Application of data analytics and machine learning on data collected by smartphones to understand human behavioural patterns

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Teena Thomas Vattukalathil

Permission to Use

Abstract

A growing number of health studies seek to leverage smartphone-based recording to continuously monitor consenting participants' health behaviours, including those related to mental health, mobility, and activity. So as to better understand health risks and the influence of the environment on human physical and mental health conditions, such studies commonly use smartphones to collect health behaviour relevant metrics such as screen state, app usage, location, activity level, browsing behaviour, etc. They also typically use survey instruments incorporating questionnaires, voice recordings, photos, multi-media content on which the user is asked to provide feedback, etc. When the data volume and variety grow substantially −−− such as is common with sensed data −−− then challenges associated with data quantity, quality, diversity, trustworthiness, etc. also increase significantly. Because most health scientists are unfamiliar with tools and concepts required for effective analysis of such high-volume and high-velocity data, it is challenging for health scientists alone to perform the computationally intensive analyses needed to secure certain types of insight from the collected data. The primary objective of this thesis is to provide computational mechanisms to support research teams associated with 3 distinct case studies utilizing smartphone-based data, so as to help obtain insights accessible to team health scientists.

The data sets for these three studies were collected from participants using a pre-existing smartphone-based application named Ethica. Such data was accumulated over a period ranging from 2 weeks to 6 months – with the study period differing across the three studies – through a set of surveys and mobile sensors such as those for the battery, screen state, GPS, etc.

This thesis addresses three significant challenges associated with the extraction and processing of smartphone data. The first is the computational burden and intricacies associated with data extraction, preprocessing and analytic steps. The second consists of a need for handling omitted and missing data points with the help of machine learning and statistical methods. The final challenge covered here is to secure valuable findings from these data sets through exploratory analysis following examination of participant adherence patterns and evaluation of the quantity and quality of the data collected. The methods applied in this thesis are useful for other studies using the Ethica platform because of the shared structure of Ethica datasets and the capacity of the code to be reused and readily adapted for other such datasets.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Nathaniel Osgood for his guidance, continuous support, motivation and constant help throughout my research and during the writing of this thesis. It was like a dream come true for me to be able to come here from India and finish my Masters under my Supervisor′s excellent guidance in this prestigious university.

I very sincerely thanks my advisory committee members Dr. Eric Neufeld, Dr. Ian McQuillan and external examiner Dr. Vivian Ramsden for their excellent guidance, meaningful suggestions, and, most importantly, their valuable time.

I submit my heartiest gratitude to my three research collaborators: 1) Dr. Anne-Linda Camerini and Laura Marciano of MEDIATICINO team, University of Lugano, Switzerland; 2) Dr. Rudy Bowen, Dr. Lloyd Balbuena, and their team from the Department of Psychiatry, University of Saskatchewan; 3) Dr. Alex Wong and his team from Saskatchewan Health Authority, for the wonderful research opportunities and for financially supporting me throughout my graduate studies. They have each not only helped me gain knowledge and research experience in this field, but also helped me improve myself on a professional level.

I humbly extend my thanks to my labmates for their sheer support during the ups and downs of my graduate studies in the past years. A very special thank you to Christine Hillis for managing our team and organizing events to keep CEPHIL active.

Last but not least, I would like to thank my family and friends for their continuous support and inspiration that encouraged me to move past the lows and march towards the highs of the future.

Dedication

I dedicate my thesis to God almighty for all of his blessings and to my family for their endless love and support.

My mother, Silu Thomas, who has always stood by my side, believed in me and inspiring me to reach my goals. She is a true role model of how caring, truthful and lovable a mother could be. And the greatest gift to my life. Thank you, Amma!. My father: Kuruvilla Thomas (deceased), for encouraging me to dream big and turn dreams to goals. He taught me to stay humble, respectful and honest in all stages of life. Thank you, Pappa!. And finally my siblings: Meenu Thomas and Royce Thomas. Thank you for believing in their older sister′s dreams and continuously encouraging her to dream big.

Contents

List of Tables

List of Figures

List of Abbreviations

JVM	Java virtual machine
HMM	Hidden Markov Model
SR	Hidden state corresponding to screen state ON and Ethica Recording ON
S/R	Hidden state corresponding to screen state ON and Ethica Recording OFF
/SR	Hidden state corresponding to screen state OFF and Ethica Recording ON
/S/R	Hidden state corresponding to screen state OFF and Ethica Recording OFF
SOnR	Hidden state corresponding to Screen turning ON and Ethica Recording ON
SOffR	Hidden state corresponding to Screen turning Off and Ethica Recording ON
ROnS	Hidden state corresponding to Ethica recording turning ON and screen state is ON
ROn/S	Hidden state corresponding to Ethica recording turning ON and screen state is OFF
TPM	Transition Probability Matrix
ABM	Agent based model
SI	Suicidal Ideation
HIV	Human Immunodeficiency Virus
PLWHIV	People Living with HIV
EMA	Ecological momentary assessment

Chapter 1

Introduction

"Data is everywhere" [1], and health data science is a rapidly evolving field in which researchers tackle real world health problems using big data-driven solutions. The techniques of data science are modern approaches that have emerged by the combination of skills from different disciplines, including computer science, health science, and statistics [2]. Such research helps to generate meaningful insights from big data streaming or are collected from different sources such as electronic medical records, clinical databases, health administrative data, social network platforms, wearables, mobile phones, etc., with larger volume and unstructured forms. Health research is an evolving field, and technological advances are necessary to understand changing lifestyles, track and anticipate disease outbreaks, elucidate the effects of unhealthy habits, better detect mental illness, etc. Survey research is a widespread traditional means of collecting data in the health sciences to understand an individual's health condition, risk factors, feedback about treatment and patient satisfaction. However, this kind of data collection technique had several challenges associated with it, such as the burden of filling out paper instruments or responding over the phone, infeasibility of doing high frequency sampling, difficulty of recalling earlier situations because of elapsed time, difficulty of communicating context [e.g., social context or geographic location] more complex descriptions of conditions, behaviour (e.g., what was eaten, degree of physical activity/sedentary behaviour), symptom (e.g., presence of a rash) etc. in case of surveys. But behavioural studies are of great importance in health research, and having efficient methods to process and understand behavioural patterns of a larger population can aid by offering insights about the influence of changing lifestyles on public health, changes needed in the public policy domains, etc. Collecting and dealing with big data from large populations are challenging and it is cumbersome to effectively draw patterns from such big data sets using traditional methods. Hence, choosing the right tools for data collection, storage, analytic processing, etc., and the use of efficient programming languages and platforms to handle big health data, maintain proper databases that can support a large volume of data and can support the parallel programming features, etc., is valuable for successful completion of studies that involve big data sets.

Recent years have witnessed an increasing proliferation of mobile technology in health research. One use of such technologies has been to collect big health data by monitoring the day to day life patterns of populations and their exposure to health hazards. These data can aid researchers in drawing insights regarding health behaviours and effects of exposures while imposing a lower burden on participants by obviating the need for participants to carry extra study devices or data collection tools. However, big data can be challenging for health science researchers to analyze using traditional techniques. Many of these challenges relate to the "4 V's", or 4 dimensions, of big data – namely Volume, Variety, Velocity and Veracity [3]. Some commentators have argued that big data is often further characterized by an additional 'V' – Value – in recognition of the valuable discoveries or scientific findings that frequently extend from big data. In this thesis, we are performing data analytics and machine learning to support the decision making on three different behavioural studies conducted by three different health research teams that leverage a specific type of big data – data that is collected nearly continuously from study participants via smartphones for a period ranging from 2 weeks to 6 months on a minute level resolution through a University of Saskatchewan-oriented mobile epidemiological data collection platform named "Ethica".

1.1 Research problem

On the one hand, large volumes of data help researchers identify valuable patterns and associations and offer considerable potential for informing intervention planning and policymaking to lower health burdens. On the other hand, while tools like Ethica support the increasingly straightforward definition and deployment of smartphone-based studies, there are challenges associated with data processing and analytics that impede many health researchers in realizing the full potential of its benefits. Reliable delivery of insight from analysis of such large-scale datasets currently relies heavily on expertise in advanced analytic tools and supporting computational concepts and methods [4]. The main objective of this thesis is to provide computational mechanisms to tackle challenges associated with big data analytics for specific case studies, and to convert this data into insights accessible by health scientists associated with those studies.

To achieve that, three major research problems are addressed in this thesis in the three case studies in Chapter 2, Chapter 3 and Chapter 4. The first challenge consists of managing the computational burden of data collected by smartphones. The second problem is to deal effectively with limitations in the quality of smartphone sensor data – particularly by using machine learning to robustly infer the underlying situation in the presence of missing data. Finally, the thesis applies data analytics to identify reveal patterns in smartphone-collected data.

The first challenge within this area relates to the computational complexities associated with the large volume of raw data streamed from the mobile sensors. This challenge can be dealt with by using proper analytic pipelines and the choice of scalable programming languages. But many of the traditional health science tools are not sufficiently scalable for analysis of such data. Some problems allow for processing by traditional tools following an initial analysis stage using other toolsets suited to effective operation with big data. The second challenge concerns the variable character of data collected from sensors. Often this set of concerns can be addressed by the implementation of proper pre-processing steps and application of statistical methods or machine learning algorithms that suit the data generation processes. However, undertaking these steps requires advanced skills to manipulate and process data that drawn on knowledge from computer science, statistics, machine learning, etc., which will help in feature extraction, missing data imputation, labelling underlying hidden states associated with data generation, data fusion, inference and classification, etc. For example, if we have to use accelerometer sensors to label a person's activities (standing, sitting, walking) during the participation period to reveal the associated activity patterns over time, a proper pre-processing of accelerometer data and the application of machine learning algorithms (e.g., deep learning or Hidden Markov models (HMM)) are essential. The third challenge lies in extracting further valuable findings from the data. This requires additional quantitative, exploratory and predictive analysis of smartphone data using interactive visualization and analysis tools, machine learning algorithms and statistical analysis (T-test, chi-square test, Kolmogorov-Smirnov test, etc.). This step can be partially performed by the health researchers after getting help from the data analyst team to aggregate the unstructured data into a structured form. But the aggregation part of big data usually requires extra efforts due to the large volume of raw data collected, and the need to summarize and reduce the dimensionality of such data to make it fit for further analysis. For example, a study that runs for six months and collects data continuously at a minute level resolution from a participant pool in the dozens to hundreds will need computational efforts even after the 1^{st} step of pre-processing. This reflects, amongst other factors, the computational burden and intricacies involved in processing and aggregating the data, in understanding the time series patterns in a daily or hourly manner, and in cross-linking sensor and survey responses. As another example, survey responses in audio/picture format needs to be transcribed to texts or to be labelled, which often requires advanced techniques based on the quantity and character of the data, such as the Google Cloud Text-to-Speech API or implementation of deep learning methods – methods whose use generally falls outside the limits of health researcher training.

1.2 Background and literature review

1.2.1 Behavioural studies using smartphone data

While exceptionally valuable in many studies, traditional means of collecting health-relevant behavioural data – such as in-person or phone-based interviews, mailed surveys, etc. – suffer from some notable limitations, including low temporal resolution and fidelity, and recall and response bias [5]. Modern smartphones are ubiquitous and possess sensors that can achieve the functionality of dedicated devices [5]. Such capacities, amongst others, have encouraged researchers to deploy smartphones as data collection tools [5]. Many research studies have collected and analyzed data from smartphone devices to understand human behavioural patterns [6–12]. Such studies use smartphone-collected survey data, sensor data, or both, to study participant activity patterns, contact patterns, mobility patterns, mental health, etc. The major sensors used within these studies include GPS, Wi-Fi, accelerometer, battery, Bluetooth, screen state, etc., with particular studies choosing the sensors employed according to the research questions being pursued. In this thesis, we have used a smartphone-based application named Ethica to collect sensor and survey data from participants during their study period. High-level information regarding this data collection tool is provided in the next subsection;

the general features of the data analytics pipeline employed are explained in subsection 1.2.3.

1.2.2 Ethica: The data collection tool

Ethica is a multi-tier system used to define, configure and deploy health studies, and to collect, store and visualize data from smartphones and wearables carried by study participants. Ethica supports defining a broad range of custom study attributes and configurations, without the need for programming. The platform originated in CEPHIL/DISCUS labs at the University of Saskatchewan within a research project named iEpi [13, 14]. From then onwards it has been successfully used by between 10,000 and 20,000 participants across more than 100 health research studies in different parts of the world, namely North America, Europe, Australia, and Asia [14,15]. The data collected from the study participants can aid research understanding of human behavioural patterns associated with mental health, activity level, sedentary behaviour, screen time usage, etc. Ethica can further help to understand several other metrics at a regional or population level, such as behaviour changes during a communicable disease outbreak, source of food poisoning, geographical barriers associated with active lifestyle or access to treatment, etc., with the support of smartphone sensors such as GPS, Wi-Fi, accelerometer, linear acceleration, screen state, battery, etc. The sensor data is collected for the entire duration of study at a minute level resolution from the study start time until its end date. The high temporal resolution of such data can aid researcher understanding of the aforementioned behavioural patterns of the study participants in a more detailed way. Also, the comparatively low cost, easy configuration and set up, ease of study monitoring, and low burden of data collection process on study participant daily routines enhances Ethica's value proposition for health researchers.

Ethica offers a user-friendly interface available to anyone who creates an account in the Ethica website, which enables them to define a study based on their needs. Every research team can access the Ethica interface and create a researcher role that allows them to set up the study specifications for their upcoming study; the defining study administrator can further be shared access to the resulting study to other collaborators or researchers who want to contribute to study setup. The creation of a study in the Ethica interface enables researchers to choose and configure study specifications such as study name, duration, start date, participant count expected, etc., can also let researchers choose the data collection sensors needed for their specific study without the need of technical expertise or computer skills. This freedom to declaratively specify sensors, aspects of study design, and graphically set up surveys, and visualize incoming data with custom tools make Ethica unique when compared to other existing smartphone-based data collection platforms.

Ethica helps researchers to set up surveys with several types of trigger logic. Perhaps the most widespread trigger mechanism consists of time-based triggering, which enables the researcher to schedule the survey trigger time associated with each survey through the Ethica interface by choosing a day/time point or window at which the particular survey should be triggered for the study participants. For example, survey1 can be configured with trigger logic to schedule it to occur daily at 8 pm, or at a random time between 7 pm and 9 pm. Secondly, researchers can set up surveys in a way that will be represented as a button on participant smartphone and web interfaces of the Ethica application; this button allows the participant to self-trigger that particular survey at any time; following the survey triggering, the participant can then populate and submit the survey responses. The third type of triggering mechanism is an advanced trigger type in which researchers can set up surveys to be triggered based on participants location using the geo-fence feature that is supported in Ethica, or due to the presence or absence of a beacon, etc. In addition to the survey trigger features, Ethica also helps researchers to set up multiple types of survey questions in their study, such as single answer, multiple answers, height/weight with unit choice, visual analogue scale, pictures, audio responses, etc., which makes Ethica a versatile platform for many common types of health studies. Another important point is the simple study set up steps and straightforward user interface design which helps researchers to set up their study with ease and supports subsequent modification of study details before or during the study launch without interrupting the ongoing data collection process or requiring software development involvement.

1.2.3 Data analytics pipeline

Big data analytics is a multi-step process that typically involves several steps – creating a hypothesis, performing and refining the analytics [4], etc. Putting aside several smaller steps, the main steps involved in generating valuable insights from big data lie in the collection, cleaning, integration, modelling and analysis, interpretation and deployment [4]. It has further been observed that the big data collected on an unprecedented scale in recent years facilitates the decision making more based on data-driven mathematical models than compared to prior assumptions about the process [16] – an observation which points towards the widespread importance of data analysis and pipelines to make the most out of big data.

Programming languages and framework used for data analytics

Choosing the right programming language and the right framework to handle big data sets will aid in addressing the first level problem of the high volume of datasets. In this project, we have used the language Scala within Apache Spark to pre-process the data and to calculate results. Python also has great machine learning and optimization libraries which can work with Apache Spark and makes the big data jobs faster. Hence for the 2^{nd} Chapter, for writing the library for an unsupervised HMM approach using the Maximum Likelihood method, the Python language is used. By contrast, for pre-processing to generate regular time series of observations from raw sensor data collected on the minute level resolution, and for aggregation of results after HMM implementation, the Scala language is used. All these were run using university computer clusters to process jobs faster. R data visualization packages, Tableau software and Excel are mainly used for plotting purposes after computing the resultant metrics using Scala and Apache Spark.

Apache Spark, Scala, Python Apache Spark (henceforth, Spark) is a multi-tier analytic engine from Apache Software Foundation designed for use with big data analytics that is difficult to conduct on a single computing system [17,18]. The features of Spark – such as in-memory processing, distributed data processing engine [17, 18], lazy execution, use of read-only datasets, etc. – elevate performance while computing with big datasets. The speed advantages of Spark over competing frameworks such as Hadoop for iterative-heavy big data analysis are furthered by several features of Spark, including generality, ease of use, generality in terms of work with Hadoop, or in a standalone or Cloud environment, a master-slave architecture which has an additional cluster manager between master and slave nodes which allocates the resources for the run, etc. [17,18]. In Spark, data initially read are stored in RAM instead of on hard disks, which makes it faster than widespread implementations of MapReduce. Also, the construction of Apache Spark atop of the scalable programming language Scala makes it easy to work with that flexible and general language, as well as with other languages like Python. Scala is chosen as the main language for the majority of the work presented in this thesis. Scala is well suited to work with Spark because of its higher level of abstraction, suitability for characterizing pipelines, and efficiency in processing data asynchronously in a parallel and distributed manner [19]. This latter feature helps Scala to perform the processing by running the analytics on clusters. Spark evolves quickly as a platform; because Scala serves as Spark's native language, use of Scala helps secure access to all new benefits of Spark features immediately upon release of a new version; by contrast, support for Python and tools such as R can be delayed due to the need to add such language support subsequently. This constitutes one of the central reasons that the majority of works presented here were performed in Scala. It is particularly notable that the Scala language was designed to support high-performance in large-scale analytics using high-level functional programming abstractions, including – but not limited to – support for stream-based processing, parallel map-reduce operations, and it has higher-level APIs supporting transparent and scalable mapping of computations over multiple cores and variable-size computation clusters, ubiquitous computation using higher-order functions, and monadic handling of errors and collections. Even though Spark supports both Python and Scala and both are major languages that support big data analytics and cluster computing [20], Scala has integrated features of object-oriented and functional languages [20], a strong and rich type system, and is a compiled language based on the Java virtual machine (JVM) and compiles to JVM byte code, making its direct execution faster than Python. But Python includes binding to many low-level libraries supporting data mining, scientific computing, machine learning and data visualization, and is easy to learn when compared to Scala, making it preferable for some optimization and statistical computation parts in this thesis.

