20, 54, 96 and 114] have mentioned that this approach may not be appropriate for detecting IER, because infrequency scales can confound IER with impression management and faking.

3.2.2.2 Inconsistency Method

Pinsoneault [91] states that the inconsistency approach is designed to use matched item pairs and compare the responses of one item with the other item. The item pairs are created in three ways, (i) direct item repetition, (ii) rational selection, and (iii) empirical selection. Researchers have measured inconsistency by incorporating repeated items into the surveys. Buechley and Ball [19] used 16 pairs of identical items in their study to identify careless responses. Wilson et al. [115] included repeated task statements in job analysis questionnaires and detected the respondents who provided inconsistent endorsement of task statements. Lucas and Baird [75] also recommend survey researchers to design very similar questions in different places of a questionnaire to check against IER. Empirical methods have also been heavily used to select item pairs for deriving inconsistency scales [18, 47, and 101]. Inconsistency scales have generally been effective at identifying random responses generated by participants with partial or no access to the questionnaire [6, 7, 14 and 112] as well as random responses generated by computer algorithm [2, 84]. However, using normal instruction has yielded mixed results in terms of the scales' effectiveness [3, 70, and 90].

3.2.2.3 Response Time Method

This method deals with measuring the overall time spent by the user to provide responses in the survey to determine IER. Curran [28] states that the time it takes for an individual to respond to a set of items is perhaps the most widely used tool for the elimination of careless responses. It is the most likely used on an intuitive basis even by those who have no knowledge of the literature. This intuitive use of response time can be independently derived by the practical extension of one simple assumption: there exists a minimum time needed to validly complete a survey. Huang et al. [54], states that the response time approach assumes shortened response time for IER than for normal responding because of the absence of cognitive processing. Normal or average response time will obviously be different for different surveys. Response time is likely to correlate with number of items, but some items take longer to complete than others. Accordingly, it is difficult to create concrete rules for response time that differ from normal outlier analysis unless the status of the participant as a C/IE responder is already known (unlikely), or experimentally manipulated [54]. This method has been used widely in research to determine careless responses [28, 54, 64 and 80] however it has its own limitations. Curran [28] states that the time that it takes one individual to respond thoughtfully may be drastically different than the time it takes another. Not quite as obvious is the fact that variation also exists in the time that it takes individuals to respond carelessly or with insufficient effort. Thus, relying only on this method is not sufficient for detecting careless responses.

3.2.2.4 Response Pattern Method

This approach in literature is known as ‘long-string analysis’ or ‘response pattern indices’ [54, 80]. This technique seems to have formally begun with Johnson's [64] use of a borrowed technique, which is later described in the work of Costa and McCrae [26]. This technique involves examining the longest string of identical responses from each participant. This may be calculated singularly, on the response option that is selected the most frequently [54, 80], or multiple times, once on each option for each participant [64]. The approach is straightforward; the assumption is that those individuals who are responding carelessly may do so by choosing the same response option to every question. The extension of this assumption is that individuals, who are responding carefully, and with sufficient effort, will not use the same response option for long periods of time. This technique also tends to be dependent on response option. That is, the typical frequency of a long-string on certain response options tends to be higher than on others; ‘Agree’ is a very popular choice on a typical scale [28, 64]. This property will vary from scale to scale and sample to sample. Because of these factors, long string analysis can be difficult to compare across different data collections without engaging in some degree of scaling. The main concern in this approach is what should be considered as a cutoff for determining whether a specific long string index indicates IER. Costa and McCrae [26] conducted a study on a sample of 983 respondents. The Likert scale used for the responses was in the range from “*Strongly Disagree*” to *“Strongly Agree”*. They found that none of their 983 cooperative participants selected the same response option more than 6, 9, 10, 14, and 9 times for the response options from strongly disagree to strongly agree, respectively. They recommended using these cutoffs to detect IER. Johnson [64] has stated that depending upon the length of the survey and the survey design the long string indices may vary for each response option used in the Likert scale.

3.2.2.5 Odd-Even Index Calculation

Many researchers have tried to use this approach of calculating odd-even consistency index from collected survey responses [54, 55, 59 and 64]. Curran [28] computes this consistency index by breaking everyone's responses on each one-dimensional subscale or facet of a measure into two sets of items: one set of responses to the even items and the other set to the odd items. For example: all positively worded items are even items and the exact opposite formation of items using negative words which would mean the same as positive ones but are worded differently are considered as odd items. Scores are calculated from the responses received for each of these items. These halved subscales are then averaged to provide an approximation of the individual's score on that subset of items. These values are paired by scale; the number of pairs produced is dependent on the number of one-dimensional groups of items that can be created in the larger scale. The lesser consistency among these scores of odd-even calculations, the responses are considered more towards IER factor. Although researchers like Jackson [59] have suggested that this approach can be used to easily detect response strings and careless responders, it does have its own limitations. The major limitation is that the odd-even index calculations can only be achieved if the survey is designed to have such components that are a contrasting pair of each other. Thus, this index calculation heavily depends upon the survey design and cannot be generalized to all survey methods.

3.2.2.6 Mahalanobi's Distance Calculation

Peck and Devore [89] state that outliers can be considered as unusual data points in relation to the entire data set distribution. Individuals who are not responding without sufficient effort are likely to differ from their thoughtful counterparts in some way and it is reasonable to believe that such responders can be considered as outliers. Basic methods of outlier analysis examine one value from a distribution relative to other values in that distribution. Mahalanobi's distance calculation is one technique that can be used for outlier analysis in statistics. As per Mahalanobis [77], this distance measure is a multivariate outlier technique, which is a simple extension of normal outlier analysis. This distance is the measure of how many standard deviations is the outlier point away from the normal distribution. By plotting all the data points and finding out the Mahalanobis distance between the data points helps to determine the outliers. These outliers are the careless responders in the surveys. This technique has been used extensively by researchers [28, 54 and 80] and has been found to be effective in finding IER when put together with other detection methods.

3.3 Methods Used in This Research

The earlier sections provide us with the review of available literature on what does careless responding mean and what are the available methods to detect such responses. Even though each method has its advantages, researchers [28, 54, 64 and 80] have shown that a combination of these methods provides better and more accurate results in identifying careless responders or low quality data compared to using individual ones. The combination can vary from one study to another as the combination of methods can be chosen depending upon the nature of survey questions being used in the study. The above methods are useful depending upon the nature of the survey questions and its design. For example, infrequency scale can be computed only when the survey has a contrasting pair of questions. This technique will not help if the survey questions are not designed in the contrasting manner. Similarly, odd-even index calculations may be done only when the survey questions are divisble into two equal halves and individual scores in relation to these halves can be calculated.

The survey questions being used in my research are derived from PROMIS short forms and thus do not support the contrast pairing of questions as expected in infrequency scale calculation. The nature of these questions does not support the survey being divided into two halves that will help us calculate the odd-even indices. Thus, taking the survey design into consideration and the limitations of each method, I have decided not to use the infrequency scale and odd-even index calculations to determine careless responses in my study. The questions used in the survey are derived from the PROMIS short forms question bank. The short forms are designed in a manner that the same content is generally not repeated in more than one question. Thus, determining inconsistency indices from this questionnaire is not feasible for the current study. My study consists of 25 questions being delivered via the mobile app that is lengthy in terms of the surveys being delivered on mobile apps. Thus, response time calculations, response pattern calculations and Mahalanobis distance calculations can be easily used as a combination of methods to determine careless responses and further low quality data submission.

To use these methods to verify the user compliance rates and quality of self-reported data, surveys in mHealth application had to be delivered using a mobile platform. The next section gives an overview regarding the PROMIS mHealth application used to conduct the experimental studies required to collect self-reported data from SCD patients. The above-mentioned research methods were then used in the form of algorithms explained in Chapter 5, to classify the collected data, set as ‘Good’ or ‘Bad’.

3.4 PROMIS mHealth App

As part of this research, all validation studies are done making use of an mHealth application developed for SCD patients using PROMIS validated measures. This application is called the SCD-PROMIS app and is available on Google Play and iTunes stores for Android and iOS respectively, as well as a web-based version. Arizona State University (ASU), in collaboration with the Pain Management Care Complex, a part of the Sheikh Zayed Institute's Pain Medicine Initiative at Children's National Health System (CNHS) in Washington, D.C., developed an mHealth platform for monitoring pain-related clinical outcomes. This research project was the development of PROMIS app for SCD. Prior to this application, the clinic relied upon paper-based survey instruments to collect data regarding patient pain intensity and burden. As described in section 2.1 from Chapter 2 this approach had several shortcomings. Patients were asked to describe pain intensity and burden after-the-fact at clinic interviews. Data was collected in the form of paper surveys, which mostly had to be transcribed later into a computer system for analysis. The main observed problem was a lack of patient compliance, as not many patients would visit the clinic as per the schedule to provide this data. To avoid all these issues, development of the SCD-PROMIS app was undertaken.

The evolution of this app dates to 2012. From 2012 to the present this app evolved through multiple versions. Fig1 provides a detailed project timeline of the SCD-PROMIS app.

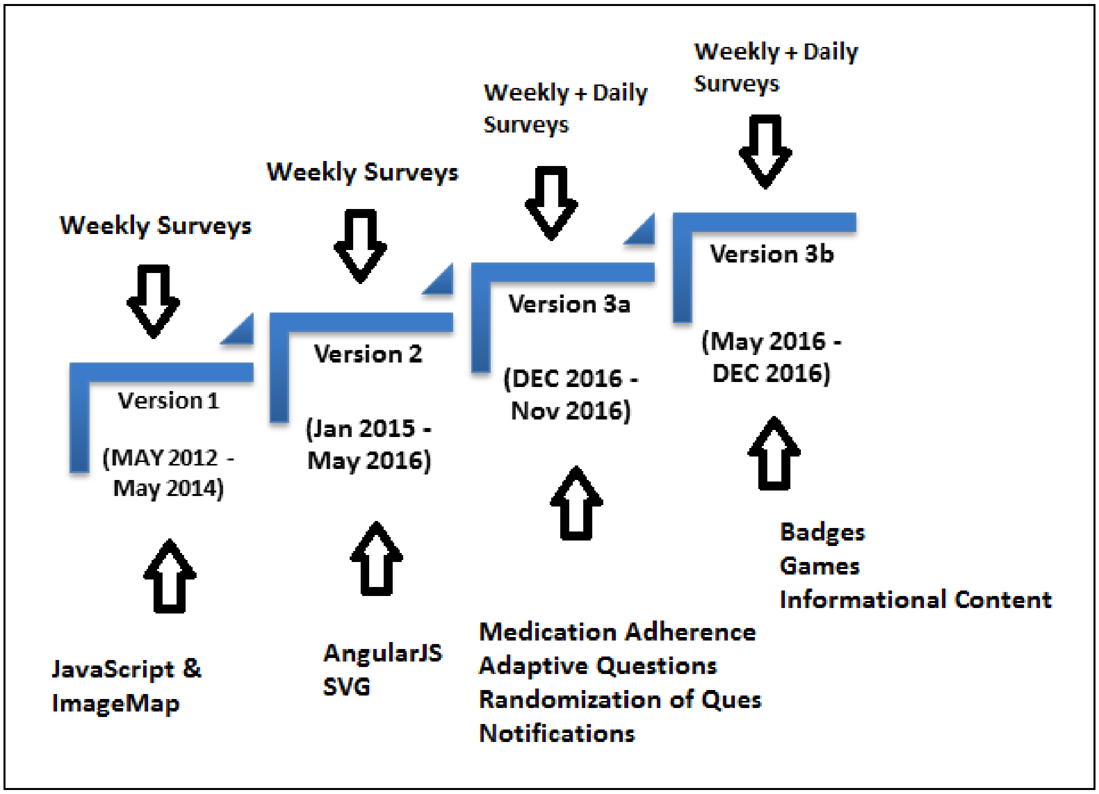


Fig 1: Evolution of PROMIS app for SCD

Version 1 of the app was developed summer 2012. This version was an Android app using JavaScript to deliver weekly surveys. Each weekly survey comprised of 4 blocks of questions formed using PROMIS validated measures for pain management along with one body pain question. The body pain question used an image map to select the appropriate body part for pain intensity. The total number of questions delivered per survey was thus 25. This version acquired an IRB approval for testing in a clinical context. In 2014, this app was deployed for a pilot study. The primary results of 75% compliance were achieved as compared to 12.5% compliance with paper surveys in the clinic. The clinical trial was stopped due to a lack of resources after only a few weeks and a handful of surveys completed, so results were promising but hardly conclusive.

The second version of the app improved the technology base by porting to AngularJS, CSS and HTML5 in a *WebView*, and enhanced native components for notifications. This version ran cross-platform in Android, iOS, and modern web browsers. This version changed the body pain question to be a Scalable Vector Graphics (SVG) question where a body image of front and back parts were displayed and clickable. A patient could click the most painful body part and indicate the pain intensity on a clickable scale ranging from 1 to 10, 10 being the most painful. This version also delivered 25 questions in deterministic order with the first question being the SVG and remaining 24 questions derived from PROMIS validated measures. These 25 questions were delivered as part of weekly surveys to the patients each week in their entire duration in the trial. Supported by philanthropic funding, this version of the application was deployed under an approved IRB for conducting a clinical trial starting in May 2016 and still active as of March 2017. Each patient in this study enrolls for 12 weeks. Further technical details about this version are provided later in section 4.3 of Chapter 4. The results from this ongoing clinical study are further discussed in detail in section 6.2 of Chapter 6.

In August 2016, CNHS and ASU received support from the Pfizer foundation for a two-year project to develop a third version of the app and conduct a clinical trial with 80 pediatric and adult SCD patients to study predictors with the aim of on reducing hospital readmission rates. The third version of the PROMIS app was developed in two phases; version 3a and 3b.

Version 3a introduced new non-PROMIS questions as well as technical changes in the entire application. From this version both weekly and daily surveys are delivered via the app. Weekly surveys now consisted of multiple-choice multiple answer questions. The SVG pain intensity question from the weekly survey was removed and embedded in the daily survey. Along with the 4 blocks of PROMIS validated questions, a few new questions related to medication adherence and adaptive questions on pain burden are included. A total of 31 questions are delivered via the weekly survey. The daily survey consisted of 3 to 4 questions related to the pain medication prescribed to the patient. Randomization of the weekly survey questions is implemented in this version. Randomization occurs between the 4 PROMIS blocks as well as within each question block, ensuring no two identical question orders are delivered to a patient in the trial period. Notifications in the app were corrected in this version. The previous version of the app would over-notify patients, leading to a form of *alarm fatigue*, causing patients to ignore the notifications and not conduct the surveys (or in the worst cases, turn off notifications on their smart devices or uninstall the app altogether). A provision to log user interactions within the app and save it to the data-store was made available as part of version 3a. All the app life cycle activities were also logged for each patient to be able to analyze user compliance and data quality measures of the self-reported data collected. Further details about this version are provided in section 4.3 of Chapter 4.

Version 3b mainly concentrated on the inclusion of intrinsic and extrinsic motivational factors within the PROMIS app to try and increase user engagement and lead to greater user compliance and data quality. Intrinsic motivations include short informational messages in the 'Did you know?' format being delivered in the app. These messages reminded the patient why it was important her her/him to faithfully complete the surveys to the best of her/his ability. Extrinsic motivations included badges and games as rewards on survey activity completion. These motivators encouraged engagement and thereby compliance through visible markers of achievement (badges) and enhanced gameplay (powerups as rewards). This version was mainly used to study the impact of these motivations on user compliance rates and data quality observed in the clinical trial. Details regarding the badges and games are provided in section 4.3 of Chapter 4.

Versions 3a and 3b together were deployed in the clinical study in December 2016. As mentioned earlier this is a 5-week study to study predictors and interventions on reducing hospital readmission rates. This is an ongoing clinical study and would end in December 2017. Patients are actively being recruited in this trial and some preliminary results are available. Details regarding this study and the results observed are explained in section 6.4 of Chapter 6.

3.5 Evolution of User Studies in This Research

This research consists of four user studies. All these studies were conducted using the PROMIS app described in section 3.4 in the university with the student consent and in the hospital using parental consent and assent from the child.

The first study was a pilot study conducted at Arizona State University in November 2015. In this trial version 2 of the PROMIS app was used. A four-week trial was setup that delivered weekly survey activities to the students enrolled in software engineering coursework SER 515. This was a controlled experiment conducted on a specific set of graduate students. The weekly activities consisted of questions based on PROMIS validated measures for anxiety control. The aim of this study was to check whether the problem of lesser user compliance and bad data quality is observed in the responses collected from mHealth application. Results and conclusions of this study are further provided in section 6.1 of Chapter 6.

A second study was conducted in the hospital at D.C on actual SCD patients. This study was of 12-week duration and it is still ongoing. This is expected to end by May 2017. Version 2 of the PROMIS app was used in this study. Weekly surveys were delivered as part of this study as well. These surveys consisted of question sets designed from validated PROMIS measures for pain management for SCD patients. The aim of this study was to validate the results obtained from study 1 on actual patient self-reported data via mHealth application. Section 6.2 of Chapter 6 provides details about this study and the results and conclusions derived.

Results obtained from study 1 and study 2 helped me validating my first research question RQ1. Response time and pattern methods described in section 5.1.1 and 5.1.2 from Chapter 5 were used to interpret these results and derive appropriate conclusions. However, there were no ground-truth values for these methods in the existing literature with respect to mHealth surveys. Hence, there was a need to derive baseline values for the PROMIS SCD app itself. Study 3 was conducted with this aim in a controlled classroom environment on students from software engineering coursework SER 421 in November 2016. Version 3a of the PROMIS app was ready by then and hence was used in this experiment. This study was conducted in a single sitting and appropriate data was collected to analyze and obtain the required baseline values. Details regarding the study and conclusions are explained further in section 6.3 of Chapter 6.

After obtaining the baseline values from study 3, a final study was conducted on SCD patients at CNHS under an approved IRB. The aim of this study was to validate the second research question RQ2 of this thesis. This study was started in December 2016 and it is ongoing with an end date of December 2017. As version 3a and 3b had all the required changes in the app to validate RQ2 (intrinsic and extrinsic motivations) both these versions were used in this study. This is a 5-week trial where both weekly and daily surveys are delivered to each patient. The results obtained so far help validate RQ2 and appropriate conclusions are drawn. These details are provided in section 6.4 of Chapter 6.

In summary, this chapter sets the context for this thesis research by providing details about the research questions, research methods derived by the combination of response time and pattern methods and finally giving details about the case studies conducted to validate these methods using PROMIS mHealth app. The next chapter discusses the iterative design process and the implementation of the PROMIS app, followed by a full presentation of the validation studies done using the research methods described in this chapter.

CHAPTER 4

DESIGN AND IMPLEMENTATION PROCESS

This chapter discusses in detail about the design and implementation of the various versions of the PROMIS mHealth app for Sickle Cell Disease (SCD) patients as described in section 3.4 of Chapter 3. It is important to note here that the design and development of this app was a combined effort of a team of software engineers under the guidance of medical professionals treating SCD. This chapter does a detailed walkthrough of all the technical components used in the development of PROMIS app platform. The first section talks about the PROMIS protocol in general. This is followed by an overview of the platform developed as part of the PROMIS app project in the second section. This chapter is largely from a paper presented at the IEEE-NIH 2016 Special Topics Conference on Healthcare Innovations and Point-of-Care Technologies [46].

I was one of the developers in the software engineering team. My involvement with the PROMIS app started from version 2 as described in section 3.4 of Chapter 3. After version 2 was deployed for the clinical study, I handled the post-production issues and derived reports of user compliance from this study. The low compliance results obtained from this trial motivated me towards this research. This led me towards forming my research questions as part of this thesis stated in section 3.1 of Chapter 3. Further my role was to design and develop version 3a and 3b which included randomization of survey questions, implementing notifications, embedding the app with intrinsic and extrinsic motivations as described in section 3.4 of Chapter 3. I was a core member for designing the database, which is used to collect the self-reported patient data from version 2 and version 3a and 3b. I was also responsible for the design and implementation of the web-portal, which was developed during version 3a implementation. This portal was developed for the clinicians to manage the ongoing clinical trials.

4.1 PROMIS Protocol

PROMIS (Pain Reported Outcomes Measurement Information System) is a set of validated self-reporting instruments for a variety of physical and non-physical health outcomes. A significant number of mHealth applications have been designed to make use of question sets from PROMIS measures to collect self-reported pain information [17-20]. Data from these self-reported surveys is collected and further analyzed by the clinicians, which helps them address each individual patient with better treatment procedures suitable to him. Thus, quantity and quality of survey data collected plays an important role in analyzing patient pain intensity levels and providing effective interventions.

Prior to the development of PROMIS app, the Pain Clinic at Children's National Health System (CNHS) relied upon paper-based survey instruments to collect data regarding patient pain intensity and burden. This had several shortcomings. Patients first were asked to describe pain intensity and burden after-the-fact at clinic interviews. Patients could complete the paper surveys at home (but few rarely did) and collected paper surveys were transcribed into a computer system for analysis. A significant obstacle was that collecting the data required visits to the clinic, which often resulted in cancelled appointments due to personal conflicts. Therefore, compliance is a significant issue in conducting any clinical trial related to pain reporting. A software platform was created through an interdisciplinary design process, identifying key requirements and desired features aimed at addressing these issues. This led to the evolution of the PROMIS SCD app as described in section 3.4 of Chapter 3.