The three studies covered in this thesis include large scale data collected from smartphones at a minute level resolution with the help of Ethica software. Hence, while processing the dataset and performing analysis, a cluster computing framework like Spark is highly valuable to avoid prolonged analysis execution time. Because of the above-mentioned reasons, Scala and Python are the languages used in this thesis for performing the Spark-analysis. Such tools provided key support in solving the first research problem of handling computational burden and intricacies of manipulating smartphone data in this thesis. The two other research problems motivating use of machine learning and data analytics for smartphone data are covered in the 3 main Chapters, namely Chapter 2, Chapter 3 and Chapter 4, and are noted in the next section.

1.3 Thesis organization

Within this thesis, we investigate analysis approaches for three case studies to analyze data from three different sets of vulnerable populations. Each such case study is associated with a separate study and constitutes a chapter within this thesis: Chapter 2, Chapter 3 and Chapter 4. The name and details of the main three chapters are described in Table 1.1 below. The three below cases use the same Ethica platform for data collection. They are each conducted in combination with a team of health scientists, and are focused on survey data and streamed sensor data collected from the mobile devices of study participants.

Chapter	Focus	Description	Partner
Chapter 2	Focused on sensor data	Analysis to support a study to understand teenagers screen time and app usage patterns and its association with mental health.	MEDIATICINO team, University of Lugano, Switzerland
Chapter 3	Focused on survey data.	Analysis to support a study to understand the mood change patterns and influence of risk factors on Suicidal ideation.	Dr. Rudy Bowen & Team, Royal University Hospital, Saskatoon, Saskatchewan, Canada
Chapter 4	Focused on Survey and sensor data.	Analysis to assess the outcomes of a feasibility study conducted on HIV patients.	Dr. Alex Wong & team, Saskatchewan Health Authority, Regina, Saskatchewan, Canada

Table 1.1: Thesis organization

It bears emphasis that the methods implemented in these chapters to overcome study-specific challenges have a potential application beyond the case studies considered here. For each case study, we examine each study design, configured data streams, data collection and analysis methods. In the first 2 case studies – appearing in Chapters 2 and 3 – the data collection and study set up was performed by the corresponding teams, reflecting the ease with which Ethica can be configured for custom study designs. Within these projects, work – and, by extension, the chapter contents – focused on the analysis following data collection to help support the health scientists. For the study covered in Chapter 4 – which was conducted in partnership with health scientists and system personnel based in the infectious clinic in Regina Qu'Appelle health region a part of Saskatchewan Health Authority – the work of the author involved substantial efforts exploring possible study configurations within Ethica, with many exploratory designs being investigated. As a result, the corresponding chapter (Chapter 4) discusses study design and recruitment through and including analysis.

In Chapter 2, the focus lies on the screen state sensor data – and specifically the use of such data to infer the underlying hidden states associated with the data generation process. Such inference can then support estimation of the actual screen time over study days for each participant by handling the missing latent states using the unsupervised version of the machine learning approach called Hidden Markov Modelling (HMM). This approach helped to infer the daily screen time patterns associated with each participant, and thus aided the team in effectively using the data collected through screen state sensor, such as by investigating its associations with participant characteristics, as well as different metrics derived from the survey and other sensor data, such as app usage patterns, activity patterns, self-reported aspects of mental health, etc. of the study participants. In light of the unsupervised nature of the HMM approach used in this chapter, another simulation modelling approach was also used to evaluate and better understand the limitations of that approach. Specifically, an agent-based model was used to represent the Ethica data generation process in a stylized fashion, and – through scenario simulation – thereby generate synthetic ground truth datasets to cross validate the unsupervised HMM approach.

To support this approach, Chapter 2 introduces a machine learning algorithm in the form of a Hidden Markov Model (HMM) for imputing and labelling the hidden states associated with the data generating process underlying the time series of screen state observations. In order to obtain greater insight into the etiology of the missing data, patterns of battery sensor data with a reliable regular recording interval are also helpful. Hence, the HMM machine learning approach is implemented here using two sequences of data observed from mobile sensors – screen_state and battery sensors – as inputs. These two time series are sampled at different times and exhibit a distinct structure. To effectively pass these into the model, a reliable sequence of data pre-processing steps was used to transform the collected battery and screen state sensor time series (the latter in the form of transitions) into a uniform time series format. Manually creating "ground truth" training and testing sets for screen state is highly tedious, particularly at the fine-grained temporal granularity being sought (with second-level resolution). For this reason, an unsupervised HMM approach using a Maximum Likelihood algorithm is implemented to estimate HMM model parameters and to infer the underlying states associated with the data recording process.

Secondly, considering the unsupervised nature of the original data set, validating the prediction accuracy is challenging. As a result, in this chapter, we seek to validate this system using "synthetic ground truth" data produced by another simulation modelling approach. Specifically, I implemented a system to simulate the underlying data generation states in a manner plausibly similar to how they happen in the data collection tool, and to output data similar to what is received from Ethica. This simulation is implemented as an agent based model in AnyLogic software. Such a model can then be used to evaluate the degree to which the HMM achieves accurate inference in a variety of different data generation regimes – including some closely according with the assumptions of the HMM, and some departing from such assumptions. To support this, the simulation model emits 1) A "synthetic ground truth" time series of data specifying the true underlying situation (state) obtaining in the simulation model, and 2) Two time series consisting of "synthetic" observations of battery and screen transitions, respectively, in a form similar to what is recorded from the Ethica data collection tool. To evaluate the accuracy of the HMM framework, the emitted observation sequences from the simulation model are passed as input observations to the multivariate HMM model described above, with the hidden states then being predicted by the HMM modelling process. These predictions are then cross validated with the known synthetic ground truth states. This synthetic ground truth approach helps to cross validate the accuracy of the unsupervised HMM model using ground truth data generated by the simulation model and to test the accuracy as HMM assumptions are violated. Overall, the model exhibits good accuracy in predicting the underlying states with a 1 second resolution.

Following a suitable evaluation of the HMM approach via cross validation, I used the Viterbi algorithm to derive the most likely sequence of hidden states for each study participant. This was then used to calculate the actual screen time for that participant. In reflection of different patterns in phone usage for different participants and types of phones – such as Android vs. iPhone, separate participant-specific HMM model parameterizations are used in this project, where each such model is trained using participant-specific data collected through mobile phones. While not covered here in detail, the capacity to use a computational cluster to simultaneously train and – separately – run the different HMM models using data for different participants and the use of programming languages such as Scala and Python with spark dataframes significantly reduces the time required to perform the computation.

In Chapter 3, we discuss several data analytic methods that are implemented to support the team in remedying some uncertainties in the collected data that were the result of lapses in study management. One trouble faced by the team was to find the missing start and end dates associated with the study duration of a few participants, who were distributed the same physical smartphone from the facility. In the dataset collected by Ethica software, a phone-based id was recorded for each participant as their user_id; this value was the same for all participants who had used the same phone. Several rounds of cross-checking were performed based on the dates recorded in Ethica, the admission and discharge dates of participants, etc. to label such data. The analysis also included adherence, additional quantitative and exploratory components. For example, adherence analysis is performed to secure the participants' adherence behaviour during the derived time participation windows; quantitative and qualitative analysis is then performed to support understanding the association between 4 variables self-reported by participants.

Chapter 4 analyzed the study adherence in terms of availability of survey data and GPS sensor data. Apart from the analysis, four different rounds of study configurations of Ethica were also created before study launch, based on modifications and suggestions by the study team, and as per the feedback from several patients and researchers involved. These include changes in the survey question designs and selection of sensors. Also, study interface testing and technical help during the recruitment of patients in Regina was also provided by the author to support the team with the data collection step. The main focus of this chapter is to check whether the study feasibility criteria set at the beginning of the study were met after the study, by performing only quantitative analysis of the data and – by Research Ethics Board stipulation – without looking into the content of responses. Aggregate counts of collected data – such as the fraction of surveys answered on a day within the study and participant basis – are calculated and discussed in this chapter. The adherence graphs created in this chapter and Chapter 3 are reusable, in the sense that the graphs can be regenerated in other similar studies created using Ethica interface, as a standard way to check the adherence patterns of the study participants.

In all the three case studies covered in this thesis, the role of the author was to handle the computational complexities associated with the extraction and processing of smartphone data. The methodologies implemented within the three case studies are therefore focussed on the survey and sensor data collected by the smartphones through Ethica software. In a division that was sometimes mandated by the University of Saskatchewan Research Ethics Board, for such studies, the author did not have access to the clinical databases containing patient demographics or participant details such as their age, gender, medical conditions, education details, etc., and hence analyses related to patient or participant demographics are not covered in this thesis. However, in future publications, the results or outcomes of the methodologies implemented in the three case studies will be joined with the clinical or other databases holding patient information by the corresponding research teams and will be used to arrive at insights or conclusions related to patient behavioural patterns. The manuscripts of two research projects using the results and outcomes arrived from the implemented methodologies discussed in Chapter 2 and Chapter 3, after joining with the participant demographics and results from other study databases by the corresponding research teams, were submitted for publication by the corresponding teams, with the author of this thesis included as one of the coauthors for each.

Chapter 2

Inferring smartphone screen states using a multivariate HMM approach

2.1 Background

Screentime patterns and their influence on the mental and physical health of teenagers and adolescents have been an area of interest for researchers for several years. One set of studies has focused on analyzing the impact of different factors associated with screentime usage amongst teenagers and adults. Examples include investigations of the role of parent-student interaction or the family environment [21], the popularity of gamification [22], and the impact of location-based games (e.g., Pokémon GO) [10]). Another set of studies has sought to examine the beneficial and adverse effects of screentime on the mental or physical health of teenagers or adolescents [23, 24]. Even though data collected through survey questionnaires can provide insights into participants smartphone usage, the proper usage of larger datasets acquired through automatic mechanisms reporting screen state, app usage, phone turning on or off events provide additional temporal resolution, essentially eliminate recall bias and lower reporting bias, and can better elucidate changes over time in such factors. However, to utilize the maximum potential of such data sources, careful attention must be given to the underlying data generation process of each automatic recording mechanism while processing the data, and proper pre-processing steps should be performed prior to analysis.

2.2 Introduction

This investigation sought to contribute methods to reliably estimate daily smartphone-based screen time patterns, using longitudinal time series collected from consenting teenage participants through a smartphone-based application called Ethica (introduced in Chapter 1.2.2). This application keeps track of screen state transitions occurring in participant smartphones by high-sampling-frequency recording by sensing screen state through operating system mechanisms. As in many other large-scale data collection processes, following the completion of data collection, the data sets were pre-processed to improve data quality and were then analyzed to arrive at research findings. In this study, after an initial pre-processing step, an unsupervised HMM approach was used to label the underlying hidden states, to calculate daily screen time usage patterns.

This HMM model helps to probabilistically label the hidden states associated with the screen state data generation process, including identifying time intervals in which screen state transitions may be missing due to a lapse in recording by the Ethica app; the resulting time series of inferred states can then be used to report estimates of screen time exposure and other components of smartphone use. To understand the accuracy of this implemented unsupervised HMM model, I validated it using ground truth data generated by a simulation model replicating data generation scenarios (Section 2.6 and Section 2.8.2). The data collection process, pre-filtering steps, HMM model, and synthetic ground truth simulation models are explained in detail in the sections below.

2.2.1 Data collection

The data used in this project was collected by a study named MEDIATICINO, conducted jointly by a group of researchers at the Institute of Health Communication at the Universitá della Svizzera Italiana (University of Lugano) in Ticino, Switzerland, and researchers at the University of Saskatchewan in Canada. This study was approved by relevant Research Ethics boards in both institutions; It was supported by the Swiss National Science Foundation (Grant no. 175874) and the University of Saskatchewan Behavioural Ethics Review file number for the study is ID 39. The author of this thesis was added as a student working on this project in the REB file associated with this study and worked closely in collaboration with members of the larger study team. The smartphone-based data collection tool called Ethica was installed on all participant smartphones to collect data continuously throughout the study period. The primary purpose of the MEDIATICINO study is to understand the role of social media and mobile phone usage in the development of young people in Switzerland. This wave of the MEDIATICINO study collected sensor and survey data from 100 participants for around 45 days. However, screen state sensor data is available only from 94 participants; hence, data from 94 participants are analyzed in this chapter. All of these participants are assenting teenage boys or girls 13 to 14 years of age, whose parents previously provided consent. The HMM approach was used to derive a daily estimate of the duration time that the smartphone screen was on and off for each participant, throughout their study period. The results were then used as part of a broader analysis – lying outside the scope of this chapter – that compares these results with other sensor and survey responses, to secure additional insights concerning mobile usage patterns and mental health associations. While the analysis conducted by the broader team together with the author for this study extended to the derivation of detailed findings regarding behaviours and exposures, this chapter focuses specifically on the methodological contributions regarding the HMM-based machine learning algorithm created and applied by the author to estimate screen time exposure.

As introduced in Chapter 1.2.2, Ethica, the data collection tool installed on participant smartphones for this study, is a sophisticated and versatile data collection platform used by a large number of research studies around the world, supports approximately 23 sensors [25] and offers study-specific custom surveys to collect data. In the investigation that forms the basis of this chapter, the focus lay on the screen time behaviour associated with the screen state sensor data collected from participants; hence, only data from the screen state and battery sensors are used for this analysis. Battery sensors are used along with the screen state sensor because of the regular nature of data collected by the battery sensor, which is set to record on every 5-minute interval. From the time that Ethica is installed on participant smartphones until the conclusion of their time in the study, the screen state transitions happening in the phone were recorded by Ethica – subject to interruptions described below – with that data subsequently being uploaded to Ethica servers and stored in a No-SQL Cassandra database for analysis.

One naive approach to calculating the participant's total screen usage time on the device during a day could consist of totalling up the accumulated time duration separating screen-turning-on and screen-turning-off events. But this calculation of screen time is affected by certain additional factors that need to be considered in a reliable estimation procedure, which are described in detail in the next section.

2.3 Problem description

A key latent distinction associated with data collection is that between data recording and data non-recording states. In the data recording state for this study, the Ethica app records every event or transition of screen state sensors; every roughly 5 minutes, it will – further – record battery sensor data. However, in the case of the non-recording states, the Ethica app data collection was non-functional; hence, the screen transitions and battery observations will not be captured, resulting in missing observations. This section discusses several scenarios that can trigger transitions between such recording and non-recording states.

The data collection process of the Ethica app is a near-continuous process underway whenever the app was running, whether in the background or foreground. But this process could be interrupted due to several reasons. Common scenarios include the following:

- Data collection is manually paused through proactive participant requests to the application by pressing Ethica's built-in "snooze"/"pause" button.
- The Ethica application is transiently unloaded from memory by Android/iPhone. Frequent reasons for such unloading are to free up memory for other applications, to reduce computational load on the phone, and to conserve battery energy.
- There is a sudden endogenously caused shutdown of the phone – for example, due to battery depletion.
- The user elects to shut down or restart the phone – for example, to conserve power when going off-grid for a prolonged period.
- While not a primary cause of an interruption in data recording, it bears noting that such a disruption can be prolonged by several factors, including continued heavy use of other applications on a phone that is currently short of memory, or a delay in the auto restart of the app, following a phone restart.

A high-level characterization of the sequence of underlying states and observed events during the data collection period is depicted in Figure 2.1 below.

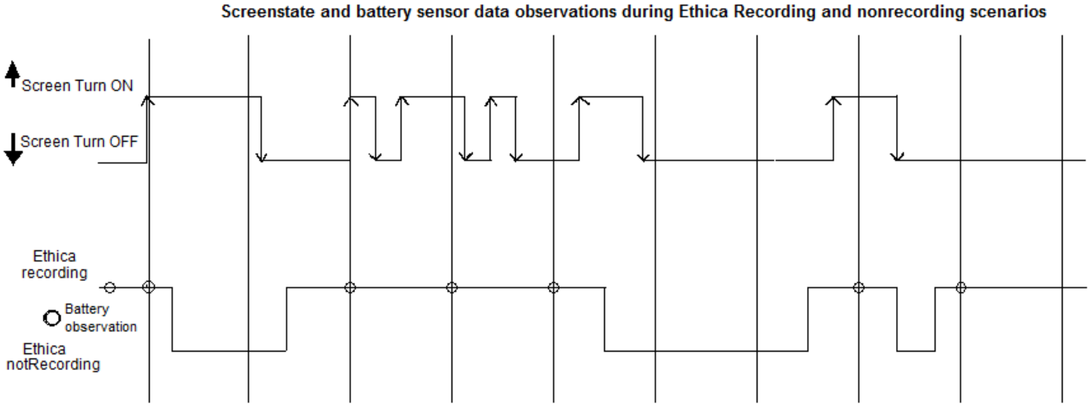

Screenstate and battery sensor data observations during Ethica Recording and nonrecording scenarios
Screen Turn ON
Screen Turn OFF
Ethica recording
Battery observation
Ethica notRecording

Figure 2.1: Screen state and battery observations during hidden Ethica recording and non-recording states

Figure 2.1 provides an illustration of a concrete scenario to illustrate these processes and their relation to one another. The top portion of the graph in Figure 2.1 above represents the underlying phone hidden states associated with the phone screen, namely intervals during which the screen is on and (by contrast) off. In this section of the diagram, the figure represents events in which the phone turns on as an upward arrow and the screen-turning-off events as downward arrows. No missing transition events were present in this top section of the diagram, because it depicts the real underlying hidden states of the system. By contrast, the bottom section of Figure 2.1 represents the (also not directly observable) recording and non-recording states of the Ethica data collection process, with a higher value of the line indicating that recording is taking place, and a lower value indicating that recording has stopped; it bears noting that ○ indicating battery observations are only present when the line is at a higher level – indicating the presence of ongoing recording. An example of a missing record scenario was depicted in the second time slot of Figure 2.1, when there was a non-recording interval in effect from the middle of the first time slot to the middle of 2^{nd} time slot, censoring one of the actual phone screen turn off events (as depicted by the downward arrow in the top section of the diagram). As per the figure, the next screen turn off event during active data collection happened in the 3^{rd} time slot, which could result in misinterpretation of screen time assumptions if we depend naively on considering the period between a screen-turning-on and its next screen-turning off event as indicative of a period when the screen was on.

To conclude, lapses in the data recording process – as depicted in Figure 2.1 – can censor measurements of observations, and omission of direct information on screen state transitions. The resumption of data collection can record further screen state transitions, which could be naively misinterpreted as a longer interval of invariant phone screen state. Also, the lapses in recording are of different lengths, and the uncertainty as to in what underlying screen state state a phone is present will grow with rising time since the recording was stopped.

Even though the chances of Ethica stopping recording and missing transitions within a given minute are low, the frequency of such censoring depends on the type and vintage of the smartphone, the aggressiveness of concurrent use of apps by the user, the phone battery levels, a user's attitudes and habits with respect to phone battery conservation measures, and several other factors which differ from participant to participant or from phone to phone.

In a naive approach, the difference between a record_time of 2 consecutive paired recordings in which the first is true (turning on) and the second is false (turning off) could be misunderstood as implying that the screen is on for the entirety of the difference in times between the pairs. Similarly, the elapsed duration of time between a successive false (turning off) and true (turning on) event could be mistaken as an interval in which the screen is continuously off. While it is tempting to come to such conclusions, the validity of the inferences is far from guaranteed: The underlying data generation process has a hidden aspect of state associated with it – whether data is being recorded – which is not observed. As a result, the data recording process could have been stopped or interrupted without showing any signs in the observed screen state records. As a result, observing consecutive true and false observations does not imply a continuous period of uniform screen state because the underlying state of the hidden data recording process between this pair of transitions is not observed. The perils of assuming that there is continuous recording between consecutive pairs of screen state observations are indicated by the fact that, in some cases, two consecutive observations indicate the same screen state transitioning – with 2 'true' entries or 2 'false' entries, with no interspersed entries of another sort. The occurrence of such a pair shows that there was a sojourn to a non-recording hidden underlying state that took place between these transitions. However, as long as the recording and non-recording component of the hidden state is not labelled, the missing screen transitioning events cannot be identified, and the interpretation of screen time based on observations are unreliable. This non-recording state sometimes happens due to restarting of data collection tool – sometimes after eviction of a program for the sake of using resources, and sometimes after the phone itself is restarted. Once the tool restarts, the condition of the screen at that particular restart time will be entered into the database as a screen turning on observation, which is indistinguishable from the other typical entry of screen turning on or turning off events. A viable solution to infer the actual screen transitions observations lies in labelling the unlabelled hidden states associated with the underlying data recording process, which is implemented in this chapter. After labelling the hidden states, the daily screen time duration for all users can be estimated.

Initial examination of the data from Ethica studies suggests that lapses in data recording cannot be safely ignored for researched studies seeking to investigate the association between smartphone usage and participant behavioural patterns, lower socioeconomic status, or mental health, because the associated censored intervals may materially impact results and misleads the actual scenarios or participants phone usage patterns – with some of the populations of greatest interest and health risk potentially having their data censored the most heavily. Based on these considerations, a methodology to robustly infer the underlying screen state was of great importance, and was explored, implemented, and evaluated in this chapter.

As emphasized above, the pattern of phone usage and phone type typically differs across participants. Hence, the inference model needs to support individual-level adaption, with the parameters of model being trained on a per-participant basis to sidestep bias that could be imposed through the use only of a single model for all participants; in this regard, it bears noting that the very long time series associated with participants will help supports such per-participant estimation. The underlying process depicted in Figure 2.1 shows that training a model capable of recognizing hidden state patterns from the observed sequence (battery and screen events) can aid in finding a solution to this type of problem.

2.4 Pre-processing and filtering

Several pre-processing steps were performed to filter and prepare data prior to HMM training and testing. Firstly, the screen state sensor data was extracted from the Cassandra database in a way that filters out non-participants (e.g., study team members). The screen data source mentioned above consists of six different fields, namely, user_id, date, device_id, record_time, timestamp and state [25]. For this analysis, we selected just three columns of interest from the screen state data source: user_id, record_time and state. The user_id was the (anonymous) Ethica unique id assigned to each participant, record_time was the time at which a screen state transition occurred, and the state column specifies which of two types of screen state transition events was observed at that time. If the state column was marked as 'true', then it represents an event in which the screen was either turning on, or recording was resuming when the screen was already on. Similarly, a 'false' entry represents an event in which the screen is turning off, or when recording is resuming when the screen is already off.

The other sensor used for HMM observation was the battery sensor, which was scheduled to record every 5-minute interval. Standard data for the Ethica battery data source included eleven columns in the original table. Because the battery data source is reliably recorded on a regular basis and is standardized across phone models, whether a battery record was present or not at a particular time slot was used to check if in order to provide information as to whether any interruption or missing observations happened in the data recording, as recorded across a 5 minute interval. It bears noting that because the battery sensor can be measured with minimal additional power consumption, and is generally associated with fewer privacy concerns than other sensors, it offers a favourable source of data for many studies. However, to address the needs for this investigation, only two columns were needed for HMM purposes – user_id and record_time. After pre-processing steps, the data consists of a single time series of uniformly-spaced, time-binned pairs of screen state and battery sensor observations, where each element of a pair indicates either the value of the observations or its absence.

The main steps of pre-processing are mentioned in the below points:

- Step 1: Firstly, the screen state sensor data table table_1 and battery sensor data table table_2 were created after extracting the data from Cassandra. Step 1 consisted of filtering out of non-participant data. Non-participants were researchers who performed testing for the application during the study design phase, to evaluate the working condition of the application. To filter out the non-participant data, a complete list of all Ethica participant ids of study participants were shared by the Swiss research team. This list is used to filter out non-participant data. Step 2 to step 8 below are performed on a participant-specific basis, by iterating through each participant's corresponding dataset, with the participant id being used to join the battery sensor data and other data tables for that participant.

- Step 2: A new participant-specific dummy table_3 was created, with records quantized into regularly-spaced bins of duration 1 second. This participant-specific data began from the first time that a screen state transition for that participant was recorded in the Cassandra database for the study, until the last time that the screen state transition was recorded for that participant. For the creation of this table, only the maximum and minimum record_time of each participant's screen state observations were used and the *explode* method of the Spark data frame was used to create rows quantized into 1 second intervals between the first and last record_time. In addition, a new column called "time slot" – which marks the 1 second timestamp labels for every record – was added. Each such 1-second interval is henceforth termed a "time slot", and the column correspondingly carries that name. This time slot column was the key column of the dummy table_3 that was used for further linkage (join) steps. table_3 has only 2 columns – Ethica id, and time slot specifying the corresponding to the 1 second interval.

- Step 3: For the case of screen state sensor data, if more than 1 transition occurs within a single 1 second time slot, then the final transition event was chosen. This analysis relied on the assumption that there was little valuable information to be obtained by considering screen state dynamics within a time slot of <1 second duration. For example, if the final transition reported in a 1 second time slot is screen off transition, then all previous screen on transitioning events within that time slot was ignored. But if the final transition of a time slot is true, then that time slot will be marked as a screen turn on event recorded time slot, so that the following time slots will be used for calculating the duration for which the screen was on. This rule is implemented by sorting and ranking the records within each 1 second time slot.

- Step 4: Based on the type of screen state transition observed, time slots containing screen state observations were correspondingly labelled as SS_True or SS_False.

- Step 5: Similarly, in the case of the battery, there are many time slots containing no battery observations. By contrast, sometimes multiple recorded entries occur within a given 1-second time slot. These two situations were handled on a per-participant basis by dichotomously aggregating the record count within 1-second time slots. If the count of battery events observed within a 1-second time slot was ≥ 1, then the battery observation for that particular time slot was marked as Bat_Present; time slots lacking any battery observations were marked as Bat_Absent.

- Step 6: The participant-specific dummy table_3 from step 2 was then linked (joined) with the screen state table modified in step 4 and battery table modified in step 5 records, based on the “user_id” and “time slot” columns. These columns were created after converting the record_time to 1-second timestamps. The table resulting from joining these 3 tables consists of a modified participant-specific table with 4 columns – Ethica id, time slot, Bat_Status (battery observation) and state_SS (screen state observation).

- Step 7: While the time between transitions in screen state on a given participant’s phone will vary, most 1-second intervals will contain no observations of screen state transitions. For example, if one screen transition happens in the 1^{st} second and the next transition happens only on the 10^{th} second, then all time slots between these two events thus far lack screen state observations; hence those time slots are associated with a null value in the data received from the previous step. Reflecting that, during this step, time slots that were missing screen state data (as given by the state_SS column) were updated to read SS_Absent.

- Step 8: Finally, the resultant participant-specific data set has 4 columns being saved for HMM implementation.

The final data table emerging after pre-processing has 4 columns – namely, user_id, ts_dummy, Bat_Status and state_SS. These indicate the user_id of the participant, the 1-second time slot in which the observation was recorded, the battery status indicating whether any battery observation took place within that time slot, and, finally, the screen status column which indicates the final (if any) of the screen transitions observed during that time. This data set then serves as a time series of observations for training the HMM model.

2.5 System description & methodology

It was noted that the transition between the phone recording and non-recording states occurs in a process with a distinct structure. The data recording process of Ethica is a technical system that first enters the recording state, then enters non-recording state upon eviction from memory, rebooting, or other triggers. This system represents an unfolding process transitioning over time between underlying states that cannot be readily observed, and thus fits naturally into a domain modelled by an HMM. Hence, in this chapter, HMM was chosen as the machine learning approach employed for inference. It bears emphasis that while analysis conducted using HMM algorithms are frequently used to infer the structure of an HMM model, within this work, in this case, strong existing theory existing concerning the structure of the underlying Markov chain, and analysis for the HMM focused purely on estimating the Markov model transition probabilities. In light of the lack of fine-grained data concerning the underlying state of the phone over time, the HMM was trained in an unsupervised fashion using data collected by Ethica using observations from battery and screen state sensor as the features of the model, and several underlying states associated with recording, non-recording and transitional stages as the hidden states.

Even though the four major hidden states associated with this model were Ethica recording or non-recording states when the screen is either in an ON or OFF states – states represented as SR, $\bar{S}$R, S$\bar{R}$, $\bar{S}\bar{R}$ in the model – there are an additional set of four ephemeral states also represented. These reflect the fact that while screen transition events are associated conceptually with *transitions* between states, traditional HMMs are limited to emitting observation based on presence within a state – and not based on undergoing a specific transition. To accommodate this limitation, beyond the four major states above, there are four ephemeral states within the HMM implemented so as to express the transient contexts in which Ethica screen transition events occur. Specifically, two additional states — SOnR and SOffR – were created to indicate the hidden states associated with the screen turning on event and screen turning off events (respectively) while Ethica is recording. Two other events – ROnS and ROn$\bar{S}$ – are designed to capture the fact that screen state events are also issued during the restarting of Ethica. In order to consider the transient nature of these instrumental states – purposefully designed specifically for the purpose of causing occurrence of screen transitioning events at the appropriate times, rather than representing any persistent phone status – a minimal residence time of 1 time slot was chosen for each. This results in a multivariate HMM model with 8 states and 2 features representing the data recording process of Ethica. The theory-based structure of the HMM supports inference of the underlying state in different phases of the data collection process. Detailed model descriptions and steps for implementation of the algorithm are explained below.

2.5.1 HMM topology & training

In Hidden Markov Models (HMMs), the emission distribution depends on (i.e., is conditioned upon) the state of the underlying Markov process [26]. Maximum likelihood estimators are commonly used to estimate transition probabilities associated with the hidden Markov chain, with the model's goodness of fit being evaluated using the model likelihood [27].

As noted above, the HMM model considered in this chapter was implemented with 8 different states; in other words, the model posits that, at any time slot t, the system represented will be in exactly one of eight underlying (and hidden) states: SR, $\bar{S}$R, S$\bar{R}$, $\bar{S}\bar{R}$, SOnR, SOffR, ROnS or ROn$\bar{S}$, as described in Table 2.1.

On every time slot (here, a 1-second interval), the system can undergo a change of state or remain in the current state. The general architecture of the HMM incorporating all knowledge about the system and instantiated with 1-second time slots is depicted in Figure 2.2 below.

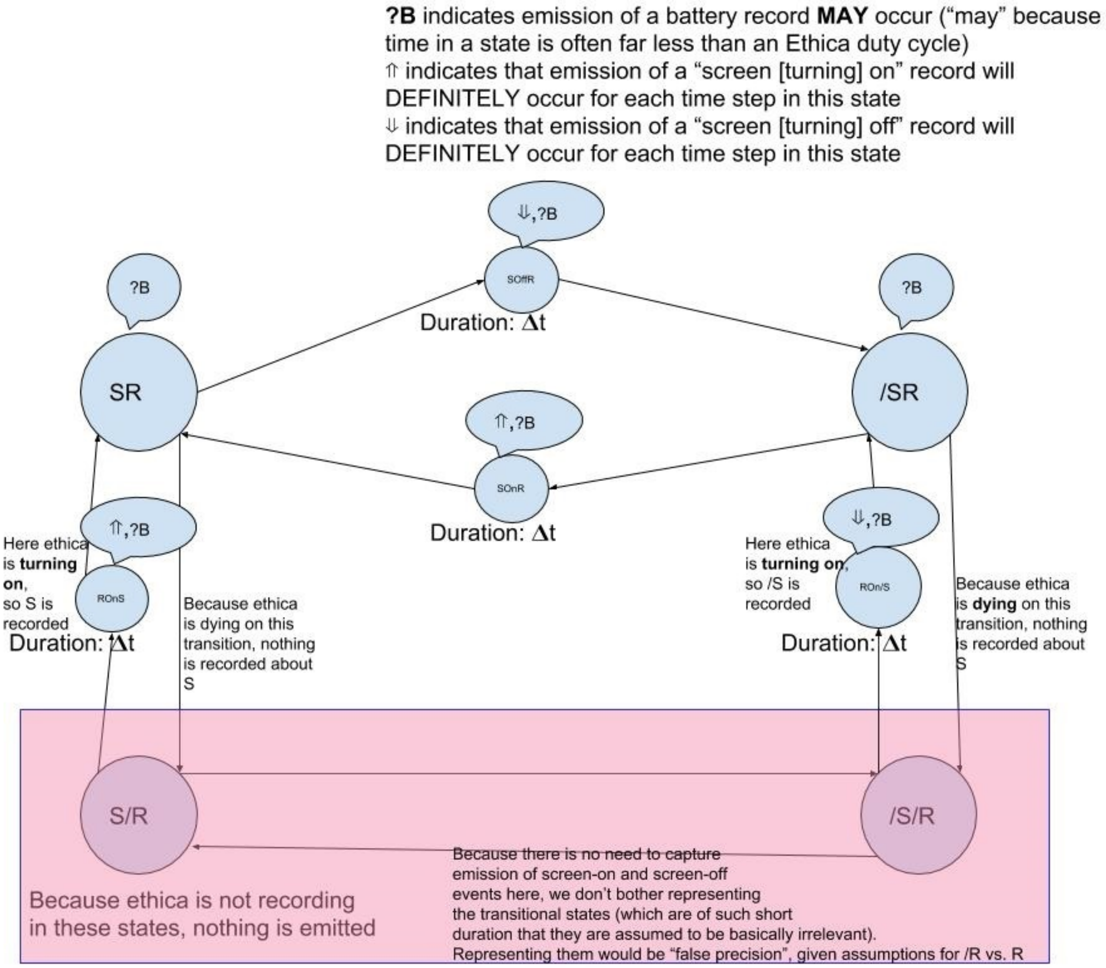

?B indicates emission of a battery record MAY occur ("may" because time in a state is often far less than an Ethica duty cycle)
⇑ indicates that emission of a "screen [turning] on" record will DEFINITELY occur for each time step in this state
⇓ indicates that emission of a "screen [turning] off" record will DEFINITELY occur for each time step in this state
⇓,?B
SOffR
Duration: Δt
?B
SR
?B
/SR
⇑,?B
SOnR
Duration: Δt
⇑,?B
Here ethica is turning on, so S is recorded
ROnS
Duration: Δt
Because ethica is dying on this transition, nothing is recorded about S
⇓,?B
Here ethica is turning on, so /S is recorded
ROn/S
Duration: Δt
Because ethica is dying on this transition, nothing is recorded about S
S/R
/S/R
Because there is no need to capture emission of screen-on and screen-off events here, we don't bother representing the transitional states (which are of such short duration that they are assumed to be basically irrelevant). Representing them would be "false precision", given assumptions for /R vs. R
Because ethica is not recording in these states, nothing is emitted

Figure 2.2: A general architecture of the HMM model with states as nodes and transitions as edges

The 8 state names are described as per Figure 2.2 in Table 2.1 below.

Symbolic state name	State description
SR	SCREEN state is ON, phone is RECORDING
$\bar{S}$R	SCREEN state is OFF, phone is RECORDING
S$\bar{R}$	SCREEN state is ON, phone is NON-RECORDING
$\bar{S}\bar{R}$	SCREEN state is OFF, phone is NON-RECORDING
SOnR	SCREEN turning ON while RECORDING is ON
SOffR	SCREEN turning OFF while RECORDING is ON
ROnS	RECORDING turning ON while SCREEN state is ON
ROn$\bar{S}$	RECORDING turning ON while SCREEN state is OFF

Table 2.1: Description of 8 states name as per Figure 2.2

As mentioned, there were 8 states ($N = 8$) for this HMM model. Of these, there were six states associated with the recording state of the Ethica app data collection, and 2 other states – S$\bar{R}$ and $\bar{S}\bar{R}$ – associated with the non-recording mode of Ethica. The time slot of the HMM was set to be 1 second (Δt) – far less than the standard 5-minute duty cycle of Ethica. It was advantageous to have this Δt as small as possible so that the transitional states associated with the transition event from the screen off($\bar{S}$) ⇔ Screen on (S) (i.e., SOffR, SOnR) would be as short as possible. During an interval with multiple screen state transitions, if the final transition happened within that time slot indicated the screen turning on event, then the screen state transition was marked as turning on. As noted above, the indications of the screen turning on or the screen turning off (specifically) were indications of *transitions*, either in terms of screen state, or in terms of recording. When the recording was OFF (i.e., R was false), then the hidden state can be changed without such observations occurring (and without an intervening transient state), but if this transition occurred in the midst of a time where the recording was ON, there was assumed to always be an indication of screen state transition emitted by the Ethica app. More details about the state sequences and the associated transition probability assumptions for the model are mentioned in the parameter initialization subsection below.