4.2 PROMIS Platform Overview

The platform architecture is a straightforward application of modern software systems engineering principles. The app, described in section 4.3, is a hybrid app with native mobile components and cross-platform web components. The app is driven from a REST Application Programming Interface (API, section 4.4) that determines what activities to deliver to the patients at given times and receives results of pain reporting activities. In this way, the platform is extensible and personalized. A second server-side component is a Pain Reporting Portal (section 4.5) that enables clinicians to provision new patients, see completed activities and basic compliance reports, and export data for detailed analysis. Figure 1 provides an overview about the PROMIS platform.

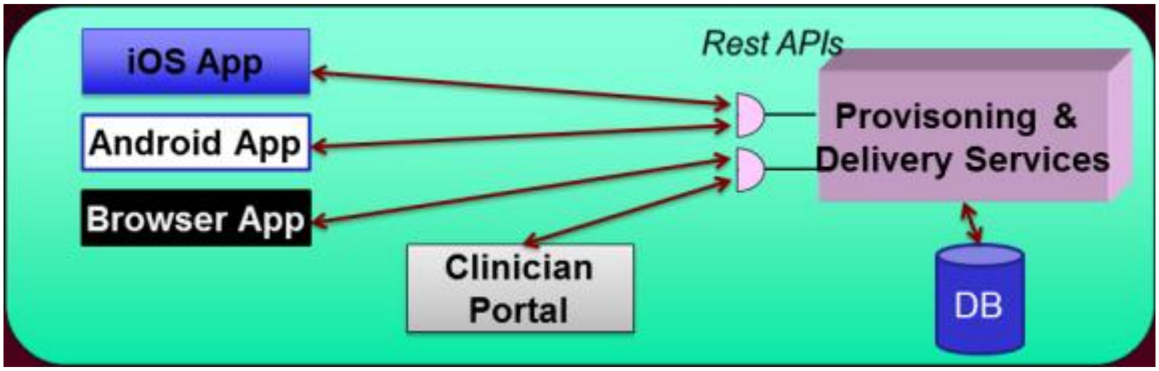
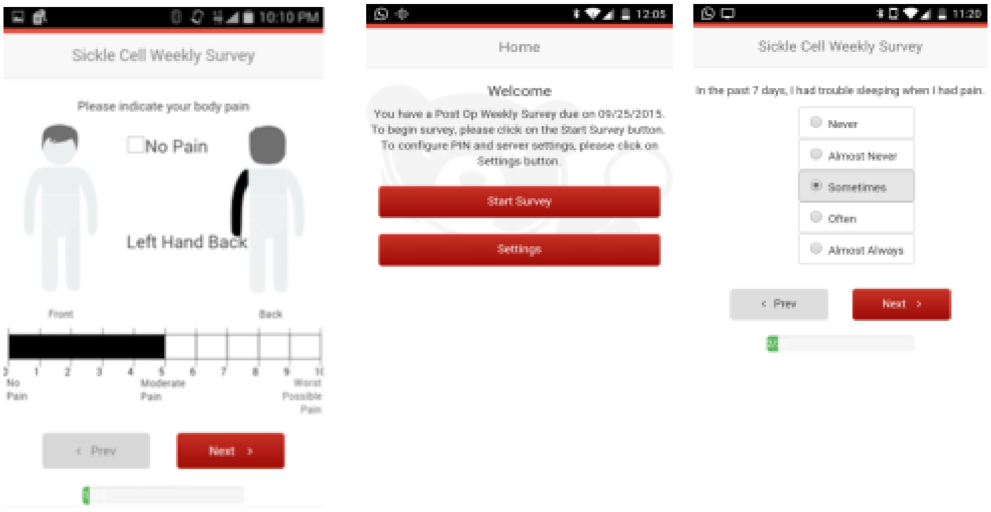


Fig1 – PROMIS Platform Overview

4.3 Pain Reporting App Based on PROMIS Measures

As mentioned in section 3.4 of Chapter 3, the PROMIS app evolved in multiple versions till date. Version 1 of the app was developed using JavaScript. This version delivered 25 questions in a weekly survey. The first question is a body pain intensity question implemented as an image-map, while the remaining 24 questions are presented as 4 blocks of PROMIS validated questions. This version was developed prior to my involvement in this research project.

Version 2 of the app was developed to rewrite the entire app using AngularJS, CSS and HTML5. This version too delivered a set of 25 questions in a weekly survey. The image-map for the body pain question was replaced by a Scalable Vector Graphic (SVG) question. This SVG had two body images displaying both front and back of a human body. The body parts that could be selected were Head Front, Head Back, Chest, Back, Abdomen, Lower Back, Left Hand Front, Left Hand Back, Right Hand Front, Right Hand Back, Left Leg Front, Left Leg Back, Right Leg Front and Right Leg Back. The remaining 24 questions were formed as 4 blocks of PROMIS validated measures using pediatric short forms 8a v1.0 surveys: pain interference, physical function (mobility, short form 10a), fatigue, and anxiety. Some questions from short forms were omitted at the discretion of the clinical lead of the project. This app had basic functionality of delivering the survey over a mobile device on a weekly basis. Below figures are the screenshot images of this version of the PROMIS app.



Sickle Cell Weekly Survey

Please indicate your body pain

□ No Pain

Left Hand Back

Home

Welcome

You have a Post Op Weekly Survey due on 09/25/2015.
To begin survey, please click on the Start Survey button.
To configure PIN and server settings, please click on Settings button.

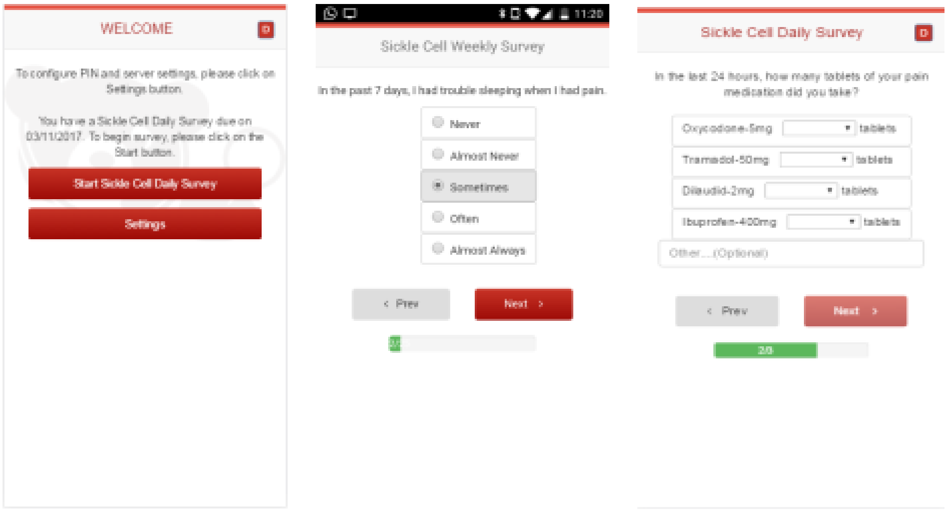
Sickle Cell Weekly Survey

In the past 7 days, I had trouble sleeping when I had pain.

○ Never
○ Almost Never
● Sometimes
○ Often
○ Almost Always

Fig 2: PROMIS app Version 2

The next version developed for the PROMIS app was version 3a. As mentioned in section 3.4 of Chapter 3, this version underwent few changes in the survey design as well as the questions which are being delivered as part of the survey. This version of the app delivers weekly as well as daily surveys to the patient. The SVG question is removed from the weekly survey and instead a multiple-choice multiple answer body pain questions are being asked. This question is followed by the 4 sets of PROMIS validated questions. Two new types of questions got introduced in this version. These are medication adherence and adaptive questions. All these together form 31 questions delivered to the patient via the weekly surveys. As part of daily surveys, 3 to 4 questions are being delivered which are related to the pain medication dosage prescribed to the patient. This version introduced randomization in the survey questions being delivered for weekly surveys. Randomization is achieved at two levels: (i) in between the 4 PROMIS blocks and (ii) within the questions of a PROMIS block. The figure below provides a view of the PROMIS app delivered as part of version 3a.



WELCOME

To configure PIN and server settings, please click on Settings button.

You have a Sickle Cell Daily Survey due on 03/11/2017. To begin survey, please click on the Start button.

Sickle Cell Weekly Survey

In the past 7 days, I had trouble sleeping when I had pain.

○ Never
○ Almost Never
● Sometimes
○ Often
○ Almost Always

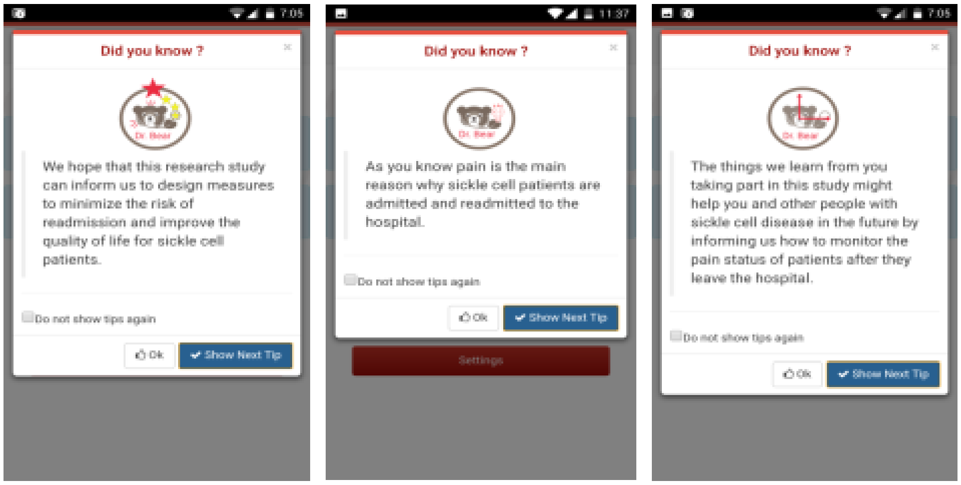
Sickle Cell Daily Survey

In the last 24 hours, how many tablets of your pain medication did you take?

Oxycodone-5mg ▾ tablets
Tramadol-50mg ▾ tablets
Dilaudid-2mg ▾ tablets
Ibuprofen-400mg ▾ tablets
Other....(Optional)

Fig 3: PROMIS app Version 3a

The next version of the app developed as per the project timeline displayed in section 3.4 of Chapter 3 is version 3b. This version introduced the intrinsic and extrinsic motivational factors into the PROMIS application. Intrinsic motivation was in the form of short informational messages educating the patient regarding SCD and the importance of responding to the surveys delivered via the app. These messages would come up as small pop-ups with an option to close them as well as to show the next tip available. Figure 4 displays app screenshots with informational content.



Did you know ?

We hope that this research study can inform us to design measures to minimize the risk of readmission and improve the quality of life for sickle cell patients.

□ Do not show tips again

Did you know ?

As you know pain is the main reason why sickle cell patients are admitted and readmitted to the hospital.

□ Do not show tips again

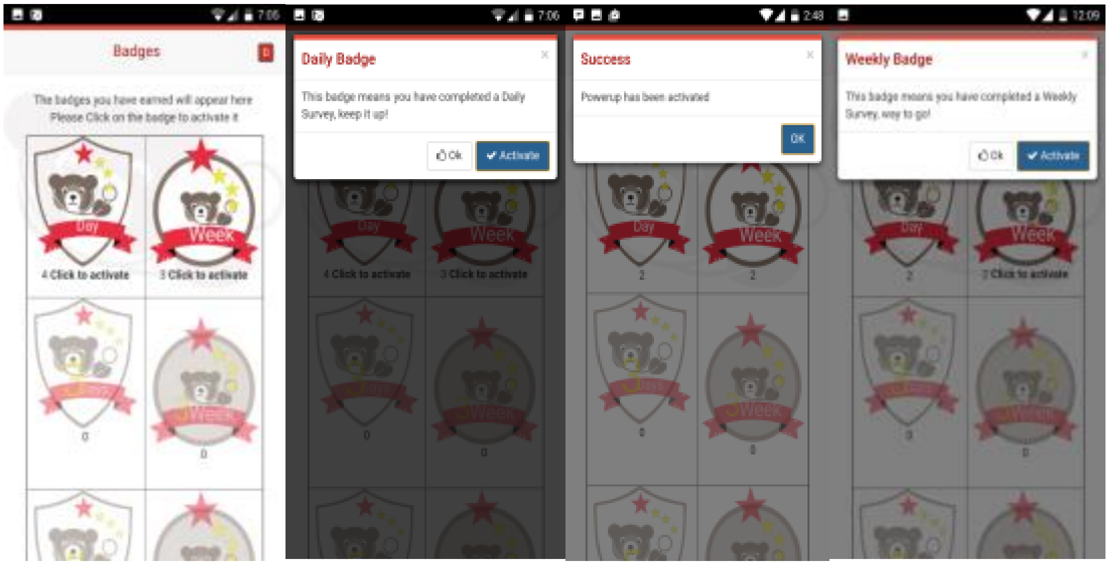
Did you know ?

The things we learn from you taking part in this study might help you and other people with sickle cell disease in the future by informing us how to monitor the pain status of patients after they leave the hospital.

□ Do not show tips again

Fig4. PROMIS app Version 3b – Intrinsic Motivators

As part of extrinsic motivation badges and games were introduced in this version. There are 6 types of badges recognizing the effort in completing weekly and daily surveys combined. When a pending activity is completed, appropriate badge associated with the activity gets listed in the patients received badge list. The badges that have been received are displayed as colored images and the ones that are yet to be achieved are greyed out. A little description regarding each badge can be seen by clicking on the badge. Each badge is associated with some powerups that help the user to have a better game play in the games center. These powerups can be activated only once by clicking the badges the patient has achieved. These powerups are then utilized during the game play to enhance the experience. The images below give a better picture of how the app looks with badges and powerups display. Below figure displays the badge board and a type of weekly and daily badge activation screenshots from the app.



Badges

The badges you have earned will appear here
Please Click on the badge to activate it

Daily Badge

This badge means you have completed a Daily Survey, keep it up!

Success

Powerup has been activated

Weekly Badge

This badge means you have completed a Weekly Survey, way to go!

Fig5. PROMIS app Version 3b – Badges and Powerups

Two open source games have been taken from Github and modified as per the PROMIS app requirements. These are developed using HTML5, JavaScript and jQuery and are embedded in the app released to the app stores. We contacted the authors of these games via email to derive their permission to use these games in the PROMIS app. Once the permissions were received then the game code was copied into the app code. The patients, who receive the enhanced content version (version 3b) of the app, have all time access to these games. Games act as part of extrinsic motivation to attract them towards using the app more often. The powerups, which get activated from the Badges page automatically, are added to each user's individual game play. The correct number of powerups are displayed and updated on their usage irrespective of the game the user chooses to play. The game play start time and end time are recorded in the database for the data analysis purposes. Below images display how the dashboard for the games and their individual screens look like.

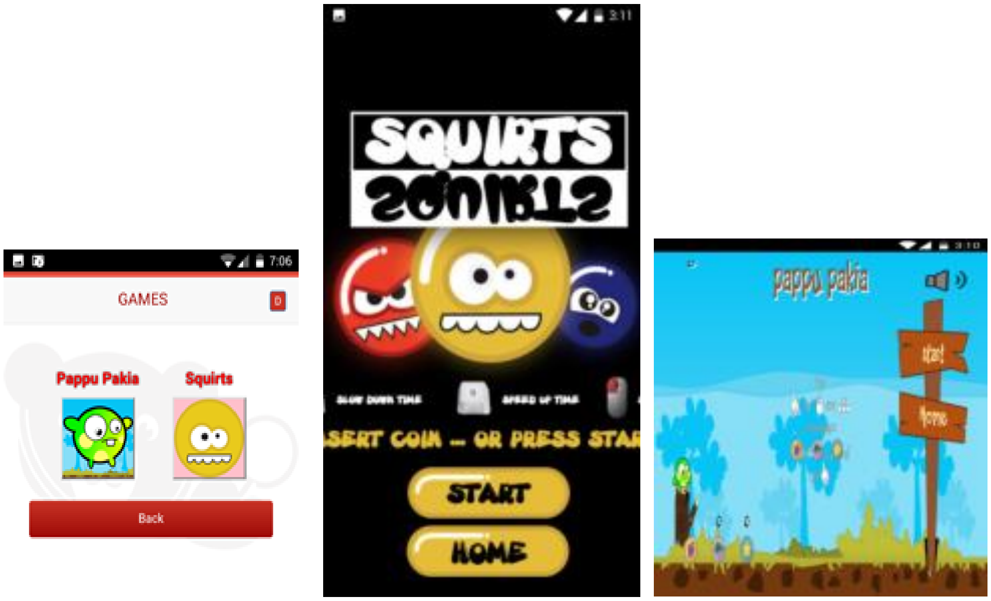


Fig6. PROMIS app Version 3b – Games

4.4 REST API

The app provides the cross-platform portability capability is by leveraging the HTML5 “stack” as described in the previous section. The other way is by driving the app using a REST API with JSON as the data language. REST is an architectural style attributed to Fielding [23]. REST is more suited to mHealth apps than heavier SOA-based architectures for mHealth [24] for several reasons, most notably it provides a lighter, more open architectural approach and that it is readily amenable to modern web-based apps. JSON (JavaScript Object Notation) is the lingua franca of client-server web communications, as it is readily consumed by front-end web centric mobile apps such as ours.

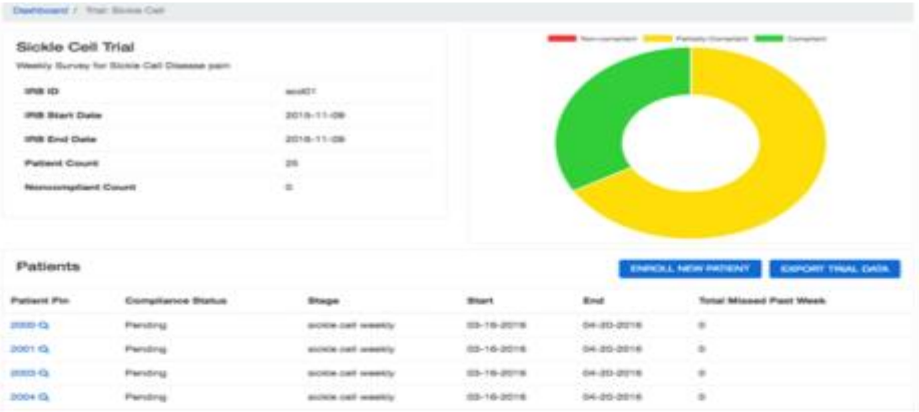
The REST API exposes server-side behaviors controlling when various activities are available to a patient, retrieving an activity definition for a patient (the set of survey questions), and submitting survey results to the server. This information is passed back and forth between client and server using JSON, and the app merely becomes a “player” of the activity information it receives.

As the PROMIS app underwent multiple changes in the progressive versions, the API was also changed to support these app changes. In version 3a new question types with multiple-choice multiple answers were introduced. API underwent changes to support these question types and save them appropriately to the database. In version 3b, badges and games were introduced. Each patient's badge data and game play had to be tracked and used the next time the patient performs any activity. This data needed to be saved in the DB for each patient. Thus, API was changed to help save this data along with respective changes in the database schema to support badging and gaming in the app.

4.5 Pain Reporting Portal

The other user-facing component of the system is a web-based portal used by clinicians and their staffs to provision new patients in a study, monitor compliance of patients already enrolled in the study, and export data for offline analyses. Along with version 2 of the PROMIS app, the development of this Portal took place. The development of this portal started as a team project as part of software engineering course SER 515. This portal was developed to replace the two older incarnations of a simple portal being used in the PROMIS project. Version 2 got deployed in an ongoing clinical study in which the clinicians utilize this portal to enroll patients into the trial and to monitor their survey activity at regular intervals. The portal works directly from the same database as the REST API developed to support version 2 of the app.

The portal provides a central place to manage patient enrollment and get quick visual reports on patient compliance under a given clinical trial. The donut chart in Figure 3 gives a visual view of compliance over the trial population, while the other built-in reports show compliance for individual patients. Softer interpretations of compliance are possible by tracking task completion using timestamps. We use this information to indicate partial compliance if a patient completes an activity but delayed from the date/time when the protocol schedule requested it be completed. We plan to add a notification feature that will proactively alert clinicians when patients are severely non-compliant, and allow for manual or automated notifications to such patients through portal settings. The portal may be accessed using any standard web browser and includes a responsive user interface design meaning it is also usable on smartphones with small screen sizes. Figure 7 shows the main compliance-monitoring screen of the study.



Sickle Cell Trial

Weekly Survey for Sickle Cell Disease pain

IRB ID	scd01
IRB Start Date	2016-11-09
IRB End Date	2016-11-09
Patient Count	25
Noncompliant Count	0

Patients

Patient Pin	Compliance Status	Stage	Start	End	Total Missed Past Week
2000	Pending	sickle cell weekly	03-16-2016	04-20-2016	0
2001	Pending	sickle cell weekly	03-16-2016	04-20-2016	0
2003	Pending	sickle cell weekly	03-16-2016	04-20-2016	0
2004	Pending	sickle cell weekly	03-16-2016	04-20-2016	0

Fig7. Clinical Trial Dashboard View

Currently the portal is undergoing changes to support the versions 3a and 3b of the PROMIS app. As part of these changes, the portal will utilize the REST API developed to support latest version of the PROMIS app, and will be completely decoupled from the database.

This chapter provided the technical details behind the evolution of the PROMIS app in various versions along with other components that support the deployment of this app in user studies. The next chapter gives an insight regarding the mixed methods approach being used in this research to evaluate the research questions. These methods are used in three different algorithms derived as part of this research.

CHAPTER 5

MIXED METHODS AND ALGORITHMS

The main contribution of this thesis is a mixed methods approach to identify and filter out the bad quality responses from the self-reported data collected via mHealth applications. This approach extends the use of clickstream data and user activity logs to determine the data quality of the response set. This chapter discusses the mixed methods approach in detail. The first section covers the methods in brief followed by sections describing the algorithms derived using these methods for assessing data quality.

5.1 Mixed Methods

In Chapter 3 section 3.2, various methods from literature survey were introduced for detecting careless responses in web and mHealth surveys. Amongst these methods, the ones that are suitable for the design and implementation of the PROMIS app used in this research are response time and response pattern methods. The next two subsections will provide details regarding how these methods will be used within the PROMIS app to detect careless responses.

5.1.1 Response Time Method

This method deals with measuring the overall time spent by the user in completing the survey activity delivered via the mobile application. As described in section 3.2.3 of Chapter 3, everyone requires some finite amount to time to grasp the question content and provide an appropriate answer option for it in a survey activity. The calculation of response time for each survey activity will help determine the data quality factors associated with it. Response time method assumes that a shortened response time can be considered to produce bad quality data as compared to normal response time. In this research, this method is used individually to determine whether a specific self-reported survey activity responded by the user in certain amount of time is considered as shortened time or normal response time. Accordingly, a data quality flag is associated to the survey activity mentioning it as good or bad quality response by the user. To calculate the response time for a survey activity, the logs collected as part of clickstream analysis described in Section 3.2.1 are be passed as input to this method. The timestamps collected in the logs help derive the total response time for that specific activity. In section 3.2.3 of Chapter 3, threshold values were mentioned for this method. These values were derived with respect to web surveys. To my knowledge, the existing literature does not provide any specifications regarding threshold values derived for response time method in the context of mHealth surveys. This lead to a need to define baseline values to use in this method. Using version 3a of PROMIS app (section 3.4 of Chapter 3), we conducted an exercise to define such a baseline (details provided in section 6.3 of Chapter 6). Once the baseline value was derived, a comparison is done between the response time calculated for a specific response with the threshold value. A data quality flag is assigned to the response as per the comparison results achieved.

5.1.2 Response Pattern Method

This method (also known as 'long-string analysis or 'response-pattern indices') involves examining the longest string of identical responses within a survey activity completed by a specific user. As discussed in Chapter 3 Section 3.2.4, this method works on a basic assumption that those individuals who respond carelessly may do so by choosing the same response to every question or most of the questions delivered as part of a single activity. The extension of this assumption is that responses with lesser longer string index value present have higher chances of being good quality data as compared to the ones with greater longer string pattern value. In this research, this method is used to analyze the survey responses provided in the self-reported survey activity by the user and determine whether the response set can be considered as good or bad quality data. The in-survey clickstream log data collected as part of Section 3.2.1 is provided as input to this method. This method further checks whether a long string pattern of same answer option by the Likert scale are being observed in individual survey activity response set. As mentioned in the earlier section, existing literature did not any threshold values to be considered for this method in the mHealth context. Chapter 6, section 6.3 discusses about the experiment conducted to derive a threshold value to be considered for the long string pattern method. Accordingly, if a survey response consists of a long string pattern with index value greater than the derived threshold then the response is tagged as ‘Bad’ quality data, otherwise it is tagged as ‘Good’.

5.1.3 Combination of Response Time and Pattern Methods

The methods described in the above sections individually help in distinguishing good quality responses from bad ones. However, when these methods are used in combination with each other, the good and bad quality data classification is improved. For example: There can be a survey activity response which has the total response time greater than or equal to the threshold level decided for response time method. Thus, as per the response time method this activity will be tagged as good quality data. If the same activity has a long string index greater than the threshold decided for response pattern approach then it will be tagged as bad quality data. The results produced by the usage of these methods individually may produce ambiguous results for certain survey responses that fall in the criteria mentioned above.

To resolve this ambiguity, I propose a solution that makes use of the combination of both the response time and response pattern methods to classify the ambiguously categorized survey responses into either good or bad data. In this method, first the survey response time is taken into consideration to filter out good and bad responses. This filtering will make use of the threshold level derived for the response time method from section 6.3 of Chapter 6. If the response time is less than the threshold level, irrespective of whether a response pattern is observed or not, a bad data quality flag is assigned for the activity response. If the response time is greater than the threshold level, then the activity response set is checked for long string pattern. If a long string pattern greater than the threshold derived for pattern method (section 6.3 from Chapter 6) is observed, then the total time spent in the long string block is calculated. If this block response time is lesser than the threshold for response time then the entire survey activity response is flagged as bad else as good quality data. Making use of a combination of these methods helps to classify the entire self-reported data set into good and bad quality responses without having any ambiguity.

5.2 Algorithms for Detecting Careless Responses

To make use of the above methods in the classification of self-reported data collected from mHealth surveys, I have designed three algorithms for each of the above method respectively. Each response set collected from the participants in the experimental study will be passed as inputs to these algorithms to derive a data quality flag for that response. The next sub-sections provide details regarding each of them.

5.2.1 Algorithm Using Response Time

Using response time as an input factor, I have derived the below algorithm to derive a data quality output flag for any response set. The inputs to this algorithm are the survey response collected via the mHealth app and the threshold values derived for response time from study 3 explained in section 6.3 of Chapter 6. The output is a data quality flag in one of three possibilities: Fast, Normal, or Slow, as per the calculations performed. The details of the algorithm are provided below. This is followed by a hypothetical example of a clinical trial scenario where the patients participate in a trial of weekly surveys and provide their responses. The data analysts make use of this algorithm to derive a data quality flag against each response collected.

5.2.1.1 Algorithm

INPUT:

1. JSON String (survey response)
2. Minimum time per question (seconds) MRT-LOW
(Derived from pilot study 3 in Chapter 6)
3. Maximum time per question (seconds) MRT–HIGH
(Derived from pilot study 3 in Chapter 6)

OUTPUT:

1. Data Quality Flag
(FAST/NORMAL/SLOW)

STEPS:

```
i)      Initialize the Boolean data_quality_flag = GOOD
ii)     Check for total time spent in answering the entire survey. This will be
        labeled as "response-time".
iii)    Calculate the average time spent per question in the survey by using the
        below formula
        Avg. Time spent per question (seconds) = "response-time" /
        "Total no of questions"
iv)     IF this calculated average is less than MRT-LOW value passed to the
        algorithm
            Return DATA_QUALITY_FLAG = FAST
v)      ELSE IF MRT-LOW < calculated value < MRT-HIGH
            Return DATA_QUALITY_FLAG = MODERATE
vi)     ELSE IF this calculated average is greater than MRT-HIGH then
            Return DATA_QUALITY_FLAG = SLOW
vii)    END
```

5.2.1.2 Example

I will explain this algorithm with an example. John is a 12-year old patient suffering from Sickle Cell Disease (SCD). He and his mother go to the pain clinic for a regular check-up. Now, the clinic is undergoing a study on pediatric SCD patients to determine the pain levels they go through over a period of 4 weeks. Under this trial, each patient receives a survey of 25 questions each week using version 2 of the PROMIS SCD app (see section 3.4 of Chapter 3 and 4.3 of Chapter 4). Each weekly activity is active in the system for 2 days. The doctor asks John if he would like to participate in this study. John is excited to be a part of this study as he was getting a $50 Amazon gift card in return for his participation. John and his mother sign the consent form. Later the clinician helps John download the SCD-PROMIS app on his mobile phone. A personal identification number is provided to John that will be used to save his survey responses in a de-identified manner.

All the settings are fed into the app and John is set to go for completing his first weekly activity in the clinic itself. The clinician helps him go through each question and select an appropriate response option. Once the survey is filled, the responses are submitted and the activity gets completed. The clinician reminds John that he would be receiving a similar survey in the next consecutive 3 weeks and he needs to submit the responses on time before the survey expires. John agrees to do so and the clinician hands him over the gift card he was looking for. John thanks the clinician and leaves the clinic along with his mother.

During this time, another patient called Jill signs up for participating in the study too. Jill is 13 years old. After a week's time from their clinic visit, both patients John and Jill receive their second weekly survey activity on their respective mobile devices via the PROMIS app. Jill opens the app and starts the activity. She reads each question carefully trying to understand what the question means. She selects an appropriate response to each one of them and submits the survey activity successfully. On the other hand, John does not pay attention to the reminder he receives regarding the weekly survey activity on his device. John misses out completing the survey on the first day. On second day, his mother reminds him about the survey and asks him to finish it before it expires. He opens the app on his device and starts the survey. However, as he was getting late for his play time, he tries to hurry up through the survey. He does not read each question carefully and simply selects any random response option for the questions. Once the survey is filled, he submits it successfully. As the survey gets submitted, John goes back to playing with his friends.

Now, at the clinic, the data analysts try to look at the responses received from both John and Jill for their respective surveys. They see that Jill has completed her survey in 140 seconds and John has completed his survey in 40 seconds. The analysts decide to use the response time algorithm to derive a data quality flag for each of these responses. The threshold value considered for the MRT-LOW is 4.8 seconds and for MRT-HIGH is 7.2 seconds.

Taking Jill's response time, below results are observed for response time algorithm. 'response-time' = 140 seconds. The average time spent per question by Jill = 140 / 25 = 5.6 seconds. The derived avg. time per question for Jill is compared with the threshold values, MRT-LOW and MRT-HIGH. As per the comparison, MRT-LOW < Jill's average time per question < MRT-HIGH. As the derived value falls within the Normal range, Jill's response is flagged as 'Normal' for data quality.

Taking John's response time, following results are observed for response time algorithm. 'response-time' = 40 seconds. The average time spent per question by Jill = 40 / 25 = 1.6 seconds. The derived avg. time per question for John is compared with the threshold values, MRT-LOW and MRT-HIGH. As per the comparison, John's average time per question < MRT-LOW. Thus, his response is flagged with 'Fast' for data quality measure.

The above example clearly explains how the analysts made use of the response time algorithm to calculate the data quality flags for both patients in the clinical trial.

5.2.2 Algorithm Using Response Pattern

Using response pattern as an input factor, I propose a modified algorithm to derive a data quality output flag for any response set. The inputs to this algorithm are the survey response collected via the mHealth app and the threshold values derived for response pattern from study 3 explained in section 6.3 of Chapter 6. The output is a data quality flag in either of the two possibilities: Good or Bad as per the calculations performed. The details of the algorithm are provided below. This is followed by a hypothetical example of a clinical trial scenario where the patients participate in a trial of weekly surveys and provide their responses. The data analysts make use of this algorithm to derive a data quality flag against each response collected.

5.2.2.1 Algorithm

INPUT:

1. JSON String (survey response)
2. Long String Index threshold value – LST
(Derived from pilot study 3 in Chapter 6)

OUTPUT:

1. Data Quality Flag
(GOOD/BAD)

STEPS:

```
i)      Initialize the Boolean DATA_QUALITY_FLAG = GOOD
ii)     Traverse through all the response options provided by the user in the order
        they were submitted
iii)    Calculate the maximum long string pattern observed in the response data.
        This value will be termed as "response-pattern"
iv)     Check if this "response-pattern" value is less than or equal to the LST value
        passed to the algorithm
v)       IF YES then

            Return DATA_QUALITY_FLAG = GOOD
vi)     ELSE
            Return DATA_QUALITY_FLAG = BAD
vii)    END
```

5.2.2.2 Example

Let us consider the below example to understand the execution of this algorithm. This example is an extension to the one described in section 5.2.1.2. There are two pediatric patients with SCD, John and Jill. They are participating in a 4-week clinical trial to understand the pain levels that these patients go through over the trial period. The version of the PROMIS app being used is version 2, which delivers weekly surveys of 25 questions each to the patients. Both John and Jill have completed their weekly trial in the clinic under the clinician's supervision.

In the second week, both patients receive their weekly survey. In the previous example, we saw that John did not pay much attention to the survey questions and tried to answer them in a hurried manner. Thus, his response set was tagged as 'Fast' with respect to data quality. Whereas Jill completed her survey in sufficient time and her response was tagged as 'Normal' with respect to response time algorithm. In this scenario, we are trying to look for any patterns in the response set collected from them.

The data analysts now take Jill's response and try to look for long string patterns. As mentioned earlier, long string patterns are formed when the same response option is selected multiple times for consecutive questions in the survey. For this example, let us consider the responses provided for the first two PROMIS blocks questions. The response options in the survey are mapped in the Likert Scale manner starting from 'Never' to 'Almost Always'. The position of these options on the screen determines the option order. Below table gives an idea about the option order for these responses.

Response Option	Option Order
Never	1
Rarely	2
Sometimes	3
Always	4
Almost Always	5

Tb1: Response Options and Their Order

Keeping this in mind, John's response set for the first 2 PROMIS block questions looks like: {1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 3}. The analysts look at this response string and apply the response time algorithm on it to derive appropriate data quality flag. To do this, they use the threshold value for response time derived from section 6.3 of Chapter 6. The value derived is greater than long string index, LST = '6'. The long string index value from John's response set for the 12 questions is '8' which is from question 4 to question 11. 'response-pattern' calculated for the collected response was 8. Now compare this value with the threshold value for pattern. As per the comparison, John's pattern index '8' is greater than LST. Thus, as per the response pattern algorithm, John's response set is flagged as 'Bad'.

Later Jill's response for the first 2 PROMIS blocks was taken for analysis. Her response set looks like: {1,1,1,3,1,4,1,3,2,1,2,1}. The analysts look at this response set and apply the pattern algorithm to it. The long string index value from Jill's response set for the 12 questions is '3'. No long string pattern is found in his response. As per the comparison with the threshold value LST, Jill's pattern index '3' < LST. Thus, as per this algorithm, Jill's response set is flagged as 'Good'.

This algorithm mainly tries to see if the participant tried to game the survey by answering in the same pattern, which might result due to tapping at the same place on the screen multiple times forming a rhythm. As explained in section 5.2.1.2, John tried to hurry up while answering the survey. This might have led to forming a rhythmic pattern of screen taps, thus making him select the same answer options multiple times for the consecutive questions delivered in the survey. This example describes how the data analysts made use of the response pattern algorithm to calculate the data quality flags for both patients in the trial.

5.2.3 Algorithm Using Combination of Time and Pattern

As mentioned in section 5.1.3, the mixed method of combining response time and pattern is derived to resolve some ambiguous cases different data quality outputs are received using time and pattern algorithms alone. This method makes use of both the algorithms in a combination that double checks the response set for time, once overall and second time within a pattern block found (if any). The inputs to this algorithm are the survey response collected via the mHealth app and the threshold values for both time and pattern derived in section 6.3 of Chapter 6. The output is a data quality flag which gives the response in either of the two cases: (i) either Fast/Normal/Slow or (ii) Good /Bad. The details of the algorithm are provided below. This is followed by a hypothetical example of a clinical trial scenario where the patients participate in a trial of weekly surveys and provide their responses. The data analysts make use of this algorithm to derive a data quality flag against each response collected.

5.2.3.1 Algorithm

```
INPUT:

1. JSON String (survey response)
2. Long String Index threshold value - LST
        (Derived from pilot study 3 in Chapter 6)
3. Minimum time per question (seconds) –MRT-LOW
        (Derived from pilot study 3 in Chapter 6)
4. Maximum time per question (seconds) - MRT–HIGH
        (Derived from pilot study 3 in Chapter 6)
```

OUTPUT:

1. Data Quality Flag
(GOOD/BAD)
(FAST/NORMAL/SLOW)

STEPS:

```
1. Check for total time spent in answering the entire survey. This will be labeled as “response-time”.
2. Calculate the average time spent per question in the survey by using the below formula
        Avg. Time spent per question = “Total no of questions” / “response-time”
3. IF this calculated average is less than MRT-LOW value passed to the algorithm

        Return DATA_QUALITY_FLAG = FAST
```

```
4. ELSE IF this calculated average is greater than MRT-HIGH then

        Return DATA_QUALITY_FLAG = SLOW
5. ELSE
        a. Traverse through all the response options provided by the user in the order they were submitted
        b. Calculate the maximum long string pattern observed in the response data. This value will be termed as “response-pattern”
        c. Check if this “response-pattern” value is less than the LST value passed to the algorithm
        d. IF YES then
                Return DATA_QUALITY_FLAG = GOOD

6. ELSE
     a. Calculate the total time taken in the block where long string pattern has been found.
```

```
      b. Calculate the average time spent per question in the pattern block by using the
      below formula
          Avg. Time spent per question = Long string pattern index value /
          response time in the block
     c. IF this calculated average is less than MRT-LOW value passed to the algorithm
               Return DATA_QUALITY_FLAG = FAST
     d. ELSE IF this calculated average is greater than MRT-HIGH then

          Return DATA_QUALITY_FLAG = SLOW
    e. ELSE
          Return DATA_QUALITY_FLAG = GOOD.
7. END
```

5.2.3.2 Example

Let us extend the example used in sections 5.2.1.2 and 5.2.2.2 to understand the execution of this algorithm. The background for this example is that there are two pediatric patients with SCD, John and Jill. They are participating in a 4-week clinical trial to understand the pain levels that these patients go through over the trial period. The version of the PROMIS app being used is version 2, which delivers weekly surveys of 25 questions each to the patients. Both John and Jill have completed their weekly trial in the clinic under the clinician's supervision.

We will modify this example slightly to understand the advantage of this algorithm. In this context, Jill's response time calculated for the entire response set is 110 seconds. Thus, average time per question for Jill will be '4.4' seconds. Threshold value for response time is with MRT-LOW = 4 seconds and MRT-HIGH = 6 seconds. Taking the response time algorithm, the data quality flag for Jill's response = 'Normal'.

The data analysts now decide to use the combination algorithm. As per this method, if the data quality flag received using response time algorithm is either 'Fast' or 'Slow' then we simply exit the algorithm and return that appropriate quality flag as output. In Jill's case, the flag returned from time algorithm is 'Normal', thus, the analysts proceed with the algorithm. The next thing it checks is whether a long string pattern is found in the response set. The response options and their order remain the same as explained in table 1 in section 5.2.2.2. Jill's response set for all 24 questions from 4 PROMIS block looks like: {1,1,1,3,1,4,1,3,2,1,2,1,1,1,1,1,1,1,1,1,2,1,2,1}. The analysts calculate the 'response-pattern' index as '9' in her response set. Compare this value with the pattern threshold value '>6'. Now, as per the combination algorithm if the response pattern index is less than the threshold value then return the data quality flag as 'Good' and exit. Else if the pattern index is greater than the threshold value, the next steps must be executed. In this case, the analysts find that pattern index for the response is greater than the threshold value and thus proceed further with the algorithm. The last and final check in combination algorithm is to derive the response time per question in the pattern block found. Once this value is obtained then it is compared with the response time threshold to finally determine the quality flag for the response set. In this case, Jill's response time for the long string pattern block from question 12 to 20 is calculated as 30 seconds. The average time spent per question in the pattern block by Jill is calculated as (total time in the pattern block) / (long string pattern index). Thus, for Jill it is (30/9) = 3.33 seconds. Now this value is compared with the response time threshold range MRT-LOW = 4 seconds and MRT-HIGH = 6 seconds. The calculated average time per question in the pattern block is < MRT-LOW for Jill's response set. Thus, as per the combination algorithm, her response is tagged as 'Fast'.

This example explains the exact reason for deriving the combination algorithm. There can be some cases where the data quality flags derived from response time and pattern algorithms are a different. In such cases, the combination of the two helps to acquire better results as it checks the response time for the overall set as well as within a pattern found (if any) in the response set.

This chapter briefly explains the mixed methods and algorithms derived from them to determine the quality of collected response data. These methods and algorithms are further used in the experimental studies explained the next chapter.

CHAPTER 6

EXPERIMENTAL STUDIES

This thesis focusses on validating the two research questions mentioned in Chapter 3, section 3.1. The first research question (RQ1) is to evaluate whether long, repetitive and intermittent nature of surveys in a self-reporting task affect user compliance and data quality in mHealth applications. A pilot study was conducted using a modified survey set in version 2 of the PROMIS app (section 3.4 from Chapter 3 and section 4.3 from Chapter 4) in November 2015 to check whether the problem of decline in compliance and data quality exists in the mHealth context. Section 6.1 provides further details about this study. In May 2016, a usability study was conducted in a clinical context under IRB approval on Sickle Cell Disease (SCD) patients using the version 2 of PROMIS app. Section 6.2 provides details about this study. Results obtained from study 1 and study 2 help validate that there is an actual problem observed in the user compliance and data quality obtained within the context of lengthy and repetitive mHealth surveys. The first research question was validated; however, I could not use the data from these studies to derive baseline values for the mixed methods approach mentioned in Section 5.1 from Chapter 5. In November 2016, a pilot study was conducted in a controlled classroom environment to derive baseline or ground truth values for the response time and response pattern algorithms. Details regarding this experiment are provided in section 6.3. Further an enhanced version of the PROMIS app was used to validate the second research question (RQ2) of this thesis. RQ2 is to evaluate whether an intervention based of intrinsic and extrinsic motivational factors implemented in mHealth platform help improve compliance and data quality. In December 2016, version 3b of the PROMIS app was used to conduct a usability study on actual SCD patients in the clinic to validate RQ2. Section 6.4 provides details about this study. This chapter analyzes all the experiments that were conducted using the mixed methods approach and the algorithms, the results of the experiments and the limitations observed with the scope for future improvement.

6.1 Study 1

6.1.1 Experimental Context

A pilot study was conducted in November 2015 with the graduate students enrolled in the software engineering coursework SER 515. The aim of this study was to check whether the problem of decline in user compliance and data quality exists in the mHealth context. This study was conducted in a controlled environment to avoid any specialization factors such as rewards for doing the surveys, or participants medical condition having an impact on the results obtained.