An important challenge associated with this model was the absence of labelled empirical data for model training or validation. Here, the only empirical data available to ground the model were the participant-specific pairs of time series observations – and for which the underlying state sequence was hidden. Since there was no empirical training data available, this chapter implemented an unsupervised HMM approach to model estimation, using a maximum likelihood algorithm. For each participant separately, and for their entire studied duration, two time series of screen state and battery observations were pre-processed as above, and saved as two time series sharing contemporaneous 1-second intervals. Even though there were no labelled states available, we had a pair of observation sequences for every time slot t, and had assumptions about the model transition matrix, emission matrix, and row vector of initial probabilities, denoted as parameters A, B and δ, respectively. According to [28], 3 basic problems of an HMM model λ = (A, B, δ) should be solved to make it useful for real-world applications [28]. Firstly, given the observation sequence x and the model parameters λ, it is necessary to efficiently compute the probability of the observation sequence given the model – i.e., to compute $P(x|\lambda)$ [28]. This likelihood computation problem can be handled by using the vectors calculated by matrix α calculated by the Forward pass of the Forward-Backward algorithm i.e., $P(x|\lambda) = \sum_{i=1}^{N} \alpha_t(i)$. This probability, which constitutes the likelihood, serves as the objective function of the maximum likelihood estimation explained later in this chapter. The second problem is to find a most probable hidden state sequence Q = $q_1, q_2, .., q_T$, given the observation sequence (x) and model parameters λ, [28]. I used the Viterbi algorithm to identify Q. In addition to that, the Forward-Backward algorithm is also implemented in this chapter to calculate the posterior marginals of all hidden state variables. The results are explained in the result section (Section 2.8). Finally, the 3^{rd} problem to be addressed was to employ a method that could adjust the model parameters λ to maximize the likelihood $P(x|\lambda)$ [28]. In this chapter, we use maximum likelihood estimation in the training step of the HMM to optimally adapt the model parameters to the entire sequence of observations, which serves as the training data.

Given the HMM formulation, the most important step in this HMM modelling work lay in finding the solution to problem 3 – estimating participant-specific values for the model parameters using the maximum likelihood algorithm. An important sub-step lay in the solution to problem 1, – the computation of $P(x|\lambda)$, the probability that the observation sequence x was generated by the model; as noted above, this was addressed by using the Forward part of the Forward-Backward algorithm [28]. After each of the previous steps was completed, the third important step lay in providing a solution to problem 2, using the Viterbi algorithm, so as to find the single most likely state sequence, Q $= q_1, q_2,$ ..., q$_T$, for the given observation sequence x $= x_1, x_2,$ ...,x$_T$ [28]. Overall, in order to achieve solutions to the above problems, four main steps were implemented in this work:

- Initialization of model parameters.
- Performing maximum likelihood estimation of model parameters in λ using optimization.
- For the possible value of λ examined during the optimization in the previous step, compute posterior probabilities of all hidden state variables for the sequence of observations using the Forward-Backward algorithm.
- Using the model parameter estimates identified in the maximum likelihood, decoding or predicting the most likely state sequence (Q) using the Viterbi algorithm.

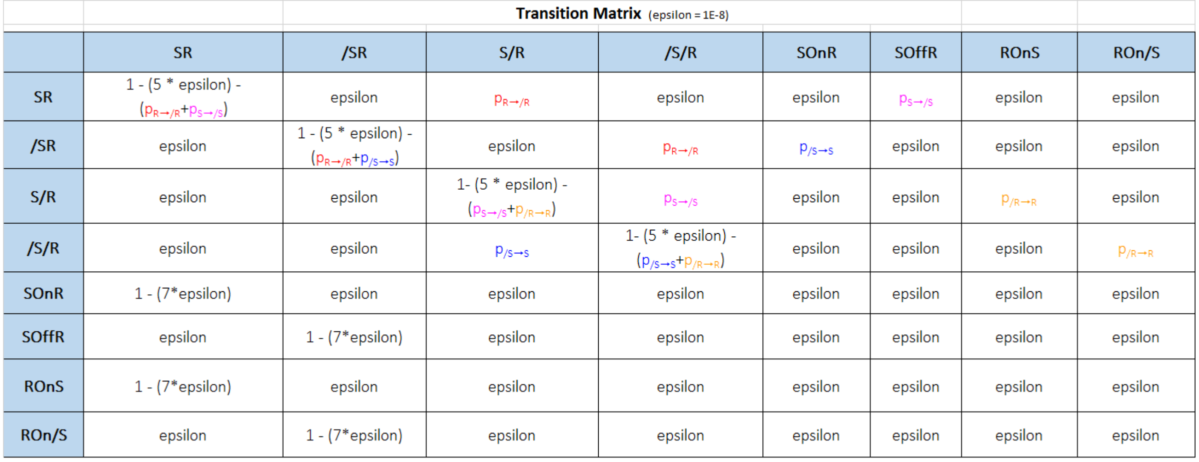
Initialize model parameters

The model parameters, represented by λ, included 2 matrices – an observation likelihood sequence B, initial state probability distribution δ, and transition probability matrix A.

Emission probabilities, B $=$ b$_i(x_t)$ is an observation likelihood sequence, representing the per-time slot probability of observing (“emitting”) an observation x_t when the system is in state i.

Initial states probability distribution (δ), where each δ_i is the probability that the system starts in that particular state i. As a distribution, the sum of the initial probabilities over the set of states must equal 1: $\sum_{i=1}^{N} \delta_i = 1$ [29]. In the case of screen state sensor recording for most of the participants, the first screen state observation recorded is screen turn off event, hence for the initial probability vector (δ), the probability corresponding to state SOffR is updated as (1-(7 epsilon)); for all other states, the initial probability is updated as epsilon. Initial probability vector $\delta =$ [epsilon, epsilon, epsilon, epsilon, epsilon, (1 - 7*epsilon), epsilon, epsilon] , where epsilon = 1E-8.

In transition probability matrix (A), each entry a_{ij} represents the per-time slot probability of moving from state i to state j, where $\sum_{j=1}^{N} a_{ij} = 1 \forall i$.
Here, we had in total 8 different states, resulting in an 8 × 8 transition probability matrix. The transition matrix used in the model is depicted in Figure 2.3 below, where the values of the symbolic parameters shown (P_SonSoff, P_Soff_Son, P_Ron_Roff, P_Roff_Ron) were estimated by maximum likelihood estimation.



Transition Matrix (epsilon = 1E-8)

	SR	/SR	S/R	/S/R	SOnR	SOffR	ROnS	ROn/S
SR	1 - (5 * epsilon) - ($p_{R\rightarrow /R}$+$p_{S\rightarrow /S}$)	epsilon	$p_{R\rightarrow /R}$	epsilon	epsilon	$p_{S\rightarrow /S}$	epsilon	epsilon
/SR	epsilon	1 - (5 * epsilon) - ($p_{R\rightarrow /R}$+$p_{/S\rightarrow S}$)	epsilon	$p_{R\rightarrow /R}$	$p_{/S\rightarrow S}$	epsilon	epsilon	epsilon
S/R	epsilon	epsilon	1- (5 * epsilon) - ($p_{S\rightarrow /S}$+$p_{/R\rightarrow R}$)	$p_{S\rightarrow /S}$	epsilon	epsilon	$p_{/R\rightarrow R}$	epsilon
/S/R	epsilon	epsilon	$p_{/S\rightarrow S}$	1- (5 * epsilon) - ($p_{/S\rightarrow S}$+$p_{/R\rightarrow R}$)	epsilon	epsilon	epsilon	$p_{/R\rightarrow R}$
SOnR	1 - (7*epsilon)	epsilon	epsilon	epsilon	epsilon	epsilon	epsilon	epsilon
SOffR	epsilon	1 - (7*epsilon)	epsilon	epsilon	epsilon	epsilon	epsilon	epsilon
ROnS	1 - (7*epsilon)	epsilon	epsilon	epsilon	epsilon	epsilon	epsilon	epsilon
ROn/S	epsilon	1 - (7*epsilon)	epsilon	epsilon	epsilon	epsilon	epsilon	epsilon

Figure 2.3: Transition matrix of HMM model

In order to reduce the search space for the maximum likelihood optimization step, we incorporated the existing theory about the underlying Markov process in estimating the parameters. The above transition matrix in Figure 2.3 contains a total of $8 \times 8 = 64$ per-time slot probabilities associated with transitioning from each of the 8 states to all other 8 states. i.e., for any of the 8 hidden states that the hidden variable can be at time t, there is a transition probability associated with it to transition from that state to any of the 8 states at time $t + 1$, for a total of $N \times N$ transition probabilities, where N is the total number of states. But from any particular state, the sum of the set of outgoing transition probabilities must sum to one. Hence, if 7 other transition probabilities of a state are known, then the 8^{th} probability must be one minus the sum of the others. There are thus a total of $N \times (N - 1)$ transition parameters that have to be figured out for the transition matrix – here, 8×7, or 56 such probabilities. But based on the below defined assumptions, several state transition probabilities were derived, resulting in a situation where the value of just four distinct parameters – P_SonSoff, P_Soff_Son, P_Ron_Roff and P_Roff_Ron – needed to be determined to completely specify the transition matrix. The assumptions relied upon when specifying the transition matrix are specified – and generally explained – below.

- As per Figure 2.2 and the transition matrix in Figure 2.3, there are two types of screen states: screen on and screen off, which can each happen within either the recording and non-recording states of Ethica. P_SonSoff – symbolically represented as $P_{S->\bar{S}}$ in Figure 2.3 – is the probability of transitioning from "screen on" states to "screen off" states. But for the Ethica recording state, there are 2 short additional instrumental states added: SOnR and SOffR. These states are transitional states designed to express the occurrence of a transition itself so as to recognize the associated event, and are therefore associated with- a minimum length duration – a duration guaranteed to be just a single one time slot in length. The transition probability $P_{S->\bar{S}}$ is updated as the probability of transitioning from SR to the ephemeral state SOffR. But in the case of non-recording states of Ethica, reflecting the fact that no observations need to be captured, there is no need for states similar to SOnR or SOffR. Moreover, I made the notable assumption that the per-time slot probability of putting down the phone (thus, turning off the screen) is similar regardless of whether Ethica is in a position to record or not. Hence, the probability $P_{S->\bar{S}}$ is treated as applying while transitioning from $S\bar{R}$ to $\bar{S}\bar{R}$.

- Similar to the above, the P_SoffSon symbolically represented as $P_{\bar{S}->S}$ in Figure 2.3 represents a probability of state transition from screen off to screen on. In the case of recording states of Ethica, the transition probability $P_{\bar{S}->S}$ serves as the probability of transitioning from $\bar{S}$R state to the ephemeral SOnR state. Similarly, in the case of non-recording state of Ethica, the probability $P_{\bar{S}->S}$ is updated while transitioning from $\bar{S}\bar{R}$ state to $S\bar{R}$ state.

- Another pair of transition probabilities that play an important role are P_Ron_Roff and P_Roff_Ron. The probability P_Ron_Roff is symbolically represented as $P_{R->\bar{R}}$ in the transition matrix in Figure 2.3, and is the per-time slot probability of transitioning from the Ethica recording state to non-recording state. Significantly, this probability was assumed to be the same regardless of whether the screen state was on or off. Thus, the probability $P_{R->\bar{R}}$ was assumed to represent the probability of transitioning from state SR to $S\bar{R}$ state, and from state $\bar{S}R$ to $\bar{S}\bar{R}$ in the transition matrix.

- Similarly, P_Roff_Ron – symbolically represented as $P_{\bar{R}->R}$ – is the transition probability from Ethica non-recording state to Ethica recording state, and was assumed to apply to characterize the probability of transitioning both from state $S\bar{R}$ to ROnS state (on the one hand) and from state $\bar{S}\bar{R}$ to state $ROn\bar{S}$ state (on the other).

- We know from the design in Figure 2.2 that screen turning on transition and Ethica recording turn on is assumed to occur only in the above mentioned state transitions: $SOnR$, $ROnS$ and $ROn\bar{S}$. Hence, all other transition probabilities were assumed to be associated with a value close to 0 (epsilon = 1E-8), with a non-zero value being imposed to protect the HMM from underflow problems.

- Four states offer the possibility of a self transition – a transition from that state to itself: SR, $\bar{S}R$, $S\bar{R}$ and $\bar{S}\bar{R}$. Recall that each row in the transition matrix represents the probability of transitioning from one particular state to all of the other states. Hence, the sum of the transition probabilities across a row is always equal to 1. Taking advantage of that constraint, the self transitioning probabilities for the aforementioned states are updated by subtracting from 1 the sum of the other (known) probabilities.

- By contrast, self-transition probabilities are considered impossible for the transient states in the design: SOnR, SOffR, ROnS and $ROn\bar{S}$; for such cases, this value is zero. But to protect the HMM from underflow problem mentioned above, this value was correspondingly assumed to be epsilon (1E-8).

The above section discussed the the assumptions and probabilities associated with the transition matrix, whose corresponding values are shown in Figure 2.3. The assumptions applied with respect to the likelihoods of making observations are explained below, and their estimation is discussed. As noted above, because observations for HMMs are limited to being conditionally dependent on the current state – not transition occurrence – transitional states SOnR, SOffR, ROnS, and ROn$\bar{S}$ (Figure 2.2) were put in place to express emission of screen turning on and screen turning off observations, specifically when Ethica is in a recording state; specifically, this occurs when Ethica is either switching between screen states while recording, or when (re-) initiating recording. Such ephemeral states are always associated with the emission of an observation of the screen turning on or off.

- For states S$\bar{R}$ & $\bar{S}\bar{R}$ (explained in Table 2.1), because there is no recording taking place, no observations of any sort can take place. Thus L($\Downarrow$|S$\bar{R}$) = L($\Downarrow$|$\bar{S}\bar{R}$) = L($\Uparrow$|S$\bar{R}$) = L($\Uparrow$|$\bar{S}\bar{R}$) = L(B|S$\bar{R}$) = L(B|$\bar{S}\bar{R}$) = 0. No other emissions probabilities will be mentioned for these states.

- Knowledge of Ethica operation suggests that all emissions of $\Downarrow$ or screen turn off events occur only when either transitioning from S $\rightarrow$ $\bar{S}$ via SOffR or during $\bar{R}$ $\rightarrow$ R (when the screen was at the off state when Ethica (re)initiated recording) via ROn$\bar{S}$. For such states, the emission is guaranteed to occur, and thus L($\Downarrow$|SOffR) = L($\Downarrow$|ROn$\bar{S}$) = 1. For all others, it is treated as having no chance of occurrence, and thus L($\Downarrow$|SR) =L($\Downarrow$|$\bar{S}$R) = L($\Downarrow$|SOnR) = L($\Downarrow$|ROnS) = 0.

- Reasoning similar to that characterized in the previous item also holds for the screen turning on event. We know that all emissions of $\Uparrow$ (or screen turn on event) should occur only in transitioning $\bar{S}$ $\rightarrow$ S via SOnR or $\bar{R}$ $\rightarrow$ R (when screen was on when Ethica (re)initiated recording) via ROnS. For such states, emission is guaranteed to occur, and thus L($\Uparrow$|SOnR) = L($\Uparrow$|ROnS) = 1. For all others, it is treated as having no chance to occur, and thus L($\Uparrow$|SR) = L($\Uparrow$|$\bar{S}$R) = L($\Uparrow$|SOffR) = L($\Uparrow$|ROn$\bar{S}$) = 0.

- In the case of battery observations, it was assumed for simplicity that battery records were received according to a memoryless stochastic process (at each time slot independently). And regarding the probabilities of battery observations in the recording states, in accordance with the memoryless assumption, we assumed that in all of these 6 recording states, each time slot of length Δt will have a probability of occurrence according to the ratio between Δt and 5 minutes, where 5 minutes was the typical length of a duty cycle in Ethica (i.e., the duration of Ethica recording epochs). For example, we assumed that if Δt were 2 minutes, then the likelihood of observation within a 2 minute period would 0.4 (2/5). i.e., L(B|SR) = L(B|$\bar{S}$R) = L(B|SOnR) = L(B|SOffR) = L(B|ROnS) = L(B|ROn$\bar{S}$) = Δt/(5 min)

- To lower the risk of singularities in computation of probabilities, we assumed that the probabilities of some emission that were in theory impossible were instead associated with a very minimal chance of occurrence. In such cases, the probability of emissions was updated as epsilon, where epsilon = 1E-8.

Based on the above assumptions, all corresponding probabilities in the emission matrices were set according to the values in Table 2.2 and Table 2.3, below.

The entries of the emission probability matrix were created by enumerating the probability of emitting observations from each corresponding state. Table 2.2 below represents emission probabilities of screen state observations. In the case of screen state observations, there were no fixed intervals scheduled, and screen states transitions were recorded whenever a screen turned on or turned off. Each of the screen state change records ($\Uparrow, \Downarrow$) indicates two things: 1) That Ethica was definitely recording at that time slot and 2) the screen state was on or off following that event, respectively. Also, emissions could happen because of two reasons, one reason was the changing from an opposite screen state to the new state; the second reason was that the screen was already in that state, but the recording was just now turned on (due to a restart of Ethica's recording). Unfortunately, it was not possible to distinguish these two types of transitions in the database.