The mHealth app used in this study is based on version 2 of the PROMIS app as mentioned in section 3.4 from Chapter 3 and section 4.3 from Chapter 4. As the participants were non-SCD patients the survey questions were redesigned using PROMIS short forms for measuring anxiety levels in the students towards the end of semester. The survey was designed to deliver a set of 25 questions per week. First question consisted of a Scalable Vector Graphics (SVG) question. The remaining 24 multiple-choice single-answer questions were picked from validated PROMIS measures. The set of questions and the order of their delivery were repeated each week for every student. The length and format of survey questions were designed to mimic the surveys delivered via the version2 of the PROMIS mHealth app for SCD patients. All questions were compulsory, and the survey would not proceed if a question were not answered. This web application could be downloaded as a mobile app on Android devices or it could be used over the browser. The surveys were taken anonymously to avoid any bias during data analysis and protect student confidentiality. Each week the responses to the surveys were recorded and saved in the database. The recorded data included information about the answer option chosen for each question, time spent on each question as well as if the question was revisited.

A total of 62 students participated in this study for a duration of 4 weeks. Data from all 62 students was included to analyze the results and derive appropriate conclusions. Out of these 62, 50 users used the web-version of the mHealth app and remaining 12 used the Android app version. Students were rewarded with an extra credit in their coursework for participating in this study.

6.1.2 Results

The self-reported data was used to analyze two important factors: user compliance and data quality. For studying user compliance, survey response data from all the 62 students was considered. Tb1 and figures 1 and 2 are the charts that display the results observed for user compliance from the collected data. The total number of students who completed all four weekly surveys was 35. Out of these 35, 10 were Android app users. Total number of students who completed three out of four weekly surveys was 12. Out of these 12, 2 were Android app users. The number of students who completed two out of 4 surveys was 7 and the ones who completed only one survey was 1.

Week	**Compliance (%)**
1	100
2	98
3	85
4	63

Tb1: Weekly Compliance Measure – Study 1

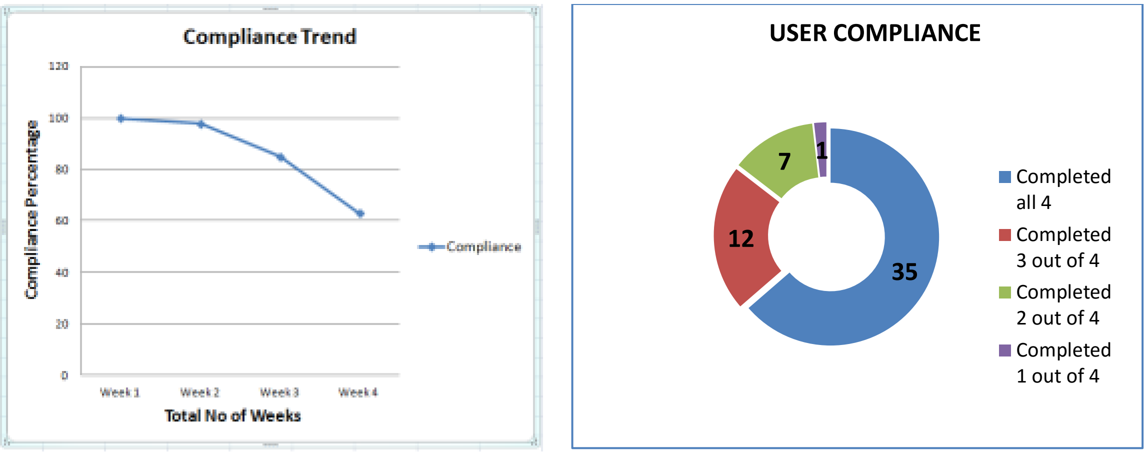


Fig 1: Line and Pie Charts Displaying User Compliance

The observations displayed above regarding user compliance clearly show that all the Android app users were more compliant in completing the surveys over the weeks as compared to web app users. One reason behind this observation could be that the Android app users were receiving push notifications on their mobile device each time a weekly survey was available to be given. Notifications were also being delivered for reminding the completion of incomplete activities during the period when survey was due.

To analyze data quality, survey data was considered of the participants who completed all 4 weekly surveys. Initially the data sets were analyzed based of the average time spent in the survey per week. Out of the 4 weekly surveys max average time spent on each question was in week 1. Week 2 to week 4 showed a major reduction in time spent per question as compared to week 1. Week 4 surveys displayed the least time spent on each question out of all the 4 weekly surveys. Tb2 provides these details.

Week	**Avg. Time Taken to Complete the Survey (in seconds)**
1	200
2	60
3	46
4	40

Tb2: Weekly Response Time – Study 1

In this study, the participants could be divided into two categories, Android app users and web users. The total time spent in the surveys by each of these two groups was compared. The average time spent for each question by these two users was calculated and there is not much of a difference between the behaviors observed in Android app users as compared to web app. Tb3 provides this data.

No	Participant group	Avg. time spent per question
1	Web application users	4 seconds
2	Android application users	3 seconds

Tb3: Time per Question Analysis Based on Participation Group

Each individual response is associated with a specific response time required to complete the survey and submit it via the app. To analyze the collected data set using response time method (as per section 5.1.1 from Chapter 5) a data set was formed which consisted of response times for each such individual instances. These data points were put together in certain response time ranges measured in seconds against the count of responses falling in that specific range. Figure 3 displays the data distribution observed.

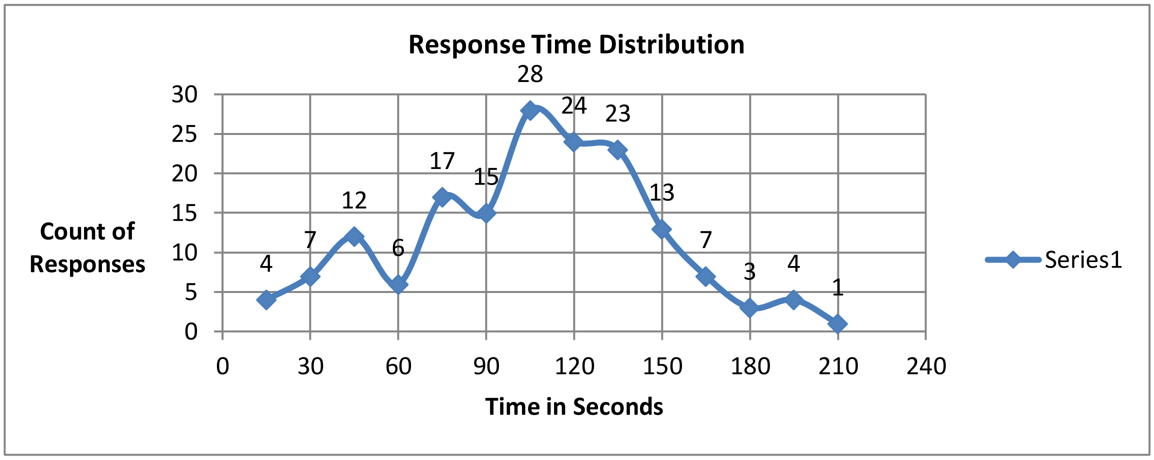


Fig3: Response Time Distribution Curve – Study 1

No explicit data clustering algorithms were used in this study to group the data points into different response time ranges. From mere observation itself, the above data distribution curve provides four distinct response time ranges. These ranges are derived with an intent to maintain real time continuity and avoid any loss of data points observed within the curve. Table 4 provides the details for response time ranges.

Response Time Range (in seconds)	**Group No**
0 to 60	G1
61 to 90	G2
91 to 150	G3
Greater than 151	G4

Tb4. Response Time Ranges Obtained – Study 1

Long string patterns can be calculated by checking the repetition of the same response selected for a sequence of survey questions answered. To analyze the data as per the response pattern method (as per section 5.1.2 from Chapter 5) each response set was monitored for any occurrence of long string patterns. The counts observed against a specific long string index value is shown in the figure 5. If there are multiple long string patterns observed in a single survey response, then each pattern is counted as 1 individual unit.

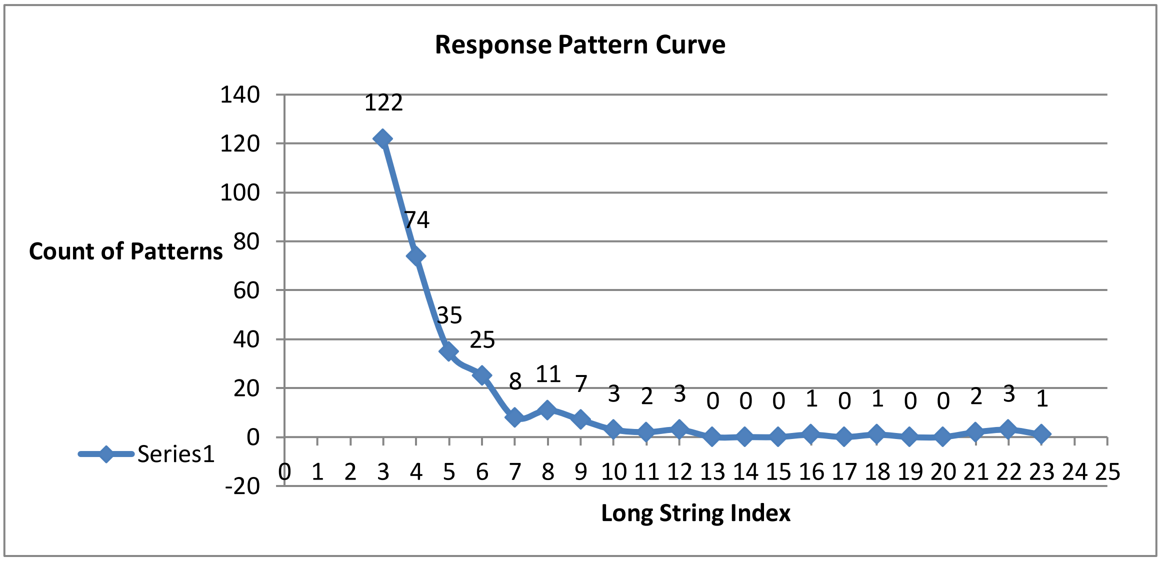


Fig5: Response Pattern Counts – Study 1

Further the response time and pattern data was used to evaluate whether there is any relation between time and pattern length observed in the collected self-reported response set. Figure 6 displays the relation derived between response time and response pattern observed.

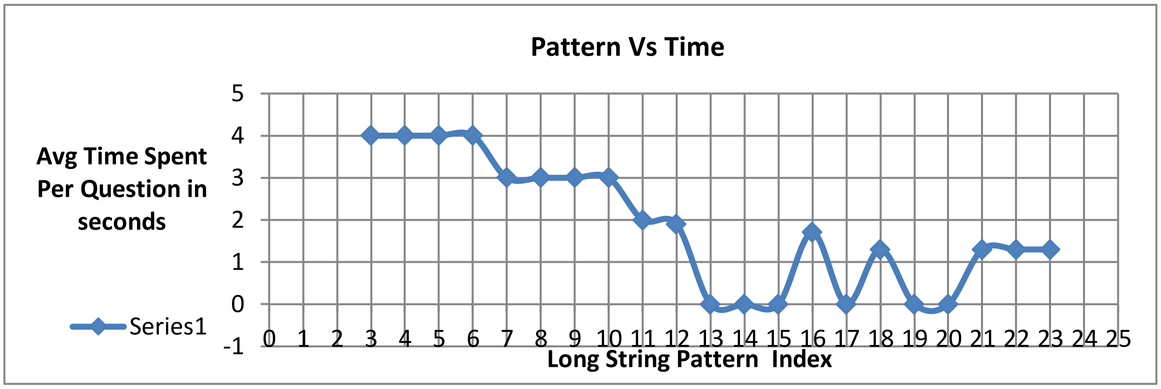


Fig6. Relation between Response Pattern and Response Time – Study1

6.1.3 Interpretations

Table1 and figures 1 and 2 provide a clear picture as to how user participation continued decreasing as weeks passed by. In the last week of experiment, the compliance rate reduced to a low of 63%. In case of a clinical trial, the decrease in user compliance at this rate will impact the clinical outcome that needs to be derived upon analysis of the patient self-reported data collected. As per table 2 and figure3, using response time method four data distribution ranges were found. As per figure 4, maximum responses fall in the range of 90 to 150 seconds and the rest are distributed in either lower ranges or higher ranges. It is noticeable that four different labels of data quality are observed in the data set collected. Hence all the data collected is not of good quality as few of the data points are taking very less time to answer the entire survey and few of them are taking too much time to answer the survey. These results suggest that long and repetitive mHealth surveys do impact the data quality of the responses collected. Similarly using response pattern method, from figure 5 one can see that a significant number of response patterns have been found. From figure 6 a relation between response pattern length and response time can be found. This figure shows that there is an indirect relation between these two entities. As the response time decreases, the length of the response pattern increases and vice-versa.

From the results and interpretations derived as part of this study, one can relate back to the behavior observed by John in section 5.2.1.2 and 5.2.2.2 from Chapter 5. When John spent lesser time to answer the survey then the quality of data collected from him was labelled ‘Bad’. Similarly, when a greater long string index was observed in his survey response then the quality flag was labelled ‘Bad’ for him. Results from this study do show that responses were observed which spent time lesser than 2 seconds to answer the survey. Responses with long string indices greater than 8 are also observed. These numbers suggest that the user compliance and data quality have been impacted due to the long and repetitive nature of the mHealth survey delivered to the participants.

6.1.4 Limitations

The limitations of this study were that out of 62 students only 12 students used mobile application for participation in the study. The participants belonged to a narrow demographic range as the students considered for the study were all from a single coursework. Student participation from other courses offered was not included in this study. The results might have been impacted if there was a variety in the demographics of the participants. The results observed for user compliance were better as compared to web users as the notifications were being delivered reminding the participants regarding the due date of the surveys. This might have been a reason behind better compliance observed in Android users as compared to web. The results observed in data quality measures were not very different from the web user participation.

6.1.5 Conclusions

In conclusion, the results obtained in experimental study 1 support the hypothesis from RQ1 that long and repetitive nature of mHealth surveys does impact the user compliance and data quality observed in the responses collected. There are limitations in the demographics observed or in the number of participants who used the mobile version of the app. However, the results derived clearly suggest that there exists a problem in user compliance and data quality observed in the responses collected from lengthy and repetitive mHealth surveys.

6.2 Study 2

6.2.1 Experimental Context

Study 1 was conducted to validate whether the problem of decline in user compliance and data quality exists in the context of mHealth surveys. The results suggest that the problem exists. However, study 1 was conducted in a controlled environment with a non-patient population. The next thing was to test the same on actual patient data collected from a clinical trial. As mentioned in section 3.4 from Chapter 3, in this research I used PROMIS mHealth application developed for conducting clinical trials on patients with Sickle Cell Disease (SCD) to validate my research questions.

In May 2016, the hospital in collaboration with Arizona State University acquired an IRB approval and started a study on patients with SCD. The clinical aim of this study was to measure and improve user compliance in the patient population undertaking the trial. This is an ongoing clinical trial that expires in May 2017. My aim from this study was to validate the conclusions derived from study 1 on actual SCD patients using an mHealth app. This study was used to check whether user compliance and data quality are impacted when the patients are provided with lengthy and repetitive surveys every week.

This study made use of version 2 of the PROMIS app as described in section 3.4 of Chapter 3 and 4.3 from Chapter 4. This version delivered a set of 25 questions every week to the patient via a mobile application downloaded on their mobile device. These survey questions consisted of 1 SVG question and 4 sets of questions based of PROMIS validated measures for collecting self-reported pain data. The set of questions and the order of delivery are repeated each week. All questions are compulsory and the survey will not proceed if a question is not answered. The survey data is de-identified. Each week the responses to the surveys are recorded and saved in the database. This survey is specifically designed to be taken as a web application that is deployed within a native Android or iOS device. Clinicians determine the user compliance rates observed in comparison to the paper-based surveys conducted earlier in the clinic analyze data collected from these surveys.

Each participant is enrolled in this study for 12 weeks. The total number of participants who enrolled for this clinical study protocol to date is 36 patients. The number of patients who were considered for this experiment was 33 as only these patients have fully completed their 12 weeks in the trial. 3 patients have been excluded from the data analysis, as their 12 weeks have not been completed. These patients are with ages ranging from 12 to 21. There were total 138 weekly responses available for all these patients combined.

6.2.2 Results

For analyzing user compliance, self-reported weekly survey data collected from the 33 patients was used and plotted on a graph. Figure 7 displays the user compliance trend observed for all the patients in the 12-week clinical trial period.

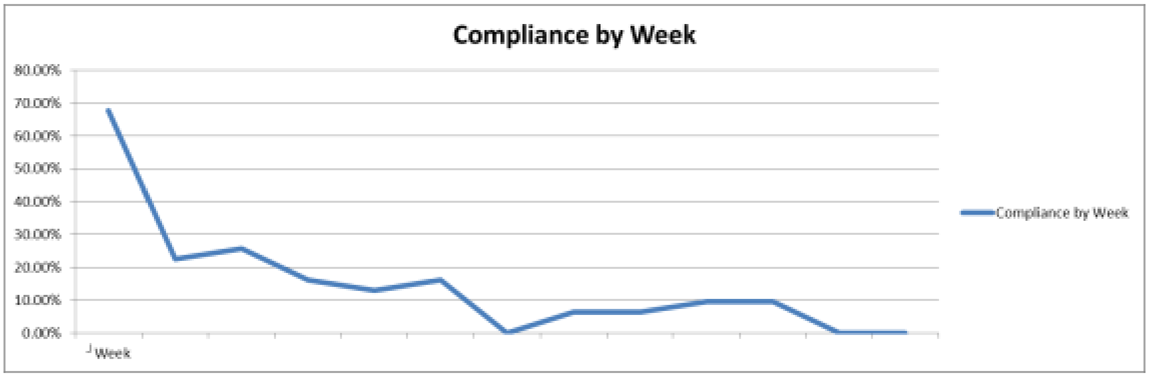


Fig7: User Compliance Trend – Study 2

Response time is calculated for each individual response collected in the entire data set. As per the response time method, all the response time data points are plotted on a graph against the count of responses observed. Figure 8 provides the details regarding the distribution curves obtained as per the response time.

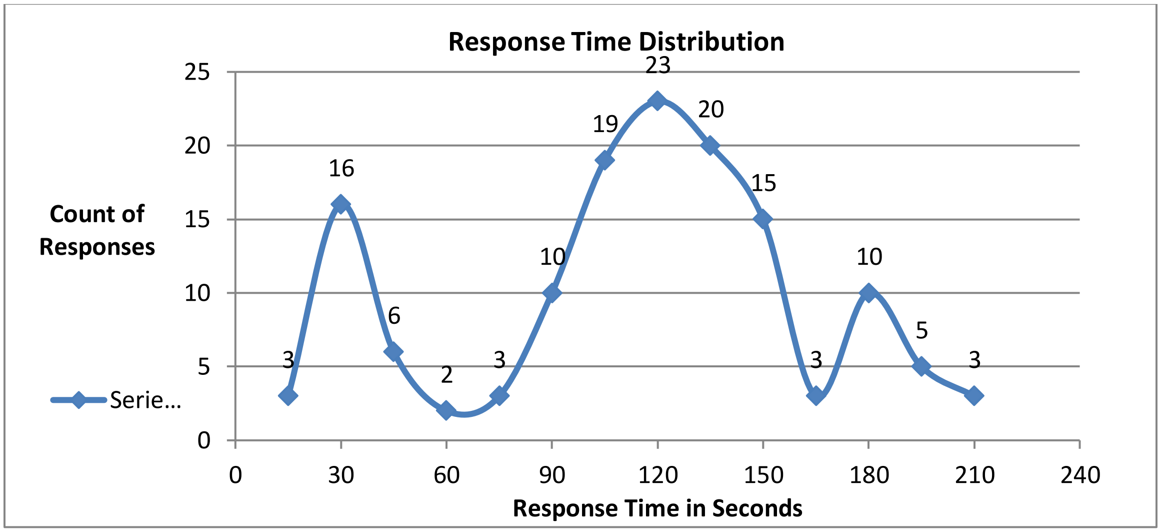


Fig8: Response Time Distribution Saddle Curve – Study 2

Like study 1, no explicit data clustering algorithms were used to group the data points for response time observation. With the help of observation, four data distribution ranges were derived. These ranges make sure that there is no loss of data obtained and each data point can fit into one of the ranges derived. Table 5 gives details about these ranges.

Response Time Range (in seconds)	Group No
0 to 45	G1
46 to 90	G2
91 to 170	G3
Greater than 171	G4

Tb5. Response Time Groups – Study 2

Long string patterns are calculated for this response set. A graph is plotted with long string index value against the count of responses in which the long string occurrence is observed. Figure 10 gives a detailed view of the long string pattern graph.