	SS_True	SS_False	SS_Absent
SR	epsilon	epsilon	(1-(2 × epsilon))
$\bar{S}$R	epsilon	epsilon	(1-(2 × epsilon))
S$\bar{R}$	epsilon	epsilon	(1-(2 × epsilon))
$\bar{S}\bar{R}$	epsilon	epsilon	(1-(2 × epsilon))
SOnR	1-(2 × epsilon)	epsilon	epsilon
SOffR	epsilon	(1-(2 × epsilon))	epsilon
ROnS	1-(2 × epsilon)	epsilon	epsilon
ROn$\bar{S}$	epsilon	1-(2 × epsilon)	epsilon

Table 2.2: Emission matrix - for screen state observations (epsilon = 1E-8)

Table 2.3 represents the emission probabilities of battery observations for each state. Every battery observation (B) was an indication that Ethica was running (and recording) in the back end. This battery-related observation was scheduled to record on each duty cycle of Ethica i.e., for every interval of approximately 5 minutes (300 seconds); while such recording takes place according to a fairly regular cycle, for simplicity, I assumed that the process was memoryless. Hence, in the emission matrix below, the probability of observing battery present records in each of the 6 recording states of Ethica was updated as once in every 300 seconds. Accordingly, the battery absent probability was updated as 1-(1/300 second) because the total probability was one, and there were only 2 possible observation possibilities for each state (here, an observation of a battery record, or its absence). In the two non-recording states – S$\bar{R}$ and $\bar{S}\bar{R}$ – the chance of observing battery records were zero, and the HMM assumes that the system will almost never emit a battery observation (as expressed by emitting a Bat_Absent observation in any time slot). Hence, the probability of emitting Bat_Present was updated as zero or epsilon (1E-8), and that of Bat_Absent was set to (1-epsilon).

	Bat_Present	Bat_Absent
SR	1/300	1-(1/300)
$\bar{S}$R	1/300	1-(1/300)
S$\bar{R}$	epsilon	1-epsilon
$\bar{S}\bar{R}$	epsilon	1-epsilon
SOnR	1/300	1-(1/300)
SOffR	1/300	1-(1/300)
ROnS	1/300	1-(1/300)
ROn$\bar{S}$	1/300	1-(1/300)

Table 2.3: Emission matrix - for battery observations (epsilon = 1E-8)

Estimation of model parameters using Optimization and Maximum Likelihood based objective function.

A maximum likelihood algorithm was used for fitting the HMM model, rather than the Baum Welch algorithm sometimes employed with HMMs [26]. Rather than estimating the high dimensional space of all HMM parameters, the size of the parameter space was reduced by incorporating knowledge about the underlying system in the form of transition and emission matrices, and by assumptions noted above. This approach required estimation of only 4 symbolic parameters of the transition matrix. Estimation of such parameters was carried out by maximizing the value of the log likelihood function while varying the values of such parameters, subject to constraints in the form of limits on the value of parameters being optimized. More details about the optimization are given below. After predicting the states, model assessment employed an observation-based confusion matrix variant created using empirical observations and predicted observations. A form of out-of-sample validation was undertaken via estimating an HMM using synthetic empirical data derived from a simulation model and testing the predictions of that HMM against a “synthetic ground truth” time series specifying the underlying state of that simulation model, where that ground truth was not itself used in HMM construction or estimation.

Optimization In this chapter, the maximum likelihood algorithm (implemented using the L-BFGS-B optimization method) was used to estimate four parameters of the transition probability matrix ($\mathcal{T}$), namely, P_SonSoff (the per-time slot probability of screen turning off from the on state), P_Soff_Son (the per-time slot probability of the screen turning on from the off state), P_Ron_Roff (the per-time slot probability of Ethica recording ceasing, given presence in a recording state) and P_Roff_Ron (the per-time slot probability of Ethica recording commencing, given presence in a non-recording state). The L-BFGS-B algorithm used here was a nondeterministic limited-memory algorithm for solving optimization problems subject to simple bounds on the variables [30]. It is based on the gradient projection method used to solve large nonlinear optimization problems with simple bounds, and uses a limited memory BFGS matrix to approximate the Hessian of the objective function [31]. Here, for each parameter i, an upper bound u_i and a lower bound l_i were imposed as constraints of the form $l_i \leq x_i \leq u_i$ was applied [31], and the boundaries of the four parameters were set as described in Table 2.4 below.

Parameter	lower bound (l_i)	upper bound (u_i)
P_SonSoff	epsilon	0.03334
P_Soff_Son	epsilon	0.00833333
P_Ron_Roff	epsilon	0.001667
P_Roff_Ron	epsilon	0.00333333

Table 2.4: Boundaries set for the optimization to estimate four transition probability matrix parameters

The ranges assumed to bound plausible values of transition probabilities were set based on assumptions about the dynamics of behaviours involved. The negative log likelihood was used as the objective function to locate the particular parameter values within this range via optimization, with details mentioned below.

Maximum Likelihood Algorithm Maximum likelihood is a widely used method for fitting HMMs [26]. To optimize the model parameters, we minimized the negative log likelihood [26], as given by the below equation, which is taken verbatim from Chapter III of [26]. If the observation sequence $x_1, x_2, \ldots, x_T$ was generated by the model, then the probability L_T of observing that sequence using an HMM with N states was treated as given by the likelihood

$$L_T = \delta P(x_1)\mathcal{T}P(x_2)\mathcal{T}P(x_3)\ldots\mathcal{T}P(x_T)1'$$

where δ denotes the row vector characterizing the initial probability distribution over states, $P(x)$ is the $N \times N$ diagonal matrix with its diagonal elements corresponding to the state-dependent probabilities for successive states or – for initial element $P(x_1)$ – density function δ, $'$ denotes the transpose operator, and thus $1'$ denotes a column vector of length m, each of whose elements is 1, and $\mathcal{T}$ is an $N \times N$ matrix, which denotes the transition probability matrix (tpm) of the Markov chain.

The computation of log likelihood ($logL_T$) implemented in this chapter as below was adapted as per the equation above and the algorithm mentioned in Chapter III of [26]. Within the implementation below, v and ϕ_t are row vectors of length N, u was a scalar, and $logL_T$ was the scalar running total in which the log-likelihood was accumulated over successive time points. The notations were also used as per [26] as below except slight changes in few cases such as $P(xbat_t)$ and $P(xscr_t)$ to fit the implemented HMM model.

Algorithm 1 Algorithm for the Negative Loglikelihood calculation

```
variables
    $T$ : Total count of observation time points
    $N$ : Count of states
    $\delta$ : Row vector of length $N$ specifying initial probability of being in each state
    $\mathcal{T}$ : State transition probability matrix ($A$) of size $N \times N$
    $v$ : Row vector of length $N$ holding a value proportional to the probability of being in each state,
            given observations until this point
    $u$ : Scalar holding the sum of $v$ over all states
    $\phi_t$ : Row vector of length $N$ of the probabilities of being in each state at time $t$,
          after normalizing $v$ by $u$
    $xbat_t$ : Emission probability matrix of battery observations of size $N \times 2$, where the 8 rows
            correspond to states, and the 2 columns correspond to each dichotomous distinct
            battery observations
    $xscr_t$ : Emission probability matrix of screen state observations of size $N \times 3$, where $N$ rows
            corresponds to $N$ distinct states and 3 columns corresponding to the three possible
            distinct screen state observations
    $P(xbat_t)$ : Battery emission probability row vector of size $N$, listing, specifically for time slot $t$,
            the per-state probabilities of observing the particular battery observation $xbat_t$ when
            in that state
    $P(xscr_t)$ : Screen state emission probability row vector of length $N$, listing, specifically for
            time slot $t$, the per-state probabilities of observing the particular screen state observation
            $xscr_t$ at time slot $t$, when in that state
    $P(x_t)$ : An $N \times N$ diagonal matrix, with diagonal values representing the outer product of two
            vectors: $P(xscr_t)$ and $P(xbat_t)$, for a particular time slot $t$
end variables
procedure PROCEDURE TO CALCULATE NEGATIVE LOGLIKELIHOOD
Initialization:
    $t \leftarrow 0$
    $LogL_T \leftarrow 0$
    $P(x_t) \leftarrow P(xbat_t) \times P(xscr_t)$
    $v \leftarrow \delta P(x_t)$
    $u \leftarrow v1'$
    $LogL_T = log(u)$
    $\phi_0 \leftarrow v/u$
```

```
loop:
    for $t = 1 \rightarrow (T-1)$ do
        $P(x_t) \leftarrow P(xbat_t)'P(xscr_t)$
        $v \leftarrow \phi_{t-1}\mathcal{T}P(x_t)$
        $u \leftarrow v \cdot 1'$
        $LogL_T = LogL_T + log(u)$
        $\phi_t \leftarrow v/u$
return: $-(LogL_T)$
```

In this chapter, the scipy.optimize.minimize function was used for optimization in python, with the objective function being set to the negative log likelihood; the resulting optimization can also be viewed as performing a maximization of the log likelihood. The required log-likelihood was represented in the equations using the final value of $LogL_T$ [26].

Forward-Backward algorithm

The Forward-Backward algorithm is used in this chapter to compute the posterior probability of being in a state given the observation sequence [26, 28]. Here, using the Forward-Backward algorithm, we create a probability matrix of length T × N, where T = the count of our observations, and $N = 8$, which represented the total number of states. This probability matrix specifies the probabilities of residence in the 8 different (hidden) HMM states at each time slot t up to the T^{th} time slot after computing both the forward probability ($\alpha_t(i)$) and backward probability($\beta_t(i)$). I implemented the algorithm from scratch in the programming language python, adapting the steps from [26]. The Forward-Backward Algorithm 2 implemented below was adapted from [26, 28, 32], and is explained below in this chapter.

Assume that the observation sequence $X = x_1, x_2, \ldots, x_T$ was generated by the model for a certain λ, and that $S = q_1, q_2, \ldots, q_T$ is the state sequence; further assume – following [26] – that $P(x_t)$ represents a diagonal matrix with the probability of observing x_t for each successive state on the diagonal, and that for a column vector x, x' denotes the transpose of x (a row vector). Then the Forward probability ($\alpha_t(i)$) and Backward probability $\beta_t(i)$, where t represents the time slot of the observation sequence, and i represents the hidden state, can be calculated inductively using 3 steps (refer Algorithm 2) below, themselves adapted from [26, 28].

To compute the state posterior probability of all state variables, this Forward and Backward variable, for each state associated with time t, computes $(\alpha_t \times \beta_t)/likelihood$. While implementing the algorithm for this chapter, we performed scaling and calculated the log of the values: $log\alpha_t$, $log\beta t$ and log-likelihood ($logL_T$). Hence the posterior marginals of all hidden states given observation sequence is calculated as: $exp((log\alpha_t + log\beta_t) - logL_T)$, where $1 \leq t \leq T$. The computation of the log-likelihood ($logL_T$) value was explained in Algorithm 1 above.

Algorithm 2 Steps to compute Forward and Backward variables

```
variables
    $x_t$ : An Observation at time $t$.
    $q_t$ : The state at time $t$.
    $\delta$ : Row vector of length $N$ specifying initial probability of being in each state
    $\mathcal{T}$ : State transition probability matrix ($A$) of size $N \times N$
    $P(x_t)$ : An $N \times N$ diagonal matrix for a particular time slot $t$, with diagonal values representing the probability of $x_t$ conditional to being in the $N$ successive states.
end variables
procedure Steps to compute Forward variable ($\alpha_t(i)$)
    $\alpha_t(i) \leftarrow P(x_1, x_2 ... x_t, q_t = S_i|\lambda)$
Initialization Step:
    $\alpha_1 \leftarrow \delta P(x_1)$, where $\alpha_1$ is a row vector of length $N$.
Induction Step:
    $\alpha_{t+1} \leftarrow \alpha_t \mathcal{T} P(x_{t+1})$, where $1 \leq t \leq T - 1$
procedure Steps to compute Backward variable ($\beta_t(i)$)
    $\beta_t(i) \leftarrow P(x_{t+1}, x_{t+2} ... x_T|q_t = S_i, \lambda)$
Initialization Step:
    $\beta_T \leftarrow 1$, where 1 is a row vector of length $N$, each of whose elements is 1, and $T$ is the observation sequence length.
Induction Step:
    $\beta'_t \leftarrow \mathcal{T} P(x_{t+1}) \beta'_{t+1}$ , where $t = T - 1, T - 2, .., 1$.
```

To aid in validation, a modified confusion matrix was created comparing the true observations and observations predicted based on the state posterior probabilities computed by the Forward-Backward algorithm at each time slot. This is created in a way different from the traditional confusion matrix. The methodology and equations of creating the matrix are explained in Section 2.7 and the results are added in the result Section 2.8.1 of this chapter. The resultant confusion matrices were computed, and are depicted in figures Figure 2.7 for the real data from Ethica across all users, and in Figure 2.9 for the synthetic observations data generated using a simulation model.

Viterbi algorithm to decode state sequence

Decoding or inferring state sequences was an important step of the HMM. The Viterbi algorithm – a widely used and common method for inferring state sequences – was used here for decoding the single most likely state sequence. This Viterbi Algorithm 3 implemented below was adapted from [28] and [26]. After training the model based on the empirical observations, model parameters (initial probabilities, transition and emission matrices as in section 2.5.1), and substituting the maximum-likelihood values of the parameters to form the trained HMM, the most likely hidden state sequences were inferred using the Viterbi algorithm. We implemented the Viterbi algorithm from scratch in python, adapting the steps from [26,28] in the algorithm 3 below.

Algorithm 3 Steps to compute Viterbi algorithm

```
variables
    $x_t$ : An observation at time t.
    $q_t$ : The state at time $t$.
    $P(x_t)$ : An $N \times N$ diagonal matrix for a particular time slot $t$, with diagonal values representing the probability of $x_t$ conditional to being in the $N$ successive states.
    $\xi_t(i)$ : A $T \times N$ matrix, that store the highest probability of the single path, at time $t$, for the first $t$ observations and ends in state $s_i$.
    $\psi_t(j)$ : A row vector of length $T$, to keep track of argument maximized for each $t$ and $j$.
end variables
procedure DEFINE $\xi_t(i)$
    $\xi_t(i) \leftarrow \max_{q_1,q_2,..,q_{t-1}} P[q_1, q_2...q_t = i, x_1, x_2, ..x_t|\lambda]$
Initialization Step:
    $\xi_1(i) \leftarrow \delta P(x_1)$
    $\psi_1 \leftarrow 0$, where 0 represents the zero row of vector of length $N$.
Recursion Step:
    $\xi_t(j) \leftarrow \max_{1\leq i\leq N}[\xi_{t-1}(i)\mathcal{T}_{ij}]P_j(x_t)$, where $2 \leq t \leq T$ and $1 \leq j \leq N$
    $\psi_t(j) \leftarrow \arg\max_{1\leq i\leq N}[\xi_{t-1}(i)\mathcal{T}_{ij}]$, where $2 \leq t \leq T$ and $1 \leq j \leq N$
Termination Step:
    $P^* \leftarrow \max_{1\leq i\leq N}[\xi_T(i)]$
    $q_T^* \leftarrow \arg\max_{1\leq i\leq N}[\xi_T(i)]$
Backtracking step:
    $q_T^* \leftarrow \psi_{t+1}(q_{t+1}^*)$, where $t = T-1, T-2, \ldots, 1$.
```

To find the single most likely hidden state sequence $Q = q_1, q_2, .., q_T$, for the given observation sequence (x) and model parameters λ, we define the quantity $\xi_t(i)$ as the highest probability along a single path at time t, for the first set of t observations, and ending in state $q_t = i$ [26,28]. To keep track of the maximized argument (most likely state sequence) for each time slot t and state j, an array $\psi_t(j)$ is used.

The Viterbi algorithm includes calculations also performed in the forward pass of the Forward-Backward algorithm, but includes an additional backtracking step and a major difference of the maximization over previous states in the induction step of Viterbi instead of the summing procedure in the Forward-Backward algorithm (Algorithm 2) explained above [28].

In addition to being assessed using the state-based posterior probabilities computed by the Forward-Backward algorithm, the accuracy of the HMM model to predict observations was also assessed using the single most likely state sequence as computed via the Viterbi algorithm. As above, this comparison used a modified confusion matrix comparing the predicted and observed observations. This is done separately for the participant-gathered data and for the synthetic dataset produced by the simulation. The methodology and the results of the modified confusion matrix for Viterbi results are added in the result section of this chapter in Section 2.8.1. The resultant confusion matrix is depicted in Figure 2.6 for the result from the real data from Ethica for all users and in Figure 2.8 for the synthetic observations data generated using the simulation model.

2.6 HMM validation using a simulation modelling approach

The unsupervised machine learning model was difficult to validate directly due to the unlabelled nature of the data sets, as a result of which understanding model precision was a challenge purely using empirical data. In this project, beyond the confusion matrices above, we used a simulation modelling approach to cross validate the HMM model accuracy by simulating the underlying Ethica data generation process, running the HMM on observations from the simulation model similar to those available empirically, and comparing the underlying ("true") state in the simulation model with what was inferred by the HMM algorithms.

Simulation models have been used in diverse studies as a successful approach for studying complex systems, by simulating scenarios after incorporating knowledge about the structure and quantitative particulars of the system into the model, and studying the patterns of the outputs over time arising from the model simulation. To both evaluate the accuracy of the HMM model and to test the effectiveness and correctness of the data processing pipeline, I built and used an agent based model (ABM) capable of plausibly mimicking an abstraction of the Ethica screen state data generation process, and of generating empirical observations in the same form as they were generated by Ethica. The simulation model was further capable of outputting the hidden state sequences underlying such observations. As mentioned in Section 2.3, whenever the Ethica process was actively running on the device, Ethica captured screen state transitions in an asynchronous fashion (that is, at irregular intervals); by contrast, battery observations were collected at approximately regular intervals of close to 5 minutes in duration. A continuous-time agent-based model was created for simulating similar situations, with a model time unit of 1 second. This ABM generated two synthetic time series of observations, namely screen state transition events and battery observations, each sharing the format exported from Ethica, and thus capable of being processed by the analysis pipeline used for this project, including HMM parameterization. The simulation model separately generates a synthetic ground truth time series reporting the hidden states associated with the data generation process represented by the simulation model; this time series was not used to inform the parameterization of the HMM, but was instead used to evaluate the accuracy of HMM-based inference as to underlying state sequence. Figure 2.4 depicts the agent based model, which was built using AnyLogic 8 Professional version 8.4.0 software.

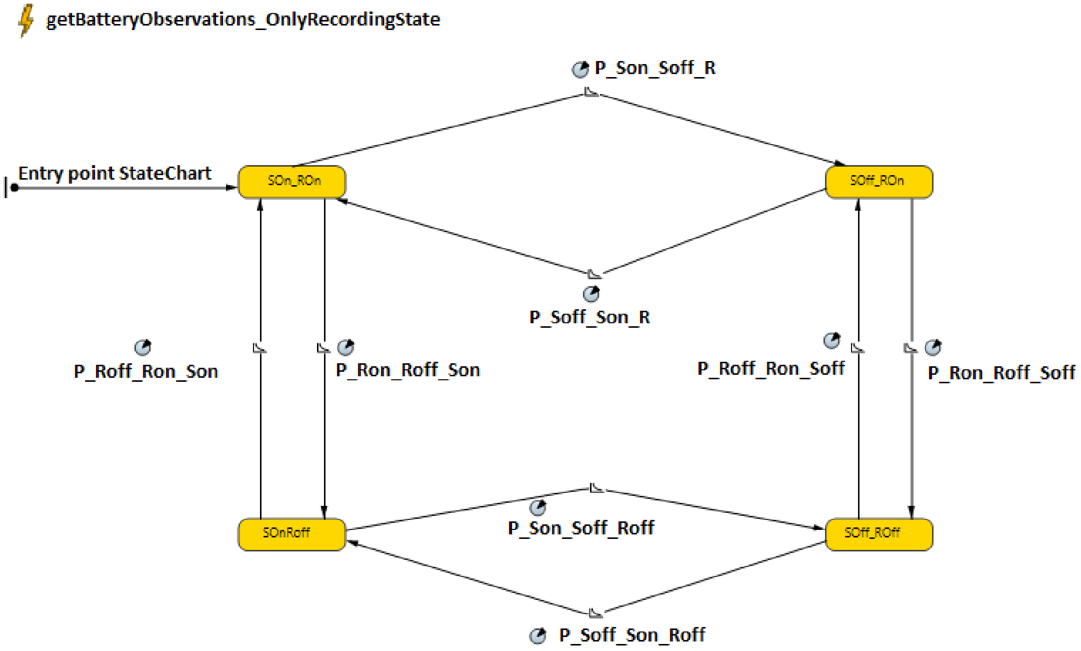


Figure 2.4: An agent based simulation model for synthetic ground truth data generation

For the purposes of testing, the screen-state and battery-related synthetic observations generated by the simulation model were used to parameterize the original HMM model implemented in the previous section 2.5.1 according to the unsupervised learning approach described above. The Forward-Backward algorithm and Viterbi algorithm were then run, with the resulting hidden states being predicted by the HMM for every (1-second duration) time slot for a duration of 45 days – the total participation duration of teenagers in the MEDIATICINO study. The time series of predicted hidden states for every 1 second time slot was then cross matched with the equal-length and contemporaneous time series of synthetic ground truth reporting of the underlying hidden states generated by the above ABM. On the basis of such mapping, the accuracy of the HMM model was assessed.

As per Figure 2.4, there were four main states defined within the agent based simulation model that roughly correspond to the four main hidden states of the data generation process, namely SOn_ROn, SOff_ROn, SOn_Roff and SOff_ROff, but with the proviso that the structure of the agent-based simulation model exhibits greater flexibility than does the HMM – for example, while the states in the HMM are restricted to be memoryless, those in the simulation model are not. The description of those the four simulation model states is as below:

- SOn_ROn : Represents a state of the phone when the screen was on and data was being recorded. Within the HMM (as depicted in Figure 2.2), this state predominantly approximated by SR in the HMM, but because of HMM limitations can also be approximated by SOnR and ROnS.

- SOff_ROn : Represents a state in which the phone screen is off and when the data recording process was running in the back end. Similar to the previous state, this is approximated by a combination of 3 other states within the HMM model structure in Figure 2.2, namely $\bar{S}$R, SOffR and ROn$\bar{S}$.

- SOn_Roff : Represents a state in which the phone screen was on, but the data recording process was not running. This is directly approximated by the state S$\bar{R}$ of the HMM model.

- SOff_ROff : Represents a state in which the screen is off and when the data recording process was not running – including situations in which the phone as a whole is off. The state is directly approximated by the state $\bar{S}\bar{R}$ of HMM model.

In this simulation model baseline experiment, the 8 transition rates between screen on and screen off states, and between the Ethica recording and non-recording states, accord closely with HMM assumptions, as reflected in the fact that the simulation model parameters correspond to the mid-points within the ranges employed when searching for the most favourable HMM model transition probabilities, as explained in the optimization section (Section 2.5.1). For example, in the baseline scenario, $P\_Son\_Soff\_R = P\_Son\_Soff\_R_Off$ = halfway between the minimum and maximum limits set for transition probability P_SonSoff of the HMM model = (0.0334/2.0). The 8 parameters values used for this baseline simulation model are described in Table 2.5 below.

Parameter Name	Value
$P\_Son\_Soff\_R$	0.03334/2.0
$P\_Soff\_Son\_R$	0.00833/2.0
$P\_Son\_Soff\_Roff$	0.03334/2.0
$P\_Soff\_Son\_Roff$	0.00833/2.0
$P\_Ron\_Roff\_Son$	0.001667/2.0
$P\_Roff\_Ron\_Son$	0.003333/2.0
$P\_Ron\_Roff\_Soff$	0.001667/2.0
$P\_Roff\_Ron\_Soff$	0.003333/2.0

Table 2.5: Table describing the parameter values of simulation model for the Baseline experiment

The AnyLogic simulation model was run for a time duration of 45 days, a period corresponding to the duration of the MEDIATICINO study. The first two states: SOn_ROn and SOff_ROn, represents the two main recording states in the simulation model. As in Ethica, both screen state transitions and battery measurements were recorded as observations when the simulation model was in either of these two states. By contrast, there were no state transitions or battery events captured from the nonrecording states in the AnyLogic simulation model; specifically, no observations were captured in states SOn_Roff and SOff_ROff.

The screen state observations "screen turning on" was captured whenever an entry to state SOn_ROn occurred within the simulation (synthetic ground truth) model. Correspondingly, a "screen turning off" event was captured whenever the simulation model entered state SOff_ROn. With respect to battery-related observations, an event that generates battery observations every five minutes (precisely) was established in the simulation model. But the battery observations from the event were captured (and later output) only when the active state in the simulation model was one of the two recording states: SOn_ROn and SOff_ROn. The time interval between such observations was chosen as five minutes because that corresponds the rough length of the Ethica duty cycle for recording battery observations (with the precise duty cycle varying over time). Two distinct time series of observations – one for each of screen state transitions and battery – were exported into Comma Separated Variable (CSV) files from AnyLogic database table(s) at the end of the simulation model execution. AnyLogic software has a feature to automatically write all essential statistics about the simulation model execution into the log files [33], which is used to log information about statechart transitions and events set up in the simulation model to database tables and to CSV files. The battery events from recording corresponding states are also collected in this way using the trace_log feature of AnyLogic software to copy all battery emissions printed as text outputs from the recording states into the CSV files [33].

Similar to the observations from Ethica, the observations generated by the simulation model also occurred at irregular intervals. Prior to delivery to the HMM, the synthetic ground truth data consequently also went through the same pre-processing step as did the empirical dataset (see Section 2.4); the use of an identical analysis pipeline for a ground-truth model further helped validate the correct operation of that pipeline. Within this analysis, there were five primary steps performed to preprocess the dataset:

- Step 1: The synthetic observations from the AnyLogic model – screen state turning on and turning off events – were collected and saved as table_1.
- Step 2: Sort the table_1 records in ascending order based on recorded time. As in the preprocessing step in Section 2.4, a new column indicating the 1 second timestamp labels was added. This column was the key column subsequently used to join with other data tables. If the screen turning on transition was recorded within a given second, it was marked as SS_True; those intervals including screen turning off observations were marked as SS_False. As in the case of screen state sensor data observations from smartphone data, if more than 1 transition occurred within a single 1 second time slot, then the final transition event was chosen (refer Step 3 of pre-processing, as described in Section 2.4).
- Step 3: For the battery observations collected from the battery emitting event set up on the simulation model, in a manner similar to step 2, a new column indicating 1 second timestamp labels was added. In accordance with step 5 of pre-processing described in Section 2.4, if there is more than 1 battery event observed within a 1-second time slot (≥ 1), then the battery observation for that particular time slot was marked as Bat_Present; time slots lacking any battery observations were marked as Bat_Absent.
- Step 4: In a fashion similar to step 2 of pre-processing in Section 2.4, a new dummy table_3 was

created, with records in 1 second regular intervals, beginning from the first time that the screen state transition was recorded in the simulation model until the last time that the screen state transition was recorded. For the creation of this table, only the maximum and minimum record_time of screen state observations were used and “explode” functionality in the Apache Spark data frame was used to create rows of 1 second intervals between the first and last record_time. Then a new column that marks the 1 second timestamp labels for every record was added. This served as the key column for table_3.

- Step 5: Dummy table_3 was then used as the main table. This table was subsequently left joined with the table_1 records using the “timestamp” column as the key value of the left join., so as to add a column in dummy table_3 with the screen state observations. The resulting table following the first left join was then, in turn, left joined with battery table_2, thus adding the column of battery observations to the resultant table.

- Step 6: The previous step resulted in a final joined table of observations, which have all of the records of screen state and battery observations subdivided according to regular 1 second intervals. The rows of the dummy table lacking any observations were marked as SS_Absent and Bat_Absent. The results are comparable to those of processing of the observation sequence after pre-processing the Ethica generated data.

The final observation sequence generated from step 6 (and originating in the AnyLogic simulation model) was used as the observation sequence in estimating and running the same HMM model described in Section 2.5.1.

As was noted above, similar to the observation sequence, a synthetic ground truth time series for the state sequence was also created from the AnyLogic model. This sequence was *not* used to inform HMM construction but was instead used later to assess the accuracy of HMM inference as to the underlying hidden state.

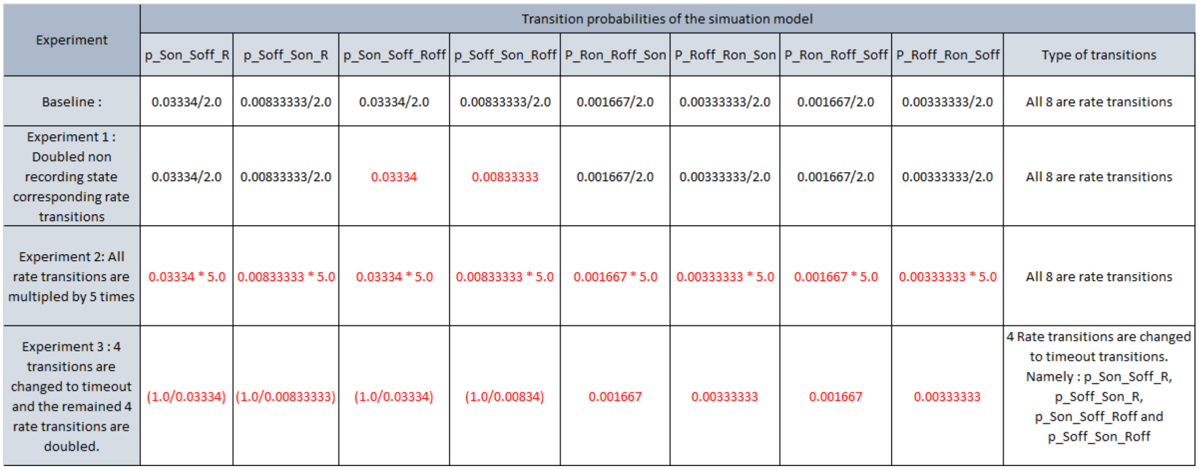
For generating these synthetic ground truth time series regarding the realized state sequence of the simulation model, a log table from AnyLogic database table(s) was used. This table logged information whenever an exit from the “active” state in the statechart of the simulation model. This table stores three types of information about the statechart: The “active” state of the state chart in the simulation model, and the entry and exit times associated with that state. This information was exported as a CSV file using the database log feature of AnyLogic software at the end of model execution. Then, using these data records, if the difference between the entry and exit time recorded for the active state was greater than 1 second, then an array of 1 second time slots between entry time and exit time associated with that active state was created and inserted as rows for each such 1 second time slots between the start time (entry time) and end time (exit time) using the explode function of spark dataframes. Then the active state recorded for the entry and exit time slots associated with the newly added time slots’ corresponding rows were filled using the active state of the simulation model marked for that interval in the log file. If more than 1 active state is recorded for the simulation model, then, as in the case of simulation model observations, the final active state at the end of the 1 second time slot was chosen. Thus, finally, a regular 1 second time slot dataset of the synthetic ground truth of the underlying states of the agent based simulation model was created and was used in the validation of the HMM models in later steps of this work.

The observation sequence generated using this simulation model was then passed to the HMM model, and the four steps mentioned in Section 2.5.1 were performed −− estimation of transition probabilities using the maximum likelihood algorithm, computation of the state posterior probability matrix at each time slot using the Forward-Backward algorithm, and the prediction of the single most probable state sequence using the Viterbi algorithm. The state inferences generated by the HMM obtained using this simulation model data are then compared using a confusion matrix against synthetic ground truth time series regarding the underlying state generated by the simulation model. The confusion matrix in Figure 2.10 in the result section 2.8.2 below was created using the synthetic ground truth states and the HMM model predicted state sequence. Two different versions of this confusion matrix were constructed to help evaluate the accuracy of HMM model predictions: One from the posterior probability predictions of the Forward-Backward algorithm, and another obtained from the Viterbi algorithm. Precision and recall values for each prediction were then calculated. The results are explained in detail in subsection 2.8.2 (Evaluation of HMM Model using Synthetic data) of the Results Section 2.8.

2.6.1 Three test experiments after changing the assumptions about the synthetic ground truth simulation model

Three different test experiments were conducted using the simulation modelling approach explained in Section 2.6 above, deliberately selecting scenarios in which the simulation model did not accord with the HMM assumptions. These tests were performed to understand the degree to which departures from HMM assumptions were tolerated, and the results of these experiments were investigated by creating confusion matrices using synthetic ground truth data, as in Section 2.8.2.

In the three experiments below, the agent-based simulation model structure remained the same as in the baseline scenario explained in above in Section 2.6 and Figure 2.4, having four main states for the data generation process in the simulation model: SOn_ROn, SOff_ROn, SOn_Roff and SOff_ROff. But in the experiments covered in this section, the transition rate in the simulation model is varied in 3 ways that depart from the HMM assumptions followed in the baseline simulation scenario, as described in Figure 2.5 and in the description below. The parameter values used for all experiments are explained in Figure 2.5 below.



Experiment	Transition probabilities of the simuation model								
	p_Son_Soff_R	p_Soff_Son_R	p_Son_Soff_Roff	p_Soff_Son_Roff	P_Ron_Roff_Son	P_Roff_Ron_Son	P_Ron_Roff_Soff	P_Roff_Ron_Soff	Type of transitions
Baseline :	0.03334/2.0	0.00833333/2.0	0.03334/2.0	0.00833333/2.0	0.001667/2.0	0.00333333/2.0	0.001667/2.0	0.00333333/2.0	All 8 are rate transitions
Experiment 1 : Doubled non recording state corresponding rate transitions	0.03334/2.0	0.00833333/2.0	0.03334	0.00833333	0.001667/2.0	0.00333333/2.0	0.001667/2.0	0.00333333/2.0	All 8 are rate transitions
Experiment 2: All rate transitions are multipled by 5 times	0.03334 * 5.0	0.00833333 * 5.0	0.03334 * 5.0	0.00833333 * 5.0	0.001667 * 5.0	0.00333333 * 5.0	0.001667 * 5.0	0.00333333 * 5.0	All 8 are rate transitions
Experiment 3 : 4 transitions are changed to timeout and the remained 4 rate transitions are doubled.	(1.0/0.03334)	(1.0/0.00833333)	(1.0/0.03334)	(1.0/0.00834)	0.001667	0.00333333	0.001667	0.00333333	4 Rate transitions are changed to timeout transitions. Namely : p_Son_Soff_R, p_Soff_Son_R, p_Son_Soff_Roff and p_Soff_Son_Roff

Figure 2.5: Table describing the parameter values of the simulation model for the baseline scenario (with transition rates of the simulation model according closely with HMM assumptions) and the 3 other experiments (placing the transition rates at a greater distance from HMM assumptions)

- Experiment 1: The rate of transitions between the screen on and screen off states in the synthetic ground truth model, when Ethica is not recording ($P\_Son\_Soff\_Roff$), is set to a value double that of rates corresponding to the Ethica recording states ($P\_Son\_Soff\_R$).

  In Experiment 1 : $P\_Son\_Soff\_Roff = 2 * P\_Son\_Soff\_R$

- Experiment 2: All rate transitions between states in the synthetic ground truth model are changed to 5 times as that of the maximum value set for optimization of the 4 corresponding HMM model transition probabilities (as characterized in Section 2.5.1).

- Experiment 3: Changed rate synthetic ground truth simulation model transitions between screen on and screen off states to timeout transitions (non-memmoryless transitions), and modified 4 parameters, namely $P\_Son\_Soff\_R$, $P\_Soff\_Son\_R$, $P\_Son\_Soff\_Roff$, $P\_Soff\_Son\_Roff$, as in Figure 2.5. The other 4 transition parameter values of the simulation model were changed to the maximum value set for optimization of the corresponding 2 transition probabilities (as explained in Section 2.5.1).

The results of all these 3 test experiments explained above along with the baseline experiment results are explained in detail in subsection 2.8.3 of the Results Section 2.8 (Results of test experiments after changing the assumptions about the synthetic ground truth simulation model).

2.7 Confusion matrix definitions

Two types of confusion matrices are created in the two sections below so as to evaluate the performance of the HMM model. The first type is the traditional confusion matrix that used to evaluate the accuracy of any classifier – usually in the context of supervised learning approaches – where each row of the matrix represents the true value of the variable (here, state) and each column represents instances of the predicted value of that variable (here, the predicted state). This can be represented as a matrix $C_1[i, j]$ where i represents the true state and j represents the predicted hidden state of the model. In Section 2.8.2, to evaluate the HMM model using synthetic data from the simulation model, we know the true states of the simulation model, and after running the HMM, the most likely predicted state sequence will be generated using the Viterbi algorithm. Hence in the case of such synthetic data, we have labelled data and accuracy of the model can be directly evaluated via traditional confusion matrices by comparing true underlying hidden states and predicted hidden states with the help of a multi-class case confusion matrix. Such a matrix is implemented and explained in Section 2.8.2.