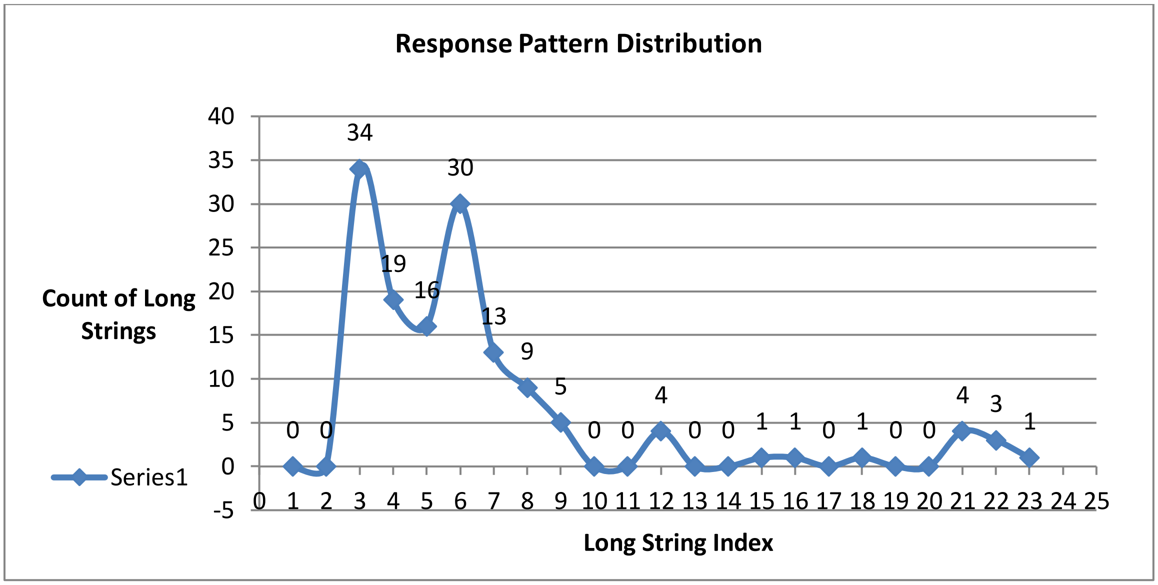


Fig10: Response Pattern Distribution – Study 2

6.2.3 Interpretations

Figure 7 provides a clear view regarding the user compliance observed in this study over a period of 12 weeks. The curve seen in this graph is a typical curve seen in mHealth apps, showing rapidly declining compliance. The graph shows a steady decline in the participation as the weeks pass by. 70 percent compliance is observed in the initial two weeks of the trial and later the trend steadily declines to a minimum of 0% after week 6. Some increase is observed in the later weeks but it reaches a maximum of 10% after week 6, which is very low comparatively. These results suggest that the user compliance is negatively impacted due to the long and repetitive surveys delivered to the patients. This was the motivation behind this research to find out measures to help improve compliance.

Figure 8 and 9 shows the data set distribution using response time and pattern methods. For response time method, a saddle curve is observed with four distribution ranges under it. Maximum number of data points were observed under the range 91 to 170 seconds which is like the range observed in study 1 (section 6.1.3). Total of 133 long string patterns were observed in the entire data set of 138 weekly responses. Out of these 133 there are distinctly two different groups of long string indices observed. One group is with index lesser than 6 and the other with greater than 6. 69 long string patterns were found with index lesser than 6 and 72 patterns with index greater than 6. Almost 50% of the total long strings found were greater than 6 and this number is a serious concern. Long string index value greater than 6 is more of a concern as each PROMIS block consists of 6 questions. If the pattern of same response observed exceeds a PROMIS block then it does indicate that there is a rhythm for tapping on the mobile device being followed by the patient to respond to the survey. These patterns do imply that the quality of self-reported data collected might be low.

These results from study 2 support the results derived from study 1 which help state that the problem regarding user compliance and data quality observed in long, repetitive mHealth surveys exists. This is a serious problem that needs attention, as low compliance and bad quality data points do not help derive accurate conclusions from the clinical studies conducted. This jeopardizes the clinical objective itself for which the mHealth apps were developed.

6.2.4 Limitations

The limitations of this study are like study 1 that the total population for obtaining the weekly response data set is small. Amongst the total participants, the patients who regularly responded to weekly surveys are very small and thus the size of the data set for analysis is less. A better participation rate would have provided with a larger data set for better classification of data points into various groups as per their response time and pattern.

6.2.5 Conclusions

This study was conducted to validate the conclusions derived from study 1 on actual patient data in a clinical context. These experiments validate my first research question (RQ1) as part of this thesis. These results do suggest that the long and repetitive nature of mHealth surveys impact the user compliance rates and quality of data obtained from the self-reported survey activities.

6.3 Study 3

6.3.1 Experimental Context

Results from study 1 and study 2 helped in validating the first research question. Response time and response pattern methods were used in these studies to observe the data distribution and to check if there were any similarities in the results obtained. Four different response time distribution ranges were derived from these results and various counts of long string patterns were also found. However, as stated in sections 3.2.2.3 and 3.2.2.4 from Chapter 3, a specific value for response time or response pattern can be considered as good or bad depending upon the nature of the surveys designed. As per my knowledge, existing literature did not provide any baseline values for response time and patterns observed in the mHealth apps. Thus, there was a need to derive baseline or ground-truth values for both response time and pattern using the PROMIS SCD app itself. To do that, a pilot study was conducted in November 2016 with senior level undergraduate and graduate students enrolled in the software engineering course SER 421. This study was conducted under an approved IRB in a controlled environment. The aim of this study was purely to derive baseline values for the response time and response pattern methods that are further used to validate the second research question using PROMIS SCD mHealth app.

Each student was provided with four-digit personal identification numbers, which identify their individual responses from the others. Every student was asked to take the survey twice. First time they were asked to respond to the survey in the fastest manner possible; without paying any attention to the questions asked by selecting a response option as quickly as possible. The second time the student was asked to respond after reading each question carefully and spend enough time understanding the question asked and select suitable response option accordingly.

By this time, the version 3a of the PROMIS SCD app was developed and ready for use. As mentioned in section 4.3 from Chapter 4, this version of the app had few major changes as compared to the previous one used in study 1 and study 2. The total number of questions delivered as part of weekly surveys was now 31. This comprised of multiple-choice single answer as well as multiple-choice multiple answer questions. This change made the survey even lengthier as few medication adherence and adaptive questions were also included along with PROMIS questions in this version. The SVG question, which was part of the previous version, was removed in this version. The order in which the questions were delivered was randomized for all 4 validated PROMIS blocks as well as within each block. This ensured that no two surveys would have the same question order. All questions were compulsory and the survey would not proceed if a question was not answered. The surveys were taken in a de-identified manner to avoid any bias during data analysis and to protect student confidentiality. The recorded data included information about the answer option chosen for each question, time spent on each question as well as if the question was revisited.

Total number of students who participated in this study was 18. The response data from all 18 students was included to obtain results for both response time and pattern. This study was conducted in one sitting and was completed within 1 hour. A pizza was provided to each student in the form of a reward for participating in the study.

6.3.2 Results

As stated above, the entire study consisted of two response sets provided by each participant. These can be labeled as response set 1 (RS1) and response set 2 (RS2). RS1 consisted of all the responses collected when the participants were asked to respond as quickly as possible without paying much attention to the content of the survey. RS2 consisted of all the responses collected when the participants were asked to respond after carefully reading and understanding the content of the survey. As there were 18 participants and each participant had to provide a response in RS1 and then in RS2, total numbers of data points in each of the response set are the same, which is 18 each. The entire set of responses collected from both the sets were collected together and plotted against a graph the below two distribution curves were observed.

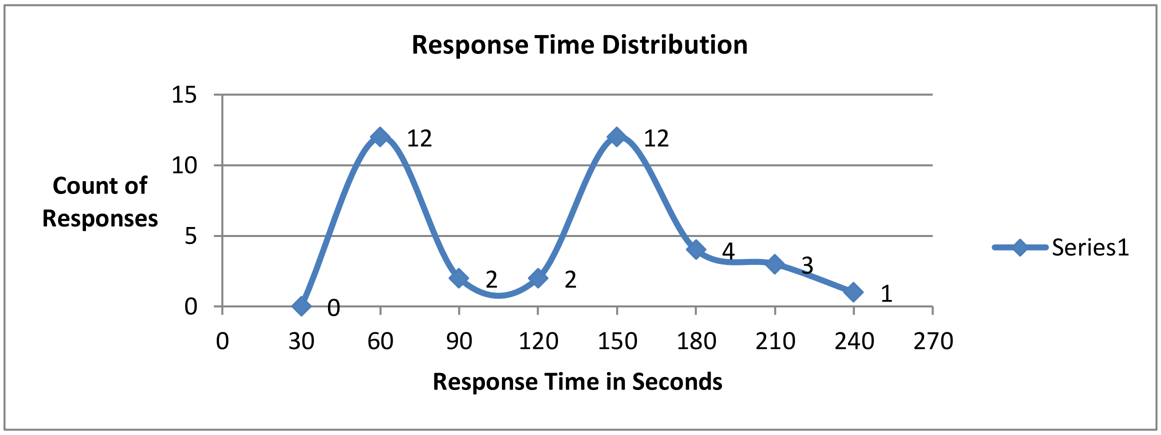


Fig11: Response Time Distribution – Study 3

From Fig 11, it can be observed that majority of the data points have been distributed under the two humps of the saddle curve. H1 depicts the data points in the range from response time 30 seconds to 90 seconds. H2 depicts all the data points in the range from response time 121 seconds to 180 seconds. For the total 36 responses collected from each of the RS1 and RS2, the data points' distribution under these two curves is as follows.

Response Set Type	H1	H2	Outside two humps
RS1	14	4	0
Rs2	0	12	6

Tb6: Data points distribution under the saddle curve – Study 3

It can be observed that 14 out of 18 data point from RS1 fall under H1 and the 4 data points fall under H2 in the saddle curve. Similarly, for RS2, 0 data points fall under H1 and 12 data points fall under H2. Remaining 6 data points of RS2 do not fall under any of the two humps observed.

After obtaining results for response time, the next thing to look at was response pattern. Long string pattern index value can be defined as the count of same response options provided for a consecutive series of questions in a single response collected. For example: If the response pattern observed for question 1 to question 6 is as "a,a,a,a,a,a" then this will be considered as a response pattern with long string = 6. In this study two response sets were collected (RS1 and RS2) with 18 individual responses in each set. These 36 responses were further analyzed to observe any long string pattern behavior in them. The below figure describes the long string patterns observed in RS1.

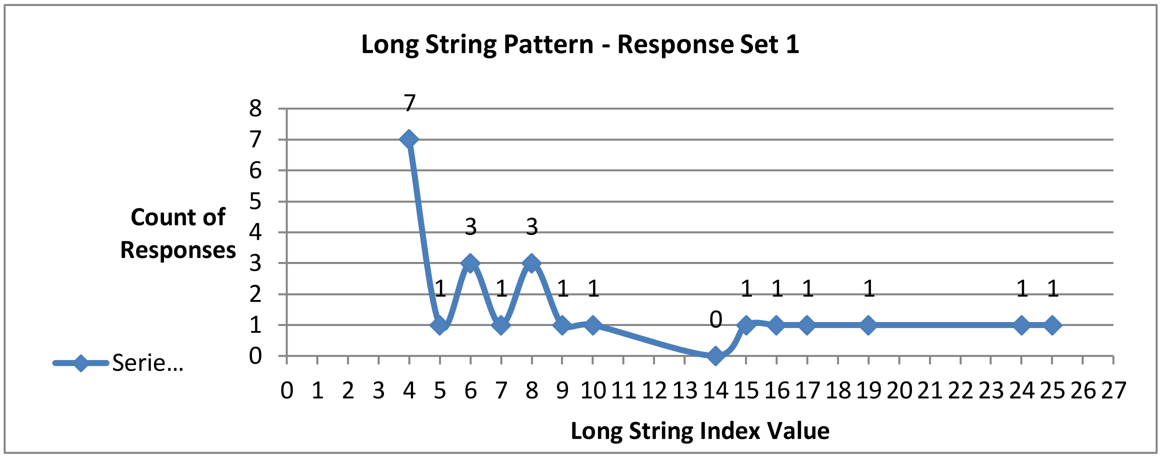


Fig12: Long String Patterns from Response Set 1 (RS1)

A total of 23 long string patterns were observed from the RS1 data points. From Fig 12 maximum number of long string patterns with index value 4 have been observed and the next maximum is with index value 6. It is also observed that the number of long string patterns observed with index value greater than 6 in RS1 is equal to 12. Below table describes this data distribution properly. Similarly, the long string patterns were calculated from RS2 as well. Fig 13 shows the long string pattern count distribution observed in RS2.

Long String Pattern Index Range	Number of Data Points
4 to 6	11
7 to 10	6
14 and above	6

Tb7: Long String Pattern Count from Response Set 1 – Study 3

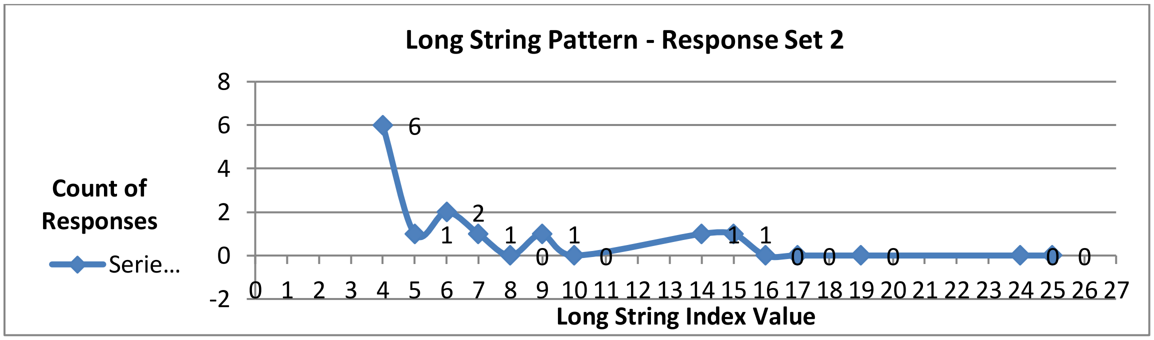


Fig13: Long String Patterns from Response Set 2 (RS2)

Total 13 long string patterns were observed from the RS2 data points. From Fig 13 it is observed that maximum number of long string patterns with index value 4 have been observed followed by long string pattern 8 and 6 respectively. The long string patterns with index value greater than 6 in RS2 are equal to 4. Below table describes this data distribution appropriately.

Long String Pattern Index Range	**Number of Data Points**
4 to 6	9
7 to 10	2
14 and above	2

Tb8: Long String Pattern Count from Response Set 2 – Study 3

Once the long string patterns have been derived, I went one step further to analyze how many responses which had these long string patterns fall under the two humps H1 and H2 observed in the saddle curve distribution of data points for response time. Each response string might contain more than one long string pattern observed in it. Each pattern is counted as an independent unit. Below table describes the distribution observed.

Long String Pattern Value	**Count in H1**	**Count in H2**
4 to 6	6	14
7 to 10	6	2
14 and above	6	2

Tb9: Long String Pattern Distribution in Response Time Curves – Study 3

6.3.3 Interpretations

After observing the results obtained from response time saddle curve in figure 11 and table 6, one can see that when the participants are asked to respond as quickly as possible without paying much attention to the survey content, then the response time taken for the entire response would fall in the range of 30 seconds to 90 seconds (H1). Similarly, when the participants are asked to respond after carefully reading and understanding the survey content, the response time taken for the entire response would fall in the range of 120 to 180 seconds (H2).

From these observations, it can be concluded that there would always be two curves with data points distributed under them whenever there is a collection of responses provided in a faster and slower manner. Thus, these two curves can be labeled as ‘Fast’ and ‘Slow’ collection of data points with respect to response time as a judging factor. There are few data points which go beyond the H2 range. Overall after observing the data distribution curves from figure 11, four labels of data distribution using response time method are described. These are described in the table below.

Response Time Range (seconds)	**Data Labels**
30 to 90	Too Fast
91 to 120	Fast
121 to 180	Normal
Greater than 180	Slow

Tb10. Data Labels as per Response Time Range – Study 3

From fig 12 and 13 it can be observed that the long string patterns have been found in both the RS1 data points as well as RS2 data points. In RS1 I found 17 long string patterns and in RS2 I found 18. This shows that a similar count of long string patterns has been found irrespective of whether the participants were responding way quicker or much slower in the survey. The data distribution of responses with long string patterns greater than 7 under H1 and long string patterns lower than 7 under H2 do indicate that the data points can be easily distributed in two groups. Group 1 would be with long string indices less than or equal to 6 and the group2 would be with long string indices greater than 6.

I hypothesize that long string patterns are inversely proportional to the response time spent in the survey. The lesser the response time spent in the survey greater would be the long string index value found in the response if any. From Tb9 one can see that the response data points with long string pattern index value greater than 14 are found majorly under Curve1 of response time. The response data points with long string pattern index value in the range of 7 to 10 are more under H1 of response time than H2. On a similar note, I also found that data points with long string pattern index value in the range of 4 to 6 were found majorly in H2 of response time distribution saddle curve. This suggests that our hypothesis is correct. If a long string pattern is observed with higher index value then the response time spent in the total survey would be lesser and vice versa.

6.3.4 Limitations

Out of the total 36 responses collected from the 18 students participating in the survey, there were 4 students who did not follow the instructions provided properly at the start of the experiment. The instructions were each student is first expected to respond to the survey as quickly as possible and then in the second turn respond to the survey after careful reading and understanding. However, few students did vice-versa of the expected. All such responses were checked and an alternate consideration was used to handle such cases. All the responses from these students' RS1 were taken with the other students' RS2 and vice-versa. This way I ensured that the data points being considered for the analysis are uniform in nature. The population belonged to the older age group as compared to the clinical trials. Demographic age range observed in SCD clinical trials are in the range of 12 to 21 or 8 to 21.

6.3.5 Conclusions

The two response sets, RS1 and RS2 have been analyzed against response time method and long string pattern method. The observations for response time method gave us two humps under a saddle curve for distributing the data points collected. I labeled these humps as H1 with time spent in the range of 30 to 90 seconds and H2 with time spent in the range of 121 to 180 seconds. With the help of these curves I could derive four labels as ‘Too Fast’, ‘Fast', ‘Normal' and ‘Slow'. These labels would help identify whether the effort taken by the participant is enough to qualify the response set as good data or bad data. Thus, it can be concluded that the response time method on its own is sufficient to label a response data set into good and bad responses.

The long string pattern method mentions that the longer the pattern index value, greater are the chances that the effort taken by the participant is less in responding to the survey thus hampering the data quality of the response. However, I saw that the long string pattern method is a weaker method on its own, as an equal distribution of data points were found from both RS1 and RS2. Later I hypothesized that there is an inverse relation between long string patterns found and the response time taken to complete the survey. I found that longer the string pattern found, lesser would be the time taken by the participant to complete that specific survey instance. This relation has been observed in study 1 and study 2 as well. Hence this relation can be considered affirmatively which will help derive better results from the clinical trial.

The aim of this experiment was to derive baseline or ground-truth values for response time method and response pattern or long string pattern method. From the above observations, one can conclude that the baseline value for response time method would be in the range of 121 to 180 seconds and baseline value for the response pattern method would long string with index greater than 6.

- Baseline value for Response Time:

  between 121 to 180 seconds for a set of 31 questions

  or

  4 to 6 seconds on average

- Baseline value for Long String Pattern > 6

These values are the key points where the behavior of the data distribution is distinctly observed in this study. These baseline values would further be used in study 4 for analyzing the data collected in study 4 against the algorithms of response time and pattern using mixed methods.

6.4 Study 4

6.4.1 Experimental Context

Study 3 helped me derive the baseline values for the mixed methods approach of using response time and pattern algorithms. The next aim was to validate the second research question of this thesis. RQ2 is to evaluate whether an intervention based of intrinsic and extrinsic motivational factors implemented in mHealth platform help improve compliance and data quality. By this time, as per the timeline provided in section 3.4 from Chapter 3, version 3b of the PROMIS app was developed and ready to use in the clinical trials. Our clinical partners mentioned as per section 3.4 from Chapter 3 started a usability study in December 2016 at the hospital. This study was organized in collaboration with ASU under an approved IRB. This study would expire on December 2017. The clinical aim of this study is to measure the hospital readmission rates of SCD patients undergoing this trial. The clinicians want to measure the reasons behind the increase in hospital readmission rates and find out measures to reduce this number. My research aim from this usability study is to check whether an intervention of intrinsic and extrinsic motivations embedded in an mHealth app helps to improve the compliance and data quality observed. Hence version 3a and version 3b of the PROMIS app were used in this study.

Each patient is enrolled for 5 weeks in this study. Version 3a and 3b of the PROMIS app is being used to conduct this experiment. As mentioned in section 4.3 from

Chapter 4, version 3a and 3b had many changes incorporated in them as compared to the previous versions of the PROMIS app. Along with weekly survey activities, daily survey activities got introduced as part of this study. A daily survey would consist of 3 to 4 multiple-choice multiple answer questions. These questions majorly focused upon gathering more information regarding the pain medication consumed by the patient daily. These questions also help the clinicians to monitor the pain medication dosage being taken by the patient and whether it matches with the prescribed medication or not. Earlier version was delivering an SVG question asking for body pain and intensity as part of weekly survey. In the current version, this question was placed in daily instead of weekly survey activity. The length of daily survey would be short compared to weekly surveys where the number of questions increased from 25 to 31 which included addition of medication adherence as well as adaptive questions getting delivered along with 4 PROMIS validated blocks of questions.

In version 3a, along with the change in the survey length; there were changes in the notifications being delivered to the patient as well. As per section 4.3 from Chapter 4, there were certain issues with the notifications being delivered in the app on iOS as well as Android devices. These were fixed as part of version 3a. Enhanced clickstream logging and user app life cycle activity logging was introduced. These helped to log every user interaction within the app right from the time when the app gets installed and activated by the patient, to attending and submitting survey activities; till the app termination (if applicable) on the patient's device. These logs play a major role while analyzing the survey response data using the mixed methods and baseline values derived from study 3 for response time and pattern.