However, in the case of an unsupervised approach, when only observations are present, and no labelled state data can be used to cross check the accuracy of state predictions, it is hard to apply a conventional confusion matrix. This is the case for the HMM when operating with the real data of 94 participants in this chapter. For such real data, the states are completely hidden, and we lack the ability to assess the accuracy of HMM predictions by recourse to privileged knowledge regarding the actual true state at any time slot. Hence, evaluation of model accuracy using the predicted Viterbi state sequence is not directly possible due to lack of knowledge of the true state sequence. But for this case, we do have sequences of actual empirical observations emitted – here, screen state and battery observations – at each step. Hence, to lend some understanding of accuracy, a construct similar to a confusion matrix is used to compare the prior predictive distribution or most likely state sequence and actual observations, with particulars as explained in the below paragraph. In order to derive these matrices (henceforth referred to as a “confusion matrix”), $C_3[i, j]$ (for Viterbi results) and $C_2[i, j]$ (for Forward-Backward results), the predicted states were decoded to generate the sequence of most probable observation sequence (or the posterior distribution over states) at each time slot and this is compared with the actual true observation sequence in a modified confusion matrix in the below section using real data (Section 2.8.1). Here, in this modified confusion matrix $C_2[i, j]$, each row of the matrix represents instances of true *observations*, and each column represents instances of *observations* decoded from the predicted states.

Algorithm to calculate modified confusion matrices $C_2[i, j]$ and $C_3[i, j]$

Both the rows and columns of the confusion matrices $C_2[i, j]$ and $C_3[i, j]$ were created using 6 different combinations of screen state and battery observations, namely

- BatteryPresent(Bt_Pr), ScreenStateTrue (SS_Tr)
- BatteryPresent(Bt_Pr), ScreenStateFalse (SS_Fl)
- BatteryPresent(Bt_Pr), ScreenStateAbsent (SS_Ab)
- BatteryAbsent(Bt_Ab), ScreenStateTrue (SS_Tr)

- BatteryAbsent(Bt_Ab), ScreenStateFalse (SS_Fl)

- BatteryAbsent(Bt_Ab), ScreenStateAbsent (SS_Ab)

where Bt_Pr indicates a time slot with battery observation present and Bt_Ab indicates a time slot with battery observation absent. ScreenStateTrue and ScreenStateFalse indicate screen state turning on and screen state turning off observations. ScreenStateAbsent was the time slot when there was no screen state observation.

Apart from the conventional confusion matrix $C_1[i,j]$ discussed in the next Section 2.8.2, for the modified confusion matrix discussed in the current Section 2.8.1, Two versions of this confusion matrix were created – one using the most likely state sequence obtained from the Viterbi algorithm ($C_3[i,j]$), and the other using the prior predictive distribution over state resulting from the Forward-Backward algorithm ($C_2[i,j]$). In case of the Viterbi variant, the confusion matrix is created using true observations and the decoded predicted observations of the most likely predicted state at each time slot. And in the case of Forward-Backward algorithm, the confusion matrix is created using true observations and prior predicted observations as characterized by the marginal of sequences of states calculated for each time slot.

The equation used to create the confusion matrix $C_2[i,j]$ (Forward-Backward algorithm) is explained as below.

$$C_2[i,j] = \sum_{t=2}^{T} I(i = x(t)) \sum_{s \in States} \frac{(\phi_{t-1} \cdot \mathcal{T})_s \cdot (\mathcal{T} \cdot \beta'_{t+1})_s}{L_T} \cdot l_s(j)) \qquad \text{(Eq. 1)}$$

where $x(t)$ is the observation (one of 6 noted above) at each time slot t, ϕ_t is the vector of probabilities of being in each state at time t (as taken from algorithm 1), $\mathcal{T}$ is the state transition matrix, $(\phi_{t-1} \cdot \mathcal{T})_s$ is the prior probability of being each state at time slot t, given observation of empirical data for time up to but not including observation i, $l_s(j)$ is the likelihood of observing j given that one is in state s, v_s represents element s of vector v; finally, as explained in Algorithm 2, β'_{t+1} (a column vector) is the transpose of β_{t+1} (a row vector) backward probability at time $t+1$, and L_T is the likelihood of the entire sequence. The indicator function $I(i = x(t))$ holds the value 1 if the value of $x(t)$ is i, and 0 otherwise.

The equation used to create confusion matrix $C_3[i,j]$ from the results of the Viterbi algorithm is added below.

$$C_3[i,j] = \sum_{t=2}^{T} I(i = x(t)) \; (\underset{s \in States}{I}[\max(\phi_{t-1} \cdot \mathcal{T})_s]) \; \cdot l_{s_{max}}(j) \qquad \text{(Eq. 2)}$$

Explanation: For a given cell in the final confusion matrix $C_2[i,j]$, for a given time slot t, where i is the true observation (row) for the actual observation at t and j is the predicted observation (column) at t, the value contributed to cell $C_2[i,j]$ represents the prior probabilities assessed by the HMM model of observing empirical datum j in light of all data up to but not including the latest observation; to the degree that the prior probability (HMM predicted without seeing the observed above) is high for the value that is in fact observed, the entries will be heavily weighted towards the diagonal entry (at which $i = j$). These values are then totalled up across all time slots.

As mentioned above, we have two variants of application of a prior prediction-based confusion matrix – one from the Forward-Backward algorithm $C_2[i, j]$ and another one from the Viterbi algorithm $C_3[i, j]$. The Viterbi algorithm gives the single most likely state sequence, whereas the Forward-Backward algorithm allows for assessing per-state probabilities. Here, in case of the Forward-Backward algorithm, we have a probability vector which is the probability of being in each of 8 different states for each time slot. Hence, in the case of the Forward-Backward algorithm, the model's probability is the sum over all states s, of $P(being\ in\ state\ s, observing\ j|x(1..t{-}1)) = P(being\ in\ state\ \text{“}s\text{”}|x(1..t{-}1)) \times P(observing\ j|being\ in\ state\ s,\ x(1..t{-}1)) \times l_s(j)$. The $l_s(j)$ is the likelihood of being in state s and observing possible observation j, where the probability of being in state s is the probability we could be in state s at time t by being in state r at time $t - 1$ and transitioning from r to s. This probability is summed over all states r at time $t - 1$, which is $(\phi_{t-1}\mathcal{T})$, where ϕ_{t-1} is a vector of probabilities of being in each possible state r at time $t - 1$ (given all of the observations up to and not including that at time t) and $\mathcal{T}$ is the Transition matrix.

$(\phi_{t-1}\mathcal{T})$ is the prior probability of being in each possible state at time t, given all of the observations up to and not including that at time. And $(\phi_{t-1}\mathcal{T})_s$ is the prior probability of being in state s at time t, given all such observations.

2.8 Results and discussion

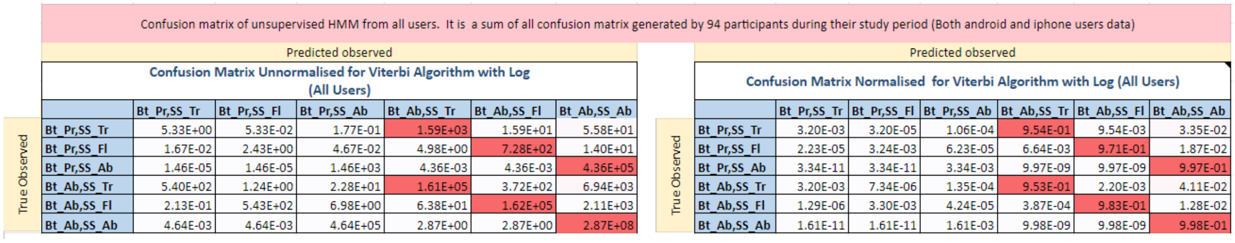
This result section includes three major subsections. The First and second sections – Section 2.8.1 and Section 2.8.2 – describe the results of running the HMM using the real data and synthetic data, respectively, and are used to evaluate the HMM model. But the third section – Section 2.8.3 – shows the result of the test experiments after altering the assumptions about the simulation model such that they differ from those for the baseline experiment explained in Section 2.8.2, and are further removed from the assumptions underlying the HMM.

According to the confusion matrix derivations described above, the patterns of results obtained from the two algorithms are studied for the empirical data from the Ethica study(Section 2.8.1), and also for the synthetic data from the simulation model(Section 2.8.2).

2.8.1 Evaluation of HMM model using Ethica study data

Assessing Viterbi algorithm results using confusion matrix (modified version: $C_3[i, j]$)

In the case of the Viterbi algorithm, the resultant confusion matrix $C_3[i, j]$ was created as per the algorithm stated in Equation Eq. 2 and explained in Section 2.7. This resultant confusion matrix is shown in Figure 2.6. For both this matrix and later confusion matrices, red colour in the figure is used to highlight cells with greater probabilities accumulated when compared to other predicted observed corresponding probabilities in the same row.



Confusion matrix of unsupervised HMM from all users. It is a sum of all confusion matrix generated by 94 participants during their study period (Both android and iphone users data)

Predicted observed

Confusion Matrix Unnormalised for Viterbi Algorithm with Log (All Users)

True Observed	Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
Bt_Pr,SS_Tr	5.33E+00	5.33E-02	1.77E-01	1.59E+03	1.59E+01	5.58E+01
Bt_Pr,SS_Fl	1.67E-02	2.43E+00	4.67E-02	4.98E+00	7.28E+02	1.40E+01
Bt_Pr,SS_Ab	1.46E-05	1.46E-05	1.46E+03	4.36E-03	4.36E-03	4.36E+05
Bt_Ab,SS_Tr	5.40E+02	1.24E+00	2.28E+01	1.61E+05	3.72E+02	6.94E+03
Bt_Ab,SS_Fl	2.13E-01	5.43E+02	6.98E+00	6.38E+01	1.62E+05	2.11E+03
Bt_Ab,SS_Ab	4.64E-03	4.64E-03	4.64E+05	2.87E+00	2.87E+00	2.87E+08

Predicted observed

Confusion Matrix Normalised for Viterbi Algorithm with Log (All Users)

True Observed	Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
Bt_Pr,SS_Tr	3.20E-03	3.20E-05	1.06E-04	9.54E-01	9.54E-03	3.35E-02
Bt_Pr,SS_Fl	2.23E-05	3.24E-03	6.23E-05	6.64E-03	9.71E-01	1.87E-02
Bt_Pr,SS_Ab	3.34E-11	3.34E-11	3.34E-03	9.97E-09	9.97E-09	9.97E-01
Bt_Ab,SS_Tr	3.20E-03	7.34E-06	1.35E-04	9.53E-01	2.20E-03	4.11E-02
Bt_Ab,SS_Fl	1.29E-06	3.30E-03	4.24E-05	3.87E-04	9.83E-01	1.28E-02
Bt_Ab,SS_Ab	1.61E-11	1.61E-11	1.61E-03	9.98E-09	9.98E-09	9.98E-01

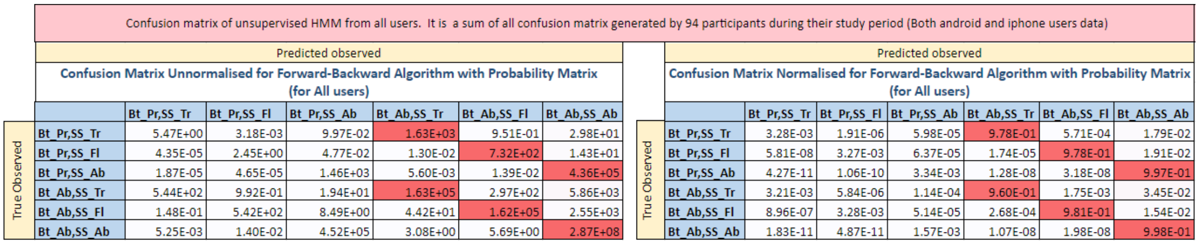
Figure 2.6: Confusion matrix unnormalized and normalized for the Viterbi algorithm (based on data from all users). In this confusion matrix, columns corresponds to predicted observations, and the rows correspond to true observations.

As mentioned in Section 2.7, combinations of screen state and battery observations that define six composite observation possibilities that are used to define the confusion matrix. The battery records appear fairly regularly, but in a fashion that cannot be readily predicted given the memoryless assumptions associated with HMM states. There is a large volume of data without battery observations in the observation time series. This reflects the fact that the time slot in the time series was chosen as seconds, and given that battery observations were only made every approximately five minutes even when Ethica is recording, in the large majority of the time slots there were no battery observations. After the confusion matrix is created as per the equations for $C_3[i, j]$ (Equation Eq. 2), it was then normalized to understand the patterns better. For the normalization step, we choose each row (i) of the unnormalized matrix – representing a particular true observation – and divide each element in that row by the sum of all elements in that row. This normalization step was performed using the normalize function of scikit-learn library [34]. Both the normalized and unnormalized versions of the confusion matrix are shown in Figure 2.6.

While examining the normalized confusion matrix depicted in Figure 2.6, it was noticed from the last 3 rows that the predictions of 3 true observed cases (battery absent), namely ((BatteryAbsent(Bt_Ab), ScreenStateTrue (SS_Tr)), (BatteryAbsent(Bt_Ab), ScreenStateFalse (SS_Fl)) and (BatteryAbsent(Bt_Ab), ScreenStateAbsent (SS_Ab)) were showing an accumulation of larger probabilities on the diagonal values when compared to non-diagonal values. By contrast, in the first 3 rows of the confusion matrix, entries corresponding to the 3 true observed (battery present) cases, namely (BatteryPresent(Bt_Pr), ScreenStateTrue (SS_Tr), (BatteryPresent(Bt_Pr), ScreenStateFalse (SS_Fl)) and (BatteryPresent(Bt_Pr), ScreenStateAbsent (SS_Ab)), there is a deviation in the accumulation of probabilities of the predicted observed on the non-diagonal when compared to the corresponding diagonal cells.

Assessment of Forward-Backward algorithm inferences using the confusion matrix (modified version: $C_2[i,j]$)

Similarly, for the Forward-Backward algorithm, on each time slot, the probability matrix with the probabilities of observing each state was calculated; after decoding the predicted state related probabilities as explained in Section 2.7, the confusion matrix $C_2[i,j]$ was created as given in Equation Eq. 1. This resultant confusion matrix for the Forward-Backward algorithm is given below in Figure 2.7.



Confusion matrix of unsupervised HMM from all users. It is a sum of all confusion matrix generated by 94 participants during their study period (Both android and iphone users data)

Predicted observed

Confusion Matrix Unnormalised for Forward-Backward Algorithm with Probability Matrix (for All users)

		Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
True Observed	Bt_Pr,SS_Tr	5.47E+00	3.18E-03	9.97E-02	1.63E+03	9.51E-01	2.98E+01
	Bt_Pr,SS_Fl	4.35E-05	2.45E+00	4.77E-02	1.30E-02	7.32E+02	1.43E+01
	Bt_Pr,SS_Ab	1.87E-05	4.65E-05	1.46E+03	5.60E-03	1.39E-02	4.36E+05
	Bt_Ab,SS_Tr	5.44E+02	9.92E-01	1.94E+01	1.63E+05	2.97E+02	5.86E+03
	Bt_Ab,SS_Fl	1.48E-01	5.42E+02	8.49E+00	4.42E+01	1.62E+05	2.55E+03
	Bt_Ab,SS_Ab	5.25E-03	1.40E-02	4.52E+05	3.08E+00	5.69E+00	2.87E+08

Predicted observed

Confusion Matrix Normalised for Forward-Backward Algorithm with Probability Matrix (for All users)

		Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
True Observed	Bt_Pr,SS_Tr	3.28E-03	1.91E-06	5.98E-05	9.78E-01	5.71E-04	1.79E-02
	Bt_Pr,SS_Fl	5.81E-08	3.27E-03	6.37E-05	1.74E-05	9.78E-01	1.91E-02
	Bt_Pr,SS_Ab	4.27E-11	1.06E-10	3.34E-03	1.28E-08	3.18E-08	9.97E-01
	Bt_Ab,SS_Tr	3.21E-03	5.84E-06	1.14E-04	9.60E-01	1.75E-03	3.45E-02
	Bt_Ab,SS_Fl	8.96E-07	3.28E-03	5.14E-05	2.68E-04	9.81E-01	1.54E-02
	Bt_Ab,SS_Ab	1.83E-11	4.87E-11	1.57E-03	1.07E-08	1.98E-08	9.98E-01

Figure 2.7: Confusion matrix unnormalized and normalized for Forward-Backward algorithm (drawing data from all users). (In this confusion matrix, columns correspond to predicted observations and rows to true observations.)

In this confusion matrix in Figure 2.7, a pattern similar to the results from the Viterbi-related confusion matrix shown in Figure 2.6 was noticed. As for that earlier matrix, red colour is used to highlight cells with larger probabilities. Some deviation is noticed in case of 3 observed predicted cases.

Overall, the results of the HMM from the Viterbi and Forward-Backward algorithms show similar patterns as per the modified confusion matrices $C_2[i,j]$ and $C_3[i,j]$ created using the results. From these confusion matrices, it was noticed that 3 predicted observations exhibit larger probabilities accumulated on the off-diagonal cells of the confusion matrix when compared to the corresponding diagonal values. But this modified version of the confusion matrices does not include any direct information about states, which makes more challenging commenting or arriving at any conclusions about the accuracy of the HMM from these matrices. Hence, to evaluate the accuracy, this HMM was applied to synthetic ground truth data generated using a simulation modelling approach (as explained in Section 2.6), and the resultant accuracy was calculated with the help of conventional confusion matrix ($C_1[i,j]$) created using the predicted states and true states. Along with that, the modified version of confusion matrices $C_2[i,j]$ and $C_3[i,j]$ using predicted observations and true observations were also created, and the pattern of the results was analyzed. This helped to evaluate the accuracy, precision, recall and F-test on the HMM results (refer Figure 2.11 in Section 2.8.2), thereby supporting evaluation of the performance of the HMM. The results are explained in the next Section 2.8.2.