Version 3b was the next immediate minor version of the PROMIS app delivered. This version included all the major changes from 3a along with inclusion of intrinsic and extrinsic motivational factors. As mentioned in section 4.3 from Chapter 4, intrinsic motivations include short informational messages regarding the SCD are being delivered via the app. As part of extrinsic motivations / rewards badges and games were introduced in this version. Two games: Pappu Pakia and Squirts are embedded in the app allowing the patient to have some game play any time in clinical trial duration. All these enhancements are exactly the interventions I require for evaluating my second research question.

In this study, each patient is provided with four-digit personal identification numbers to identify their responses from other patients. The patient gets enrolled in the trial for 5 weeks. In this duration, he receives 6 weekly and 36 daily surveys. Both versions 3a and 3b have been used in this study. Few patients are provided with version 3a and few are provided with 3b. This distribution of multiple versions of the app helped to compare the difference in behavior observed by various patients using the app.

The total number of patients who participated in this study till date are 18. All 18 patients are being used to monitor weekly and daily compliance over the 5 weeks' trial period. However only 6 patients' data is available for monitoring data quality and for analysis using the mixed methods from Chapter 5. This inclusion criteria were derived taking into consideration a bug which was found after the release of version 3a. The require user interaction logging was not getting saved to the DB and thus no logging data is available for the 12 patients excluded from data quality analysis. From the 6 patients under consideration, 4 received version 3a of the PROMIS app (non-enhanced) and remaining 2 received version 3b (enhanced) app.

6.4.2 Results for User Compliance

Measuring user compliance of self-reported response set means counting the number of responses collected per week from the patients who have been enrolled into the clinical trial. In this specific study both weekly and daily surveys were delivered for both sets of patients, the ones using non-enhanced as well as enhanced version of the PROMIS app. User compliance is observed with respect to weekly as well as daily surveys. For measuring user compliance data from all 18 patients are taken into consideration.

Fig14 and Fig15 below provide an idea regarding the trend observed in user compliance rate on a week-by-week basis. As mentioned earlier each patient received 6 weekly surveys as part of this clinical trial.

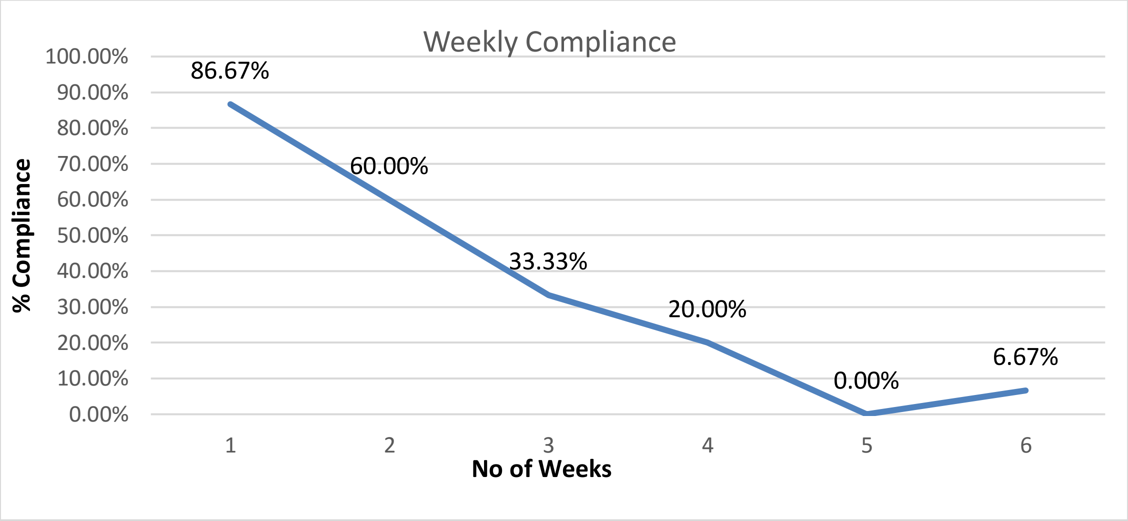


Fig14: Weekly Compliance Chart

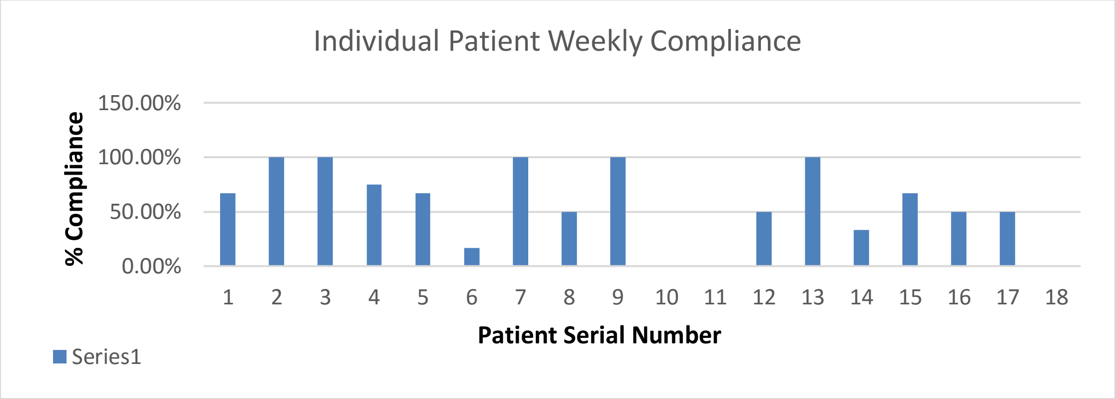


Fig15: Individual Patient Weekly Compliance Chart

Like the weekly compliance charts, I have computed the daily compliance charts as well for each individual patient based upon their participation in the daily surveys delivered via the PROMIS app. As mentioned earlier each patient received 36 daily surveys as part of this clinical trial.

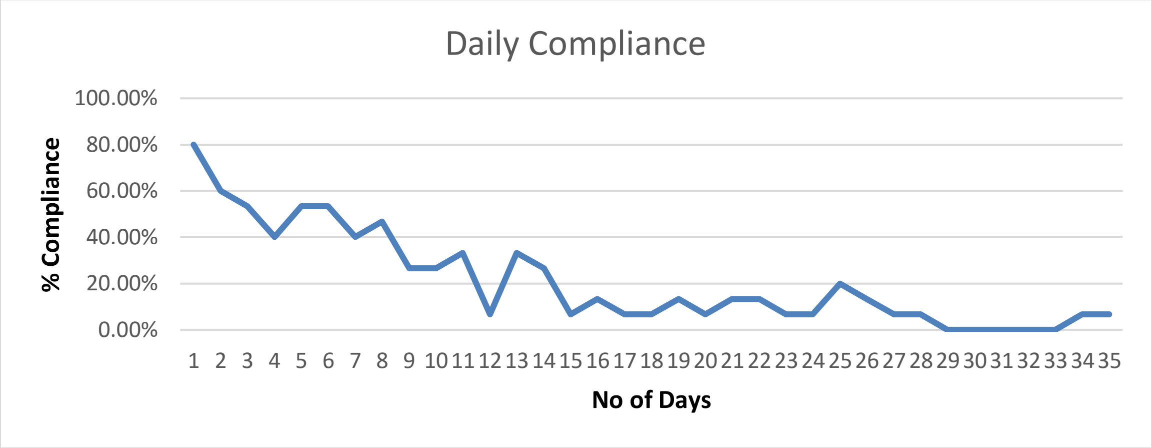


Fig16: Daily Compliance Chart

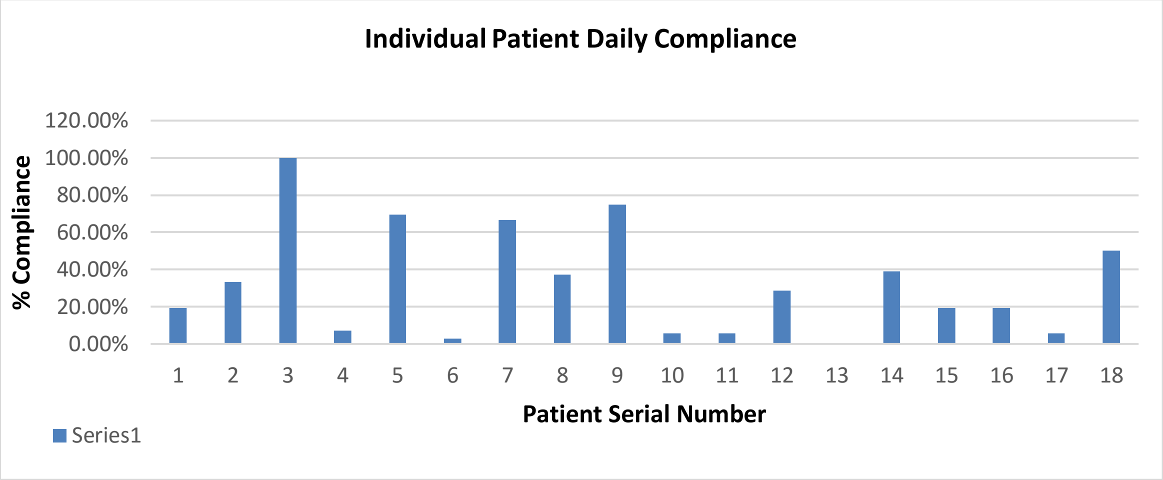


Fig17: Individual Patient Daily Compliance Chart

Out of these 18 patients in the clinical study, patient number '9' and '15' are the ones with enhanced content delivered via the app in the form of informational messages as well as badges and games embedded in the PROMIS app. Below table highlights their weekly and daily compliance rates.

Patient Serial Number	Weekly Compliance %	Daily Compliance %
9	100	75.00
15	66.67	19.44

Tb11: Weekly and Daily Compliance Observation for Enhanced Content Patients

6.4.3 Interpretations for User Compliance

Looking at figure 16 and 17, one can easily say that the user compliance trend has significantly improved as compared to pilot study 1 and 2. In a 5-week trial, now the minimum compliance is attained at week 4 as compared to the earlier clinical trials where the compliance trend reached to this minimum level right in the second week itself. We do see a steep curve from week 0 to week 3, but further down in week 3 and week 4 some consistency is observed in the compliance rates displayed. The changes that took place in this clinical trial as compared to the previous ones was that I introduced notifications for the incomplete as well as new surveys being delivered by the app to the patient. These notifications might have been helpful in continuously reminding the patient about the survey activity which needs to be completed before it gets expired. These notifications might have also helped the patients to get the app in the foreground whenever he is notified regarding the survey.

Notifications would be delivered to the patient only when he uses the Android or iOS device to download the app from app stores and complete their surveys using the app. Out of these 18 patients, 72.22% participants were using either iOS or Android smart device to download the app and give their surveys. No explicit information about device usage is mentioned for the remaining 27.77% of patients. 72.22% participants were continuously reminded about their upcoming and pending surveys which helped them to keep track of their activities and complete them before they expire. Thus, notifications did play a major role in improving user compliance rates in this clinical trial as compared to the earlier ones.

Let us observe the two-enhanced content patient’s performance with relation to compliance in both weekly and daily surveys. Patient with serial number ‘9’ has shown far better performance in both weekly and daily surveys as compared to patient ‘15’. There are few points where these two patients stand out in their behavior.

Patient ‘9’ has been very active in all the 3 weeks he was enrolled in the trial. After completion of 3 weeks the patient was readmitted to the hospital and thus the patient was immediately deactivated from the system. In these 3 weeks, the patient completed all 3 weekly surveys as well as 80% of the daily surveys provided to him. As mentioned in Chapter 4, after completion of certain weekly and daily activities badges are awarded to the patient. This patient received in total of 13 badges (both weekly and daily) combined. The patient activity log data also shows that all the badges except for the last weekly badge have been activated to receive powerups which could be further used in the game center while playing the two games embedded in the app. This shows that the patient did visit the Badges dashboard page each time he received a badge and paid attention to his achievements. This patient was using the PROMIS app on an iOS device. Thus, he also received regular notifications for the surveys he was supposed to respond to. All these factors point towards one conclusion that notifications and badges did motivate this user to perform better in both weekly and daily surveys.

Even though this patient activated all the badges he achieved as part of survey activity completion, the patient did not visit the game center even once in the 2 weeks of clinical trial enrollment. This is strange and thus it cannot be said that games acted as a motivational factor for this patient’s good compliance rates observed.

Patient ‘15’ has been active in the clinical trial for all 5 weeks. However, his compliance rates are less as compared to patient ‘9’ in both weekly and daily surveys. One reason could be that this patient was not using the PROMIS app on either of the iOS or Android smart phone devices. There are chances that this patient was using the web version of the app and thus there was no chance of any notifications being delivered to this patient. This can be verified from patient log data where no log has been created for issuing a notification even once. Lack of notifications and reminders might be a reason for the decline in the compliance rates. Regarding badges, this patient completed lesser number of survey activities and thus received only 9 badges (both weekly and daily) combined. Out of these 9 none of the badges were activated for powerups. However, this patient did visit the game center once and played both the games for in total of 20 minutes. This happened during the patient's second week in the trial. Thus, we can say that even though notifications were not getting delivered the patient did get attracted to the games and attempted to play them at least once in the entire trial period. This helps us to say that badges did play a major role in attracting patients' attention towards the app as compared to the games. However, games were also not completely ignored.

With the help of the above observations one can say that introduction of badges and games as a form of motivational factor helped to achieve better compliance rates among patients using mHealth app.

6.4.4 Results for Data Quality

Measuring data quality of self-reported response means to understand the reliability of data collected. If the data is of good quality then it can be used further to derive certain conclusions regarding treatment measures for the patients. If the data is of poor quality then the results derived from it would not be correct and this would ultimately hinder the clinical protocol. To monitor the data quality of self-reported mHealth responses, I will be using the methods discussed in Chapter 5. There are two methods: response time and long string pattern. Three algorithms have been derived using these methods to classify the data into various categories of data quality. For monitoring data quality, only weekly responses are taken into consideration. This is because the weekly surveys comprise of 31 questions delivered in the app page by page, which brings it in the required (long and repetitive) format for this thesis research. Daily surveys being short in length are excluded from the data quality analysis. As mentioned in section 6.4.1 for monitoring data quality response data from 6 patients is considered. These responses are monitored using the response time and long string pattern methods against the 3 algorithms derived as part of this thesis along with the baselines values derived from study 3.

6.4.4.1 Response Time Method (Algorithm 1)

As stated in the experimental context, the clinical trial population was split up into two groups: one with enhanced version and the other with non-enhanced version of the app. These groups can be labeled as NEN and EN. NEN consists of all the patients who received non-enhanced version and EN consists of patients who received enhanced version of the app respectively. NEN has 4 patients and EN has 2 patients. The entire set of responses collected from both the groups were put together and plotted against a graph below which shows the data distribution as per the response time method.

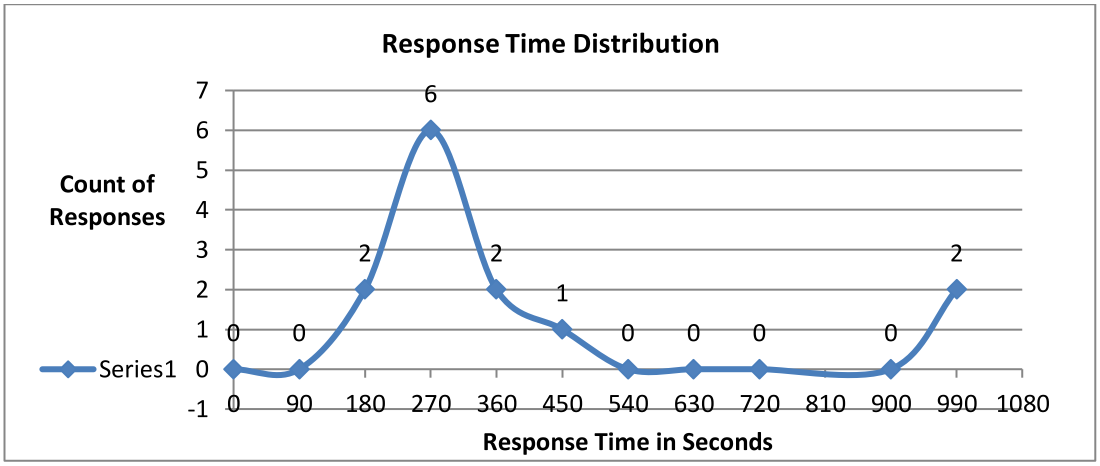


Fig18: Response Time Distribution

6.4.4.2 Long String Pattern Method (Algorithm 2)

Both groups; NEN and EN responses were taken and monitored for any long string patterns observed in the response set. Below figure provides us with the data distribution observed over the entire set of responses from both the groups using response pattern method.

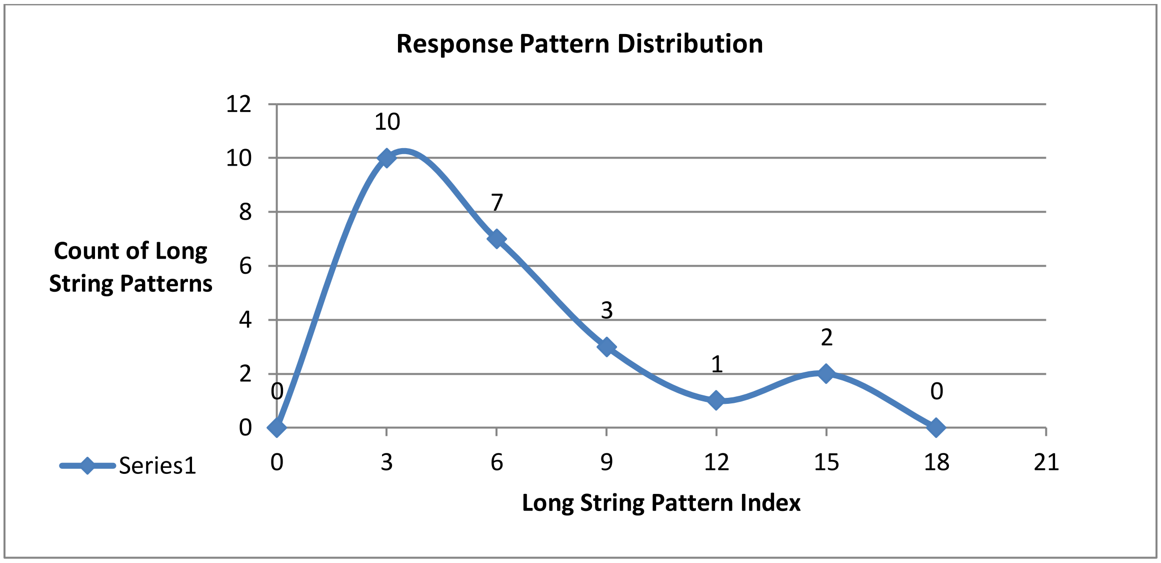


Fig19: Response Pattern Data Distribution

The above figure provides an overall count of responses, which are found with certain long string pattern index values. The internal distribution of these responses for non-enhanced and non-enhanced versions is provided below.

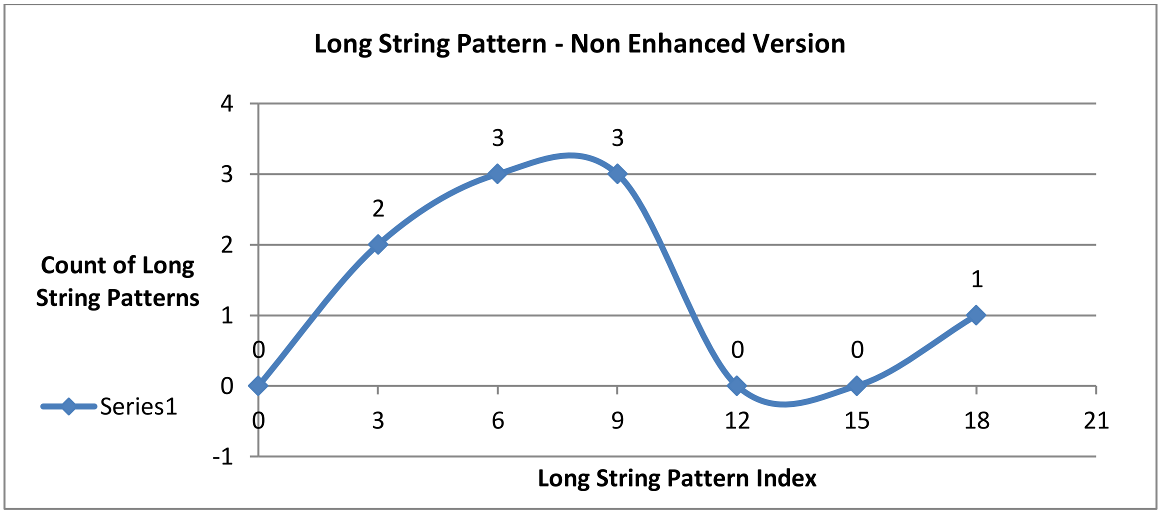


Fig20: Long String Pattern Distribution in Non-Enhanced Version Response Set

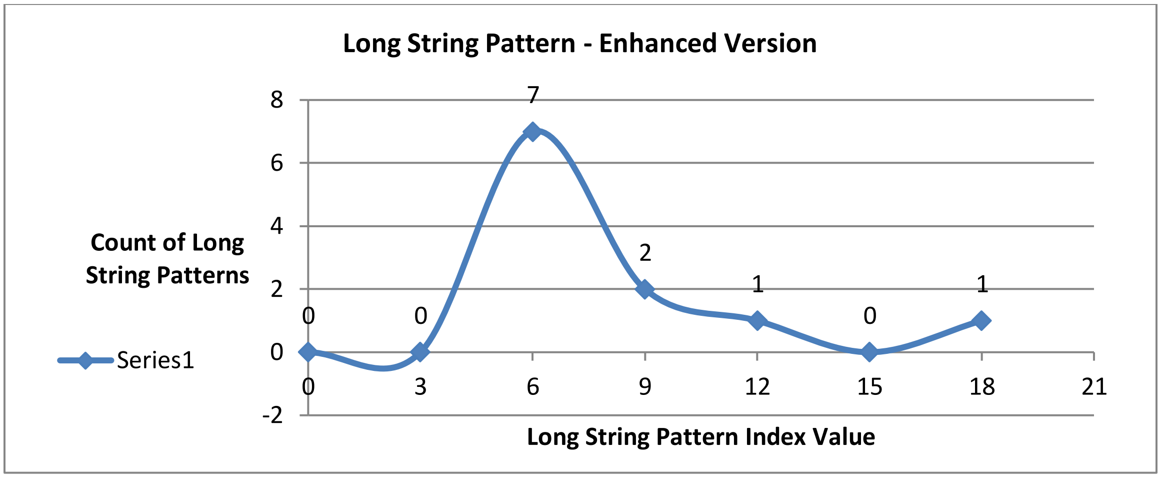


Fig21: Long String Pattern Distribution in Enhanced Version Response Set

6.4.5 Interpretations for Data Quality

6.4.5.1 For Response Time Method

From the figure 18, one can see that majority of the data points are under the distribution of 180 seconds to 360 seconds. The data points exceeding 360 seconds have taken more time as per the rest of the group that is under the curve. As per the baseline values and labels derived from study 3 in section 6.2.3, the normal range for good quality of response data is 121 to 180 seconds. Beyond 180 seconds study 3 determines the data point as 'Slow'.

However, baseline values in study 3 were derived in a controlled environment. The participants were provided with the PROMIS SCD mHealth survey questions. These questions were not significant for the participants to be able to read and provide the best suitable answer. The participants were mainly reading the question and simply selecting an answer option from the available ones. The participants were senior graduate students who mainly belong in the higher end of the population age range with respect to study 4. Age matters for being able to read and cognitively understand the question and provide an appropriate response. This study deals with pediatric patients ranging from age 8 to 21. Due to the de-identified data collected from the patients, one cannot explicitly assign an age against a response collected. The shape of the curve formed in this graph holds with the baseline, but the results must be scaled due to the above-mentioned confounding factors. Thus, I need to assume that there might be an 8-year-old patient who is trying to read and understand the questions and thus might need more time as compared to an adult reading the same. Taking these factors into consideration for response time analysis in this study, I decided to extend the baseline value of 'Normal' response time to be from 181 to 360 seconds. Above 360 will be considered 'Slow' and 30 to 90 as 'Too Fast' and 91 to 180 as 'Fast'. Below table provides an updated baseline chart being considered for this study.

Response Time Range (seconds)	**Data Labels**
30 to 90	Too Fast
91 to 180	Fast
181 to 360	Normal
Greater than 360	Slow

Tb12. Improved Baseline Values for Study 4

Taking the baseline values from table 11 into consideration, appropriate data labels were assigned to the data points from figure 18. The response time distribution for enhanced and non-enhanced version is shown below.

Response Time (in seconds)	**Data Label Assigned**
247	Normal
274	Normal
317	Normal
339	Normal
342	Normal
438	Slow

Tb13. Data Labels for Responses Collected from Non-Enhanced Version

Response Time (in seconds)	Data Label Assigned
216	Normal
270	Normal
271	Normal
511	Slow
1650	Slow
2800	Slow

Tb14. Data Labels for Responses Collected from Enhanced Version

As per the tables 13 and 14, we have achieved data points in the 'Normal' and 'Slow' range. No data point was grouped into 'Too Fast' or 'Fast' responses. For NEN response set, 83.33% of the total were recorded as 'Normal' responses or 'Good' quality responses. For EN response set, 50% responses are recorded as 'Normal' or 'Good' quality responses. By using the response time algorithm, I could identify good and bad quality responses. This data analysis would help the clinicians to determine which data points must be considered or excluded by their statisticians to derive clinical conclusions.

6.4.5.2 For Response Pattern Method

After observing the graphs from figures 19, 20 and 21 the data points I assigned labels to them using the baseline values derived from study 3 for response pattern method. The baseline value derived for long string pattern method was string with index greater than value=6.

It is observed that long string patterns have been found starting with minimum index value as '4' and maximum index value as '18'. 4 long string patterns are observed with index value greater than the baseline value '6' in non-enhanced version and 4 long string patterns are observed with index value greater than '6' in enhanced version response sets. Later I took the total number of responses in non-enhanced and enhanced to monitor how many of these do have long string patterns with index greater than threshold value to label them as 'Good' or 'Bad' quality data.

Response Set	**Good**	**Bad**
NEN	5	4
EN	7	4

Tb15: Data Labels Using Response Pattern – Study 4

From table 15, we can see that 5 out of 9 data points are labelled as 'Good' for non-enhanced response set and 7 out of 11 data points are labelled as 'Good' for enhanced response set. This method too helped in identifying the good from the bad response sets thus providing a distinct idea of which responses should be ignored by the clinicians to derive their clinical outcomes from.

After looking at the plotted graphs and data statistics for long string pattern method, some difference in the data quality is observed between responses collected in two different versions. The enhanced response set did give better quality data as compared to the non-enhanced version response set. This method did show significant difference towards improvement of data quality measure in the collected response set using enhanced version of the app.

6.4.5.3 Mixed Methods of Response Time and Pattern (Algorithm 3)

As part of this algorithm, I am trying to combine the positives of both the methods; response time and long string pattern to derive better conclusions regarding the data quality factor. As discussed in pilot study 3, response time independently can be used as a judging factor for measuring quality of data. Long string pattern alone is not sufficient for deriving conclusions towards the quality of data obtained. If combined with response time then it will produce better results rather than using pattern alone to derive a conclusion regarding data quality. This method tries to exploit this specific relation between long string pattern and response time spent in the block where a pattern is observed.

As already checked in response time method, I initially checked for the response sets from both groups NEN and EN to check whether they meet the response time criteria or not. Once it is concluded that none of the responses met the ideal condition, I went one step ahead and checked whether long string patterns were found in these responses. As discussed in the earlier section, long string patterns were observed in responses from both the groups, NEN & EN. Now as part of mixed methods algorithm, I will go one step further and check the response time for the block in which a specific pattern is found. If the response time of that block is greater than the baseline value of response time per question then the quality of data is considered good else bad.

As per the analysis of pilot study 3, moderate quality of data is achieved when the time spent on an average for each question is between ***4 & 6 seconds***. However, as part of response time analysis in this study we improvised these baseline values to be between 180 and 360 seconds for 'Good' quality data (section 6.4.5.1). If the total number of questions in the weekly survey is 31, then the new average time spent per question in the survey would be between **6 & 11 seconds**. This would be used as the improvised version of the baseline value derived for the combination of response time and pattern method.

Taking this baseline value into consideration, I derived the average time spent per question in the long string pattern block using the below formula:

$$\text{Avg. Response Time Per Question} = \frac{\text{Total Response Time for Pattern Block}}{Long\ String\ Pattern\ Index\ Value}$$

Using this formula, I calculated the average time per question from the two sets of responses NEN and EN in which long string pattern above baseline value was found. Below tables provide a brief idea about the counts of responses and the label assigned to them as per the mixed method algorithm.

Average Time Per Question in Pattern Block (in seconds)	**Data Label**
11.45	Good
4.78	Bad
14	Bad
10.72	Good
6.12	Good
5.37	Bad

Tb16. Data Labeling of Non-Enhanced Version Response Set using Mixed Method

Average Time Per Question in Pattern Block (in seconds)	**Data Label**
4	Bad
7.50	Good
6.83	Good
9.875	Good

Tb17. Data Labeling of Enhanced Version Response Set using Mixed Method

As observed in tables 16 and 17, using mixed method 3 out of 6 responses from non-enhanced response set are found, with long string pattern are labeled as 'Bad' as per the baseline value obtained for mixed method algorithm. Similarly, 1 out of 4 responses from enhanced response set, with long string pattern are labeled as 'Bad' as per the mixed method algorithm output. The data labels assigned to the data points from both non-enhanced and enhanced response set do show that the combination method is proved more useful as compared to each individual method of response time and pattern to identify bad quality responses from the data set. As this method utilizes the advantages of both the methods and rechecks for the average time spent in the pattern block as well, the results obtained can be more accurate as compared to each individual method. Looking at the data labels obtained, one can say that the enhanced version response set did contain better quality of responses as compared to the non-enhanced version. This implies that the usage of intrinsic and extrinsic motivational factors did help in the improvement of data quality obtained as compared to the non-enhanced version of the app. Using mixed method algorithm as well, a significant difference in the data quality is observed when the responses were using enhanced version of the mobile app.

6.4.6 Limitations

As the data set using enhanced content of the app was limited in number, all the conclusions for user compliance have been derived using the limited data set available. If the data set would have been larger, the conclusions would have been better justified.

I have derived different quality labels for each of the response set collected from both the groups NEN (non-enhanced) and EN (enhanced) using all three methods: response time, pattern and mixed method. However, this derivation of labels was done taking into consideration the limited amount of response data set available as part of this clinical study. As this data is live data on SCD patients who volunteer to participate in the clinical trial, we cannot control the number of patients who enroll into the study. As per the current situation limited numbers of patients have enrolled for the trial and thus the entire data analysis is driven based off this data set. The conclusions may provide few more categories of data points displaying a better distribution of data points.

6.4.7 Conclusion

The overall conclusion of the experiment can be derived that introduction of enhanced content did show a significant impact in the improvement of user compliance and data quality of survey responses collected. However due to the limitation of data set value, a variety of data classification was not possible to achieve. The methods used to monitor quality of data: response time, response pattern and mixed method have proved to be useful to identify the bad quality responses from the good ones. This identification and filtration of responses would help the biostatisticians and the clinicians to derive better analysis regarding reasons for hospital readmission using only the good quality self-reported patient data.

The results of this experiment do evaluate my second research question which states that the interventions in the form of intrinsic and extrinsic motivations do help in improving user compliance and data quality in long and repetitive mHealth applications. The next chapter provides details regarding the conclusions derived as part of the research conducted in this thesis and puts forward the future scope to take this research forward.

Results and interpretations derived for user compliance does show an improvement in the compliance rates observed as compared to earlier studies. Both non-enhanced and enhanced version patient groups showed an improvement in the weekly and daily compliance as well. As per section 6.4.3, notifications and badges can be the reason behind a good compliance rate observed in this study. Out of the two patients with the enhanced version of the app, one used games. Even though there is a limited amount of data for the enhanced version as compared to the non-enhanced version, the conclusions are well justified. One can say that user compliance rates are improved with the usage of intrinsic and extrinsic motivational factors in the mHealth app. If a larger data set can be obtained for the enhanced version, the results would help strengthen this derived conclusion.

For data quality measure, the mixed methods approach was used. Quality was measured for both enhanced and non-enhanced versions using all 3 algorithms. Baseline values derived from study 3 were used and improvised as per this study's experimental context and participation demographics. Results obtained from all the 3 algorithms did show that these methods help filter the bad quality responses from the good ones. The combination method of response time and pattern displayed a better classification of good and bad quality responses from both the participation groups. This method is more useful than their individual counterparts as it utilizes the advantages of each method and double checks the response data against time and pattern observed. All 3 methods displayed that the enhanced version response set achieved better quality responses as compared to the non-enhanced version. One can derive this conclusion that by using an intervention of intrinsic and extrinsic motivational factors data quality can be improved in the self-reported patient data.

The conclusion of this experiment helps validate the second research question of this thesis. The results and interpretations from this study do suggest that the usage of intrinsic and extrinsic motivational factors can achieve better compliance rates and quality of self-reported data in an mHealth application. As both research questions have been validated using all 4 experimental studies the next chapter provides the conclusion and future scope for this thesis.

CHAPTER 7

CONCLUSIONS AND FUTURE SCOPE

This chapter will discuss the lessons learnt and future work of this research. The discussion will start with a summary of the research methods and outcomes followed by the future work possible in this area.

7.1 Conclusions

This research studies the impact of intrinsic and extrinsic motivational factors on user compliance and data quality of self-reported mHealth surveys. The main contribution of this thesis is to find methods to detect careless responses and insufficient effort responding in mHealth surveys. The response time, response pattern and a mixed method of both time and pattern approach extends the use of surveys by utilizing user interaction logs to determine user compliance, and clickstream analysis which is used to detect long string patterns indicating data quality. The objective of this research is to evaluate whether an intervention based off intrinsic and extrinsic motivational factors does improve the user compliance and data quality in long and repetitive mHealth surveys. To that end, first contribution of this thesis is a case study in participatory design of a mHealth app based of PROMIS anxiety measures on graduate students. This research focuses on validating whether long and repetitive nature of surveys does impact compliance and data quality obtained. As a part of this case study, mHealth app delivering lengthy and repetitive surveys were tested on participants for a period of 4 weeks. The user interactions and log data were captured to monitor the activities carried out by a user in this trial period. This was a preliminary study and the number of participants was fewer than required to do a conclusive analysis. The study was small in scope and could not be done on SCD patient population, which was the objective. The results obtained from this study validated the first research question as part of this thesis. The preliminary study showed that lengthy and repetitive mHealth surveys negatively impact compliance and quality of data.

To validate the results obtained from study 1 on SCD population, a second user study was conducted on actual patient population. Participants downloaded an mHealth app from the Android and iOS app stores to fill out the surveys delivered to them. These surveys were delivered on a weekly basis for a trial period of 12 weeks. The user interactions with the app and the clickstream data were captured. This data along with the survey responses were captured and saved for further data analysis. Following the previous study, the results obtained from this study as well displayed similar trends in user compliance and data distribution patterns. In this study, a major decline in user compliance was observed right from second week itself as compared to study 1. Data quality as well was monitored as low in comparison to previous study. Thus, it showed that the situation is comparatively worse when an actual patient population is taken into consideration. The results obtained from this study strongly supported the conclusions derived from study 1 that lengthy and repetitive nature of surveys does impact the compliance rates and quality measures in self-reported patient data.

After obtaining results from study 1 and study 2, first part of thesis was completed. The next steps in this thesis were to identify methods that help detect insufficient effort responding (IER). A literature review provided methods to identify careless responses. Amongst these I made use of response time and response pattern as research methods to identify IER. Using these methods, I derived a combination of the response time and pattern, which helps in better classification of IER data from the rest. Algorithms were derived utilizing all these research methods details of which are provided in Chapter 4. To use these methods, certain baseline values need to be decided for response time and long string pattern. With this aim, a third study was conducted on students from software engineering to obtain baseline values for response time and long string pattern. Students were provided with the mHealth app that delivered lengthy surveys to them. They were made to take the same survey twice, once with responding as fast as possible and second time with responding after carefully reading and understanding the survey question. Results from both the survey attempts were analyzed and charts were plotted which provided distinct classification of data points using both response time and response pattern methods. Using these results a baseline value of 120 to 180 seconds for response time and long string pattern of greater than 6 were derived. These baseline values were further used in the final study to identify careless responding and IER.

The second research question of my thesis asks whether an intervention of intrinsic and extrinsic motivational factors embedded in mHealth application helps improve compliance and quality. To validate this question, a third study was conducted on SCD patients using an improvised version of mHealth app with informational messages as intrinsic and badges and games as extrinsic motivations embedded in the app. Few participants were made to use the non-enhanced version and few were made to use the enhanced version of the app. These participants were enrolled in the study for 5 weeks where each participant was provided with weekly and daily surveys for the entire duration. Badges were provided as rewards for activity completion and games were made available all the time to encourage user participation in the survey activities. User interaction logs and clickstream data along with responses from surveys were saved to monitor the participation of patients. Responses were analyzed using research methods and algorithms derived in Chapter 4. Using these methods, results were obtained for user compliance rates and data quality measures. Significant improvement was observed in user compliance rates as compared to study 1 and study 2 due to notifications and usage of badges and games. Similarly, the research methods helped identifying IER from the response set. Similarly, better quality of responses was also observed as compared to study 1 and study 2 when motivational factors were introduced in the app. The results from this study point out that usage of notifications, intrinsic and extrinsic motivations do help improving user compliance and data quality of responses collected. This helped me validate my second research question in this thesis.

In summary, this thesis' contributions are a case study in participatory design of an mHealth app embedded with intrinsic and extrinsic motivational factors, and a novel method to identify IER in mHealth surveys. To my knowledge, this is the first study of its kind where a combination of response time and response pattern methods was used to identify IER in clinical outcomes. From a clinical protocol compliance perspective, PROMIS app appears highly useful for improving user compliance and quality of data to be obtained for self-reported survey activities. This is a promising result for the future of mHealth apps for engaging patients in responding to survey activities with long and repetitive nature of surveys.

After the completion of patient's duration in clinical study 4, we asked them to respond to a usability survey. This survey is designed to ask questions regarding the usability factors of the app and whether the participant would like to use it in future. The survey questions are included in Appendix A. Positive responses have been received from both child patients as well as their parent proxies using the version 3a and 3b of the SCD-PROMIS app. Around 70% of the patient participants who completed this survey responded positively about the app. They mentioned that they liked the design of the app and it thought it was easy to use. They stated that they would like to use this app in the future. From parent participation group, all of them indicated that this app was easy to use and helpful for their child's well-being. These results indicate that the SCD-PROMIS app is helpful in improving user engagement and compliance rates; further improving clinical protocol.

Although these contributions are limited to a single domain, protocol and app, the outcomes are of interest due to the pain management domain, the nature of interventions (notifications), the use of mHealth app to help improve clinical compliance and the integration of innovative design such as intrinsic and extrinsic motivations (badges, games) resulting in improved user compliance and better data quality.

7.2 Future Work

From a software engineering point of view, the introduction of intrinsic and extrinsic motivational factors along with the combination of response time and pattern methods to detect careless responding discussed in this thesis can be applied to a variety of applications in mHealth domain. It was observed in the literature (section 3.2.2) that a variety of combination of methods for identifying IER gives better classification of data into good and bad quality. Combination of these methods utilizes the advantages of all the methods included to derive better results possible. Combination of response time and pattern can surely be applied to other mHealth applications used in remote pain management activities or used in the apps to collect self-reported pain data. Similarly, usage of intrinsic and extrinsic motivations can also be applied in a variety of apps from mHealth domain. However, generalizing the use of such methods and motivations for improvement in participation is a challenging task. The methods like long string patterns as well as clickstream analysis may not be applied to all the mHealth apps. Every app is unique of development and the choice of methods used to evaluate compliance and data quality depends largely on the context of use, type of users and the clinical goals of the mHealth app. Applying the combination of methods and motivations to other apps will give a better insight of how the participation of users for the apps can be improved in mHealth domain.

The way this thesis makes use of badges and few simple games to encourage user participation, future researchers can work in the direction to identify various aspects of games that attract participants. Accordingly, further research can be done in this direction to find out which aspects in the game attract participants and which aspects are not useful in improving compliance. This thesis also concentrating on the pediatric population, which showed improved results on the usage of badges and games. Accordingly, future researchers can try to identify various games and badging techniques to attract population from higher age groups as well. Researchers in this area will have to understand the nature of the app usage patterns, their data collection mechanism and accordingly design the motivational factors in the app.

One of the key findings in this research was the distinction between partially compliant and compliant users. The partially compliant users are of special interest as they are a significant number of participants who are almost compliant however due to some minor factors their usage of the app is less. Understanding the app usage of these partially compliant users will help the app designers to come up with better design to encourage this population towards taking the next to become compliant users. This will help future developers of the app to design motivational factors as per the user population and encourage participation from all the sectors.

In summary, I hope this research contributes to a growing multidisciplinary need to connect clinical research outcomes with software engineering processes.

REFERENCES

1. Ali, Eskinder Eshetu, Lita Chew, and Kevin Yi-Lwern Yap. "Evolution and current status of mhealth research: a systematic review." BMJ Innovations 2.1 (2016): 33-40.

2. Archer, Robert P., and David E. Elkins. "Identification of random responding on the MMPI-A." *Journal of Personality Assessment* 73.3 (1999): 407-421.

3. Archer, Robert P., Jennifer Fontaine, and Robert R. McCrae. "Effects of two MMPI-2 validity scales on basic scale relations to external criteria." *Journal of Personality Assessment* 70.1 (1998): 87-102.

4. Ashburn, Michael A., and Peter S. Staats. "Management of chronic pain. "The Lancet 353.9167 (1999): 1865-1869.

5. Ashley-Koch, Allison, Quanhe Yang, and Richard S. Olney. "Sickle hemoglobin (Hb S) allele and sickle cell disease: a huge review." *American Journal of Epidemiology* 151.9 (2000): 839-845.

6. Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W., "Detection of Random Responding on the MMPI--A." *Journal of Personality Assessment* 68.1 (1997): 139-151.

7. Baer, R. A., Kroll, L. S., Rinaldo, J., & Ballenger, J., "Detecting and discriminating between random responding and overreporting on the MMPI-A." *Journal of Personality Assessment* 72.2 (1999): 308-320.

8. Bagby, R. Michael, J. Roy Gillis, and Richard Rogers. "Effectiveness of the Millon Clinical Multiaxial Inventory Validity Index in the detection of random responding." Psychological Assessment: *A Journal of Consulting and Clinical Psychology* 3.2 (1991): 285

9. Ballas, Samir K., Kalpna Gupta, and Patricia Adams-Graves. "Sickle cell pain: a critical reappraisal." Blood 120.18 (2012): 3647-3656.

10. Bandura, Albert. "Perceived self-efficacy in cognitive development and functioning." Educational psychologist 28.2 (1993): 117-148.

11. Bandura, Albert. "Social foundations of thought and action: A social cognitive perspective." Englewood Cliffs, NJ: Princeton-Hall (1986).

12. Bashshur, R., Shannon, G., Krupinski, E., & Grigsby, J.,"The taxonomy of telemedicine." Telemedicine and e-Health 17.6 (2011): 484-494.

13. Beach, Daniel A. "Identifying the random responder." *The Journal of Psychology* 123.1 (1989): 101-103.

14. Berry, David T.R., Ruth A. Baer, and Monica J. Harris. "Detection of malingering on the MMPI: A meta-analysis." Clinical Psychology Review 11.5 (1991): 585-598.

15. Biau, David Jean, Solen Kernéis, and Raphaël Porcher. "Statistics in brief: the importance of sample size in the planning and interpretation of medical research." Clinical orthopaedics and related research 466.9 (2008): 2282-2288.

16. Bogen, Karen. "The effect of questionnaire length on response rates: a review of the literature." Proceedings of the Section on Survey Research Methods. American Statistical Association, Alexandria, VA, 1996.

17. Bradburn, Norman. "Respondent burden." Proceedings of the Survey Research Methods Section of the American Statistical Association. Vol. 35. 1978..

18. Bruehl, S., Lofland, K. R., Sherman, J. J., & Carlson, C. R., "The Variable Responding Scale for detection of random responding on the Multidimensional Pain Inventory." Psychological Assessment 10.1 (1998): 3.

19. Buechley, Robert, and Harry Ball. "A new test of" validity" for the group MMPI." *Journal of Consulting Psychology* 16.4 (1952): 299

20. Butcher, James N. "Minnesota multiphasic personality inventory." Corsini Encyclopedia of Psychology (1989).

21. Cannell, Charles F., and Robert L. Kahn. "Interviewing." The handbook of social psychology 2 (1968): 526-595.

22. Cechanowicz, Jared, Carl Gutwin, Briana Brownell, and Larry Goodfellow., "Effects of gamification on participation and data quality in a real-world market research domain." Proceedings of the First International Conference on Gameful Design, Research, and Applications. ACM, 2013.

23. Cella, David, Susan Yount, Nan Rothrock, Richard Gershon, Karon Cook, Bryce Reeve, Deborah Ader, James F. Fries, Bonnie Bruce, and Mattias Rose., "The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years." Medical care 45.5 Suppl 1 (2007): S3.

24. Cheng, Chihwen, R. Clark Brown, Lindsey L. Cohen, Janani Venugopalan, Todd H. Stokes, and May D. Wang.,"iACT-An interactive mHealth monitoring system to enhance psychotherapy for adolescents with sickle cell disease." Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE. IEEE, 2013.

25. Chrons, Otto, and Sami Sundell. "Digitalkoot: Making Old Archives Accessible Using Crowdsourcing." Human Computation. 2011.

26. Costa, Paul T., and Robert R. McCrae. "The revised neo personality inventory (neo-pi-r)." The SAGE handbook of personality theory and assessment 2 (2008): 179-198.

27. Crawford, Scott D., Mick P. Couper, and Mark J. Lamias. "Web surveys: Perceptions of burden." Social science computer review 19.2 (2001): 146-162

28. Curran, Paul G. "Methods for the detection of carelessly invalid responses in survey data." *Journal of Experimental Social Psychology* 66 (2016): 4-19.

29. Dan, Oana M., and Jennie W. Lai. "How am I doing? The effects of gamification and social sharing on user engagement." Proc. 68th Ann. Conf. American Assoc. for Public Opinion Research. 2013.

30. Dansie, E. J., and Dennis C. Turk. "Assessment of patients with chronic pain." *British Journal of Anaesthesia* 111.1 (2013): 19-25.

31. Davis, Jonathan R., Vivian P. Nolan, Janet Woodcock, and Ronald W. Estabrook, "Assuring data quality and validity in clinical trials for regulatory decision making." Workshop Report, Roundtable on Research and Development of Drugs, Biologics, and Medical Devices, Division of Health Sciences Policy, Washington. 1999.

32. Deci, Edward L. "8: Ryan, RM (1985). Intrinsic motivation and self-determination in human behavior." New York and London: Plenum (86).

33. Deci, Edward L., and Richard M. Ryan. "Self-determination theory: A macrotheory of human motivation, development, and health." Canadian psychology/Psychologie canadienne 49.3 (2008): 182.

34. Deterding, Khaled. "Nacke, ua (2011) Gamification: Toward a Definition³." CHI 2011 Gamification Workshop Proceedings. Vancouver, BC, Canada.

35. Deterding, Sebastian. "Gamification: designing for motivation." interactions 19.4 (2012): 14-17.

36. Deterding, Sebastian, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon, "Gamification. using game-design elements in non-gaming contexts." CHI'11 Extended Abstracts on Human Factors in Computing Systems. ACM, 2011.

37. Estrin, Deborah, and Ida Sim. "Open mHealth architecture: an engine for health care innovation." Science 330.6005 (2010): 759-760
38. Farlex Partner Medical Dictionary ,2012, www.medicaldictionary.thefreedictionary.com

39. Fiordelli, Maddalena, Nicola Diviani, and Peter J. Schulz. "Mapping mHealth research: a decade of evolution." *Journal of Medical Internet Research* 15.5 (2013): e95

40. Galesic, Mirta, and Michael Bosnjak. "Effects of questionnaire length on participation and indicators of response quality in a web survey." Public opinion quarterly 73.2 (2009): 349-360.

41. García-Gómez, Juan M., Isabel de la Torre-Díez, Javier Vicente, Montserrat Robles, Miguel López-Coronado, and Joel J. Rodrigues, "Analysis of mobile

health applications for a broad spectrum of consumers: a user experience approach." *Health Informatics Journal* 20.1 (2014): 74-84.

42. Garris, Rosemary, Robert Ahlers, and James E. Driskell. "Games, motivation, and learning: A research and practice model." Simulation & gaming 33.4 (2002): 441-467

43. Gary Kevin, PhD, Cleary Kevin, PhD, Hinds Pamela, RN, PhD; Streisand Randi, PhD, Guerrera Michael, MD, Wang Jichuan, PhD, Finkel Julia, MD., "SCD-PROMIS: A software platform that enhances self-efficacy and patient-provider engagement with the goal of reducing readmission rates in children with sickle cell pain Grant ID number: 20083487" Organization 15: 17

44. Garry, K. A., Rallabhandi, P., Walek, E., Nettleton, M., Ahmed, I., Wang, J., ... & Quezado, Z.,"An mHealth hybrid app for self-reporting pain measures for sickle cell disease." Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT), 2016 IEEE. IEEE, 2016.

45. Gibson, David, Nathaniel Ostashewski, Kim Flintoff, Sheryl Grant, and Erin Knight.,"Digital badges in education." Education and Information Technologies 20.2 (2015): 403-410.

46. Goel, Sonu, Nidhi Bhatnagar, Deepak Sharma, and Amarjeet Singh. "Bridging the human resource gap in primary health care delivery systems of developing countries with mHealth: Narrative literature review." JMIR mHealth and uHealth 1.2 (2013): e25

47. Greene, Roger L. "An empirically derived MMPI carelessness scale." *Journal of Clinical Psychology* 34.2 (1978): 407-410.

48. Green, Samuel B., and Thomas Stutzman. "An evaluation of methods to select respondents to structured job-analysis questionnaires." Personnel Psychology 39.3 (1986): 543-564.
49. Habgood, MP Jacob, and Shaaron E. Ainsworth. "Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games." *The Journal of the Learning Sciences* 20.2 (2011): 169-206.

50. Hamari, Juho, and Veikko Eranti. "Framework for designing and evaluating game achievements." Proc. DiGRA 2011: Think Design Play 115.115 (2011): 122-134.

51. Han, Jennifer, and Chris Barker. "Comparison of Intrinsically Motivating Factors in Educational Games."

52. Heerwegh, Dirk, and Geert Loosveldt. "Face-to-face versus web surveying in a high-internet-coverage population differences in response quality." Public Opinion Quarterly 72.5 (2008): 836-846.

53. Hinds, Pamela S., Suzanne L. Nuss, Kathleen S. Ruccione, Janice S. Withycombe, Shana Jacobs, Holly DeLuca, Charisse Faulkner, "PROMIS pediatric measures in

pediatric oncology: Valid and clinically feasible indicators of patient-reported outcomes." Pediatric blood & cancer 60.3 (2013): 402-408.

54. Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. "Detecting and deterring insufficient effort responding to surveys." *Journal of Business and Psychology* 27.1 (2012): 99-114.

55. Huang, J. L., Bowling, N. A., Liu, M., & Li, Y., "Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions." *Journal of Business and Psychology* 30.2 (2015): 299-311.

56. Isakovic, M., Cijan, J., Sedlar, U., Volk, M., & Bester, J., "The Role of mHealth applications in societal and social challenges of the future." Information Technology-New Generations (ITNG), 2015 12th International Conference on. IEEE, 2015.

57. Istepanian, R., Laxminarayan, S., & Pattichis, C. S. (2006). M-Health: Emerging Mobile Health Systems. M-Health: Emerging Mobile Health Systems. Retrieved from http://adsabs.harvard.edu/abs/2006mhem.book.

58. Istepanian, Robert, Swamy Laxminarayan, and Constantinos S. Pattichis. M-health. New York, NY: Springer Science+ Business Media, Incorporated, 2006.

59. Jackson vocational interest survey: manual. London, Ont.: Research Psychologists Press, 1977.

60. Jacob, Eufemia, Carol Pavlish, Joana Duran, Jennifer Stinson, Mary Ann Lewis, and Lonnie Zeltzer.,"Facilitating pediatric patient-provider communications using wireless technology in children and adolescents with sickle cell disease." *Journal of Pediatric Health Care* 27.4 (2013): 284-292.

61. Jacob, Eufemia, Joana Duran, Jennifer Stinson, Mary Ann Lewis, and Lonnie Zeltzer."Remote monitoring of pain and symptoms using wireless technology in children and adolescents with sickle cell disease." *Journal of the American Academy of Nurse Practitioners* 25.1 (2013): 42-54.

62. Jacob, Eufemia, Jennifer Stinson, Joana Duran, Ankur Gupta, Mario Gerla, Mary Ann Lewis, and Lonnie Zeltzer, "Usability testing of a Smartphone for accessing a web-based e-diary for self-monitoring of pain and symptoms in sickle cell disease." *Journal of Pediatric Hematology/Oncology* 34.5 (2012): 326

63. Jakobsson, Mikael. "The achievement machine: Understanding Xbox 360 achievements in gaming practices." Game Studies 11.1 (2011): 1-22.

64. Johnson, John A. "Ascertaining the validity of individual protocols from web-based personality inventories." *Journal of Research in Personality* 39.1 (2005): 103-129.

65. Johnson, W. Russell, Nicholas A. Sieveking, and Earl S. Clanton. "Effects of alternative positioning of open-ended questions in multiple-choice questionnaires." *Journal of Applied Psychology* 59.6 (1974): 776..

66. Kay, Misha, Jonathan Santos, and Marina Takane. "mHealth: New horizons for health through mobile technologies." World Health Organization 3 (2011): 66-71.

67. Kraut, Allen I., Alan D. Wolfson, and Alan Rothenberg. "Some effects of position on opinion survey items." *Journal of Applied Psychology* 60.6 (1975): 774.

68. Krishnankutty, Binny, Shantala Bellary, BR Naveen Kumar, and Latha S. Moodahadu, "Data management in clinical research: an overview." *Indian Journal of Pharmacology* 44.2 (2012): 168.

69. Krosnick, Jon A. "Response strategies for coping with the cognitive demands of attitude measures in surveys." Applied cognitive psychology 5.3 (1991): 213-236

70. Kurtz, John E., and Catherine L. Parrish. "Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R." *Journal of Personality Assessment* 76.2 (2001): 315-332.

71. Lai, Jennie W., Michael W. Link, and Lorelle Vanno. "Emerging techniques of respondent engagement: Leveraging game and social mechanics for mobile application research." 67th conference of the American Association for Public Opinion Research, Orlando, FL, May. 2012..

72. Laitinen, Sauli. "Do usability expert evaluation and test provide novel and useful data for game development?" *Journal of Usability Studies* 1.2 (2006): 64-75.

73. Lalloo, Chitra, Lindsay A. Jibb, Jordan Rivera, Arnav Agarwal, and Jennifer N. Stinson, "There's a pain app for that": Review of patient-targeted smartphone applications for pain management." *The Clinical Journal of Pain* 31.6 (2015): 557-563.

74. Leiner, Dominik J. "Too fast, too straight, too weird: Post hoc identification of meaningless data in internet surveys." Too Straight, Too Weird: Post Hoc Identification of Meaningless Data in Internet Surveys (November 30, 2013) (2013).

75. Lucas, Richard E., and Brendan M. Baird. "Global Self-Assessment." (2006).

76. Lugtigheid, A., and S. Rathod. "Questionnaire Length and Response Quality: Myth or Reality?" Stamford, CT: Survey Sampling International (2005).

77. Mahalanobis, Prasanta Chandra. "On the generalized distance in statistics." Proceedings of the National Institute of Sciences (Calcutta) 2 (1936): 49-55.

78. Malhotra, Neil. "Completion time and response order effects in web surveys." Public Opinion Quarterly 72.5 (2008): 914-934.

79. Mavletova, Aigul. "Data quality in PC and mobile web surveys." Social Science Computer Review 31.6 (2013): 725-743.

80. Meade, Adam W., and S. Bartholomew Craig. "Identifying careless responses in survey data." Psychological methods 17.3 (2012): 437.

81. Miller, Aaron S., Joseph A. Cafazzo, and Emily Seto. "A game plan: Gamification design principles in mHealth applications for chronic disease management." *Health Informatics Journal* 22.2 (2016): 184-193.

82. Miller, Jeff. "Burke Panel Quality R and D." Cincinnati: Burke, Inc (2008).

83. Montola, Markus, Timo Nummenmaa, Andrés Lucero, Marion Boberg, and Hannu Korhonen., "Applying game achievement systems to enhance user experience in a photo sharing service." Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era. ACM, 2009.

84. Morey, Leslie C., and Christopher J. Hopwood. "Efficiency of a strategy for detecting back random responding on the personality assessment inventory." Psychological Assessment 16.2 (2004): 197.

85. National Heart, Lung, and Blood Institute. The Management of Sickle Cell Disease. 4th ed. Bethesda, MD: National Institutes of Health; 2002:15–18. NIH publication 02-2117

86. Nielsen, Jakob. "Usability inspection methods." Conference companion on Human factors in computing systems. ACM, 1994.

87. Nielsen, Jakob, and Rolf Molich. "Heuristic evaluation of user interfaces." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 1990.

88. Park, Linda G., Jill Howie-Esquivel, and Kathleen Dracup. "A quantitative systematic review of the efficacy of mobile phone interventions to improve medication adherence." *Journal of Advanced Nursing* 70.9 (2014): 1932-1953.

89. Peck, Roxy, and Jay L. Devore. Statistics: The exploration & analysis of data. Cengage Learning, 2011.

90. Piedmont, Ralph L., Robert R. McCrae, Rainer Riemann, and Alois Angleitner., "On the invalidity of validity scales: evidence from self-reports and observer ratings in volunteer samples." *Journal of Personality and Social Psychology* 78.3 (2000): 582.

91. Pinsoneault, Terry B. "A Variable Response Inconsistency Scale and a True Response Inconsistency Scale for the Jesness Inventory." Psychological Assessment 10.1 (1998): 21.

92. Platt, Orah S., Bruce D. Thorington, Donald J. Brambilla, Paul F. Milner, Wendell F. Rosse, Elliott Vichinsky, and Thomas R. Kinney, "Pain in sickle cell disease: rates and risk factors." *New England Journal of Medicine* 325.1 (1991): 11-16.

93. Preist, Chris, and Robert Jones. "The use of games as extrinisic motivation in education." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015.

94. Price, Matthew, Erica K. Yuen, Elizabeth M. Goetter, James D. Herbert, Evan M. Forman, Ron Acierno, and Kenneth J. Ruggiero., "mHealth: a mechanism to deliver more accessible, more effective mental health care." Clinical psychology & psychotherapy 21.5 (2014): 427-436.

95. Riley, William T., Daniel E. Rivera, Audie A. Atienza, Wendy Nilsen, Susannah M. Allison, and Robin Mermelstein, "Health behavior models in the age of mobile interventions: are our theories up to the task?." Translational behavioral medicine 1.1 (2011): 53-71.

96. Rosse, Joseph G., Robert A. Levin, and Margaret D. Nowicki. "Assessing the impact of faking on job performance and counter-productive job behaviors." P. Sackett (Chair), New empirical research on social desirability in personality measurement. Symposium conducted at the 14th annual meeting of the Society of Industrial Organizational Psychology, Atlanta, GA. 1999

97. Ryan, Richard M. "Psychological needs and the facilitation of integrative processes." *Journal of Personality* 63.3 (1995): 397-427.

98. Ryan, Richard M., and Edward L. Deci. "Intrinsic and extrinsic motivations: Classic definitions and new directions." Contemporary educational psychology 25.1 (2000)

99. Ryan, Richard M., and Edward L. Deci. "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." American psychologist 55.1 (2000)

100. Seligman, Martin EP. Helplessness: On depression, development, and death. WH Freeman/Times Books/Henry Holt & Co, 1975.

101. Schinka, John A., Bill N. Kinder, and Thomas Kremer. "Research validity scales for the NEO--PI--R: Development and initial validation." *Journal of Personality Assessment* 68.1 (1997): 127-138.

102. Shah, Nirmish, Jude Jonassaint, and Laura De Castro. "Patients welcome the sickle cell disease mobile application to record symptoms via technology (SMART)." Hemoglobin 38.2 (2014): 99-103.

103. Singh, Arunjot, Sarah Wilkinson, and Sandra Braganza. "Smartphones and pediatric apps to mobilize the medical home." *The Journal of Pediatrics* 165.3 (2014): 606-610.

104. Smith, Wally R., Lynne T. Penberthy, Viktor E. Bovbjerg, Donna K. McClish, John D. Roberts, Bassam Dahman, Imoigele P. Aisiku, James L. Levenson, and

Susan D. Roseff., "Daily assessment of pain in adults with sickle cell disease." Annals of internal medicine 148.2 (2008): 94-101.

105. Spring, Bonnie, Marientina Gotsis, Ana Paiva, and Donna Spruijt-Metz., "Healthy apps: mobile devices for continuous monitoring and intervention." IEEE pulse 4.6 (2013): 34-40.

106. Stone, Arthur A., Saul Shiffman, Joseph E. Schwartz, Joan E. Broderick, and Michael R. Hufford, "Patient non-compliance with paper diaries." Bmj 324.7347 (2002): 1193-1194.

107. Union IT (2010) The world in 2010: ICT facts and figures. International Telecommunication Union.

108. Von Korff, Michael, Johan Ormel, Francis J. Keefe, and Samuel F. Dworkin, "Grading the severity of chronic pain." Pain 50.2 (1992): 133-149

109. Weinstein, James N., and Richard A. Deyo. "Clinical research: issues in data collection." Spine 25.24 (2000): 3104-3109.

110. Wenemark, Marika. The respondent's perspective in health-related surveys. Diss. Linköping University Electronic Press, 2010.

111. Wenemark, Marika, Andreas Persson, Helle Noorlind Brage, Tommy Svensson, and Margareta Kristenson,"Applying motivation theory to achieve increased respondent satisfaction, response rate and data quality in a self-administered survey." *Journal of Official Statistics* 27.2 (2011): 393-414.

112. Wetter, Martha W., Ruth A. Baer, David TR Berry, Gregory T. Smith, and Lene H. Larsen, "Sensitivity of MMPI-2 validity scales to random responding and malingering." Psychological Assessment 4.3 (1992): 369.

113. World Health Organization (WHO). "mHealth: new horizons for health through mobile technologies: second global survey on eHealth." Switzerland: Global Observatory for eHealth, 2011

114. Williams, C. L., J. R. Graham, R. P. Archer, A. Tellegen, Y. S. Ben-Porath, and B. Kaemmer. "MMPI-A: Minnesota Multiphasic Personality Inventory-Adolescent: Manual for Administration, Scoring, and Interpretation." (1992).

115. Wilson, Mark A., Robert J. Harvey, and Barry A. Macy. "Repeating items to estimate the test-retest reliability of task inventory ratings." Journal of Applied Psychology 75.2 (1990): 158.

116. Yang, Sungwon, Eufemia Jacob, and Mario Gerla. "Web-based mobile e-Diary for youth with Sickle Cell Disease." Consumer Communications and Networking Conference (CCNC). IEEE, 2012.

APPENDIX

A USABILITY SURVEY

Usability Survey Questionnaire

Delivered to Patients and Parents Enrolled in Study 4 (patient/parent)

Below are the survey questions delivered at the end of the 5-week trial period to the patient / proxy parent who were enrolled in study 4. These questions ask their feedback regarding the enhancements in the app and survey design.

Part 1: Survey Delivered to Patients

QUESTIONS	OPTION GROUP
I like using the app	1
How many times did you stop using the app because you were bored?	2
How many times did you stop using the app because you were confused?	2
I want to keep using the app.	1
I thought the app was easy to use	1

Part 2: Survey Delivered to Parents

QUESTIONS	OPTION GROUP
Do you think it is boring to work with the app?	1
Would you like to continue using the app?	1
Did you think the app was easy to use?	1
Do you feel this app helped you participate in pain management for your child?	1
Do you think this app helped in your child's well-being?	1

Part 3: Response Options:

1. Likert Scale 1 to 5 options: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree
2. 0, 1 or 2 times, 3 or more times