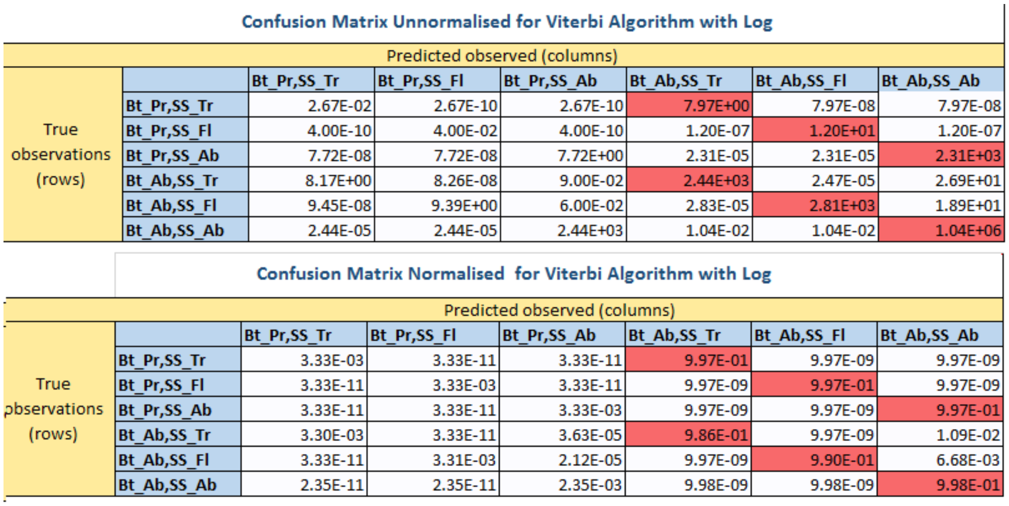
2.8.2 Evaluation of HMM model using synthetic data

This section characterizes the results from the model built and evaluated using the synthetic dataset characterized in Section 2.6. Similar confusion matrices in Figure 2.6 and Figure 2.7 from both the Viterbi and the

Forward-Backward algorithm was generated. These were generated to check whether the patterns of Viterbi and Forward-Backward algorithm resulting from synthetic data differ from what is seen in the other pair of confusion matrices for the dataset from the Ethica study across all 94 participants.

Assessment of Viterbi algorithm results using observation-based confusion matrix (modified version: $C_3[i, j]$)

As mentioned above, the synthetic data generated from the baseline experiment of the simulation model was used in this Section for HMM evaluation. Firstly, the Viterbi algorithm was used to predict the most likely state sequence, and – as for the case of the Ethica dataset – a resultant modified confusion matrix($C_3[i, j]$) using the Viterbi results was created, as is shown in Figure 2.8.



Confusion Matrix Unnormalised for Viterbi Algorithm with Log

		Predicted observed (columns)					
		Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
True observations (rows)	Bt_Pr,SS_Tr	2.67E-02	2.67E-10	2.67E-10	7.97E+00	7.97E-08	7.97E-08
	Bt_Pr,SS_Fl	4.00E-10	4.00E-02	4.00E-10	1.20E-07	1.20E+01	1.20E-07
	Bt_Pr,SS_Ab	7.72E-08	7.72E-08	7.72E+00	2.31E-05	2.31E-05	2.31E+03
	Bt_Ab,SS_Tr	8.17E+00	8.26E-08	9.00E-02	2.44E+03	2.47E-05	2.69E+01
	Bt_Ab,SS_Fl	9.45E-08	9.39E+00	6.00E-02	2.83E-05	2.81E+03	1.89E+01
	Bt_Ab,SS_Ab	2.44E-05	2.44E-05	2.44E+03	1.04E-02	1.04E-02	1.04E+06

Confusion Matrix Normalised for Viterbi Algorithm with Log

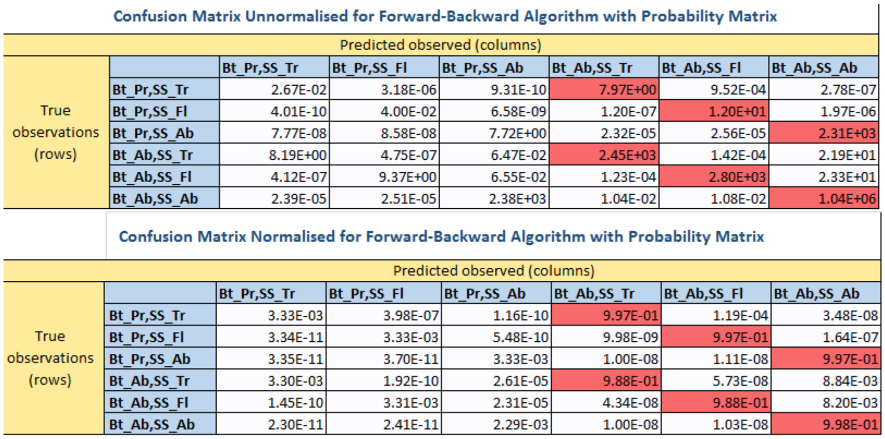
		Predicted observed (columns)					
		Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
True observations (rows)	Bt_Pr,SS_Tr	3.33E-03	3.33E-11	3.33E-11	9.97E-01	9.97E-09	9.97E-09
	Bt_Pr,SS_Fl	3.33E-11	3.33E-03	3.33E-11	9.97E-09	9.97E-01	9.97E-09
	Bt_Pr,SS_Ab	3.33E-11	3.33E-11	3.33E-03	9.97E-09	9.97E-09	9.97E-01
	Bt_Ab,SS_Tr	3.30E-03	3.33E-11	3.63E-05	9.86E-01	9.97E-09	1.09E-02
	Bt_Ab,SS_Fl	3.33E-11	3.31E-03	2.12E-05	9.97E-09	9.90E-01	6.68E-03
	Bt_Ab,SS_Ab	2.35E-11	2.35E-11	2.35E-03	9.98E-09	9.98E-09	9.98E-01

Figure 2.8: Unnormalized (top) and normalized (bottom) confusion matrix using predictions from the Viterbi algorithm, based on synthetic screen state and battery time series. In this confusion matrix, columns correspond to predicted observations and rows correspond to true observations.

This confusion matrix 2.8, drawing on the results of the Viterbi algorithm, was created in the same way as above matrix with rows (i) representing true observations, and columns (j) representing predicted observations using 6 different combinations of observations in a manner highly similar to that applied in the previous section. An important exception lies in the fact that the confusion matrices in the case of empirical datasets were created by summing up the values of all confusion matrices created for all 94 participants, while here the matrix was created using synthetic data generated from a single ABM run of 45 days in length. The patterns shown in the confusion matrix created using results from synthetic data were very similar to those associated with the Viterbi result from the original dataset. As in the case of other confusion matrices above, red colour is used here to highlight the cells accumulating larger probabilities when compared to other predicted observed corresponding probability values accumulated in other cells of the same row. The top confusion matrix in Figure 2.8 was unnormalized, and the bottom one was normalized.

Assessment of Forward-Backward algorithm results using observation-based confusion matrix(modified version: $C_2[i,j]$)

A Forward-Backward algorithm approach was also implemented for the results using a synthetic dataset to rigorously check the accuracy of predictions of observations on the basis of the Forward-Backward algorithm-assessed prior probabilities of underlying state in each time slot. And here also, a confusion matrix $C_2[i,j]$ was created using 6 different combinations of observations in a manner highly similar to that applied in the previous section to cross check the pattern observed for the Forward-Backward result. The $[T*N]$ probability matrix obtained from the Forward-Backward algorithm was used as the predicted state probabilities, where T = total length of observations and N = total number of hidden states (8).



Confusion Matrix Unnormalised for Forward-Backward Algorithm with Probability Matrix

		Predicted observed (columns)					
		Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
True observations (rows)	Bt_Pr,SS_Tr	2.67E-02	3.18E-06	9.31E-10	7.97E+00	9.52E-04	2.78E-07
	Bt_Pr,SS_Fl	4.01E-10	4.00E-02	6.58E-09	1.20E-07	1.20E+01	1.97E-06
	Bt_Pr,SS_Ab	7.77E-08	8.58E-08	7.72E+00	2.32E-05	2.56E-05	2.31E+03
	Bt_Ab,SS_Tr	8.19E+00	4.75E-07	6.47E-02	2.45E+03	1.42E-04	2.19E+01
	Bt_Ab,SS_Fl	4.12E-07	9.37E+00	6.55E-02	1.23E-04	2.80E+03	2.33E+01
	Bt_Ab,SS_Ab	2.39E-05	2.51E-05	2.38E+03	1.04E-02	1.08E-02	1.04E+06

Confusion Matrix Normalised for Forward-Backward Algorithm with Probability Matrix

		Predicted observed (columns)					
		Bt_Pr,SS_Tr	Bt_Pr,SS_Fl	Bt_Pr,SS_Ab	Bt_Ab,SS_Tr	Bt_Ab,SS_Fl	Bt_Ab,SS_Ab
True observations (rows)	Bt_Pr,SS_Tr	3.33E-03	3.98E-07	1.16E-10	9.97E-01	1.19E-04	3.48E-08
	Bt_Pr,SS_Fl	3.34E-11	3.33E-03	5.48E-10	9.98E-09	9.97E-01	1.64E-07
	Bt_Pr,SS_Ab	3.35E-11	3.70E-11	3.33E-03	1.00E-08	1.11E-08	9.97E-01
	Bt_Ab,SS_Tr	3.30E-03	1.92E-10	2.61E-05	9.88E-01	5.73E-08	8.84E-03
	Bt_Ab,SS_Fl	1.45E-10	3.31E-03	2.31E-05	4.34E-08	9.88E-01	8.20E-03
	Bt_Ab,SS_Ab	2.30E-11	2.41E-11	2.29E-03	1.00E-08	1.03E-08	9.98E-01

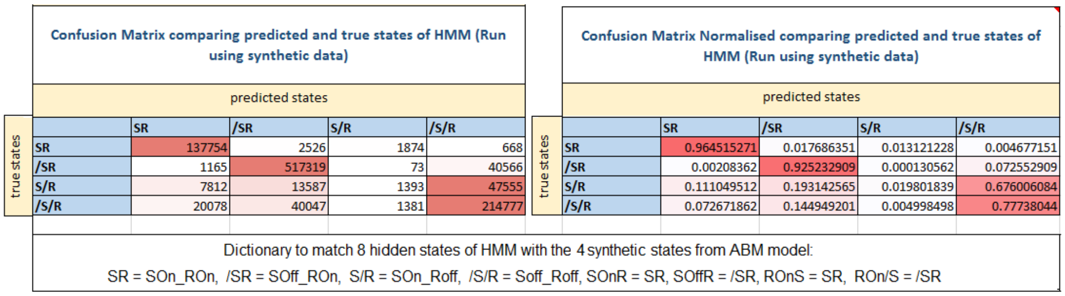
Figure 2.9: Confusion matrix unnormalized (top) and normalized (bottom) variants for the Forward-Backward algorithm, based on synthetic screen state and battery data based-observations. In this confusion matrix, columns correspond to predicted observations, and rows correspond to true observations.

In the case of the Forward-Backward algorithm, the same pattern shows as in the case of the corresponding confusion matrix for the empirical dataset (in Figure 2.7). As in the case of other modified confusion matrices explained above, the values of the diagonal elements of the confusion matrix in the first 3 rows – corresponding to the cases of predicted observations including Battery Present – are smaller than some non-diagonal values of the same row.

To conclude, it is noticed that the modified confusion matrices $C_2[i,j]$ of the Viterbi results and $C_3[i,j]$ of Forward-Backward algorithm results using the synthetic datasets exhibit a similar pattern to that observed in the corresponding confusion matrices generated by the empirical datasets. But to enhance the depth of conclusions regarding the performance of the classifier, greater insight can be secured by creating and analyzing conventional confusion matrices ($C_1[i,j]$) – a task that constitutes the focus of the next section.

Ground truth-based cross validation of hidden states predicted by the HMM: validation method 2 (conventional confusion matrix $(C_1[i,j])$)

For the synthetic data case, synthetic ground truth data was available, providing the true states at each time slot. Hence, after running the HMM, the state sequence predicted by the Viterbi algorithm at each time slot was cross matched with the synthetic true state at that time slot, and a confusion matrix was created. Here, because the true underlying states from the synthetic ground truth simulation model are known, it was possible to create a conventional confusion matrix by comparing the true and predicted state sequences. That matrix is depicted in Figure 2.10; in contrast to the confusion matrices using prior predictive values – which have rows and columns corresponding to observations – this confusion matrix has rows and columns corresponding to states.



Confusion Matrix comparing predicted and true states of HMM (Run using synthetic data)

true states	predicted states SR	/SR	S/R	/S/R
SR	137754	2526	1874	668
/SR	1165	517319	73	40566
S/R	7812	13587	1393	47555
/S/R	20078	40047	1381	214777

Confusion Matrix Normalised comparing predicted and true states of HMM (Run using synthetic data)

true states	predicted states SR	/SR	S/R	/S/R
SR	0.964515271	0.017686351	0.013121228	0.004677151
/SR	0.00208362	0.925232909	0.000130562	0.072552909
S/R	0.111049512	0.193142565	0.019801839	0.676006084
/S/R	0.072671862	0.144949201	0.004998498	0.77738044

Dictionary to match 8 hidden states of HMM with the 4 synthetic states from ABM model:

SR = SOn_ROn, /SR = SOff_ROn, S/R = SOn_Roff, /S/R = Soff_Roff, SOnR = SR, SOffR = /SR, ROnS = SR, ROn/S = /SR

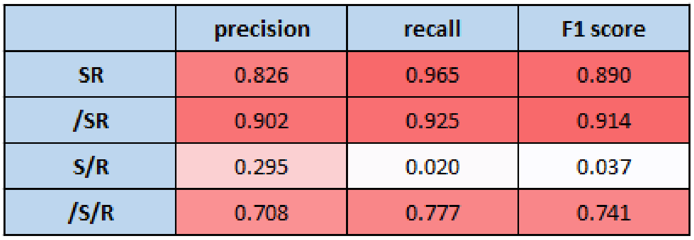
Figure 2.10: Confusion matrix created comparing synthetic ground truth data for hidden state and contemporaneous model predicted hidden state). In this confusion matrix, columns correspond predicted states and rows correspond to true states.

The overall accuracy calculated from the confusion matrix is 83%. It is calculated by dividing the sum of diagonal values of the unnormalized confusion matrix in Figure 2.10 by the total sum of all values of that matrix (refer Section 2.8.2).

Precision and recall for the HMM results using synthetic ground truth data

HMM model performance was also evaluated using standard precision and recall methods and F_1 score, as applied to a particular state. The precision was calculated by dividing the true positive by the sum of true positive and false positive values [35], when considered for a given state. For example, here to calculate the precision of SR, from the unnormalized confusion matrix on the left side of Figure 2.10, the true positive value of SR $= 137754$ was divided by the sum of true positive and false negative values corresponding to the SR. i.e., the precision of SR was considered to be $P_{SR} = TP_{SR}/(TP_{SR} + FP_{SR}) = 137754/(137754 + (1165 + 7812 + 20078)) = 0.82$. Similarly, recall values corresponding to hidden states SR, $\bar{S}$R, S$\bar{R}$ and $\bar{S}\bar{R}$ were also calculated as R1, R2, R3 and R4. The recall values were calculated by dividing true positive by the sum of true positive and false negative values. For example, in case of same SR, $R_{SR} = TP_{SR}/(TP_{SR} + FN_{SR}) =$

$137754/(137754 + (2526 + 1874 + 668)) = 0.96$. The recall values were the same values shown in the diagonal values of the normalized confusion matrix. Precision and recall values are depicted in Figure 2.11.



	precision	recall	F1 score
SR	0.826	0.965	0.890
/SR	0.902	0.925	0.914
S/R	0.295	0.020	0.037
/S/R	0.708	0.777	0.741

Figure 2.11: The precision, recall and F_1 score values for the 4 main predicted states.

It was noticed that precision values of $\bar{S}$R, SR and $\bar{S}\bar{R}$ were showing higher values. However, in the case of S$\bar{R}$, the precision and recall values were lower, indicating the difficulty that the algorithm had in predicting those hidden states. This reflects the challenges of predicting the underlying screen state (and thus the appropriate hidden state) when observations are not available for prolonged periods.

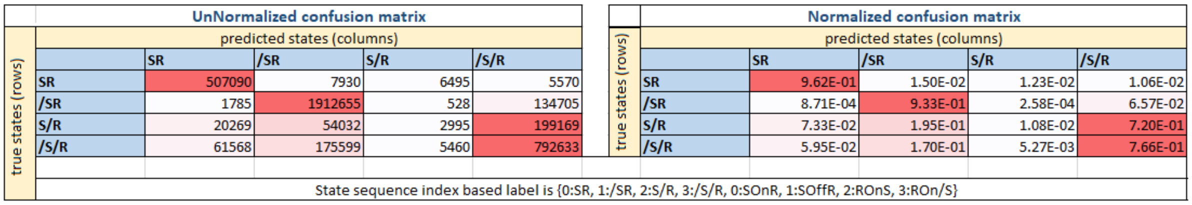
The F_1 score values were also calculated for the 4 major states to better understand the HMM accuracy, particularly in light of the non-dichotomous classification involved. The F_1 score is the harmonic mean calculated using the recall and precision values for the 4 main predicted states [36]. i.e., the F_1 score of SR was calculated as $F_{1SR} = 2 * ((P_{SR} * R_{SR})/(P_{SR} + R_{SR})) = 2 * ((0.826 * 0.965)/(0.826 + 0.965)) = 0.890$.

2.8.3 Result of test experiments after changing the assumptions about the synthetic ground truth simulation model

Below I explain the evaluation of the results generated using data collected from three successive test experiments of the simulation model, as explained in Section 2.6.1.

Experiment 1: Doubled Ethica non-recording state corresponding transition rates between on and off states

The rates associated with 2 rate transitions $P\_Son\_Soff\_Roff$ and $P\_Soff\_Son\_Roff$ between 2 Ethica non-recording corresponding – states namely SOnRoff and SOff_ROff – in Figure 2.4 are doubled. The rate associated with these rate transitions in the baseline scenario was set to be identical to the corresponding one of 2 other rate transitions between the corresponding Ethica recording states: $P\_Son\_Soff\_R$ and $P\_Soff\_Son\_R$, where $P\_Son\_Soff\_R = (0.03334/2.0) = 0.01667$ and $P\_Soff\_Son\_R = (0.00833333/2.0) = 0.00416$. But in Experiment 1, the $P\_Son\_Soff\_R$ and $P\_Soff\_Son\_R$ retain their original value of 0.017, and the rates associated with the transitions between non-recording states are set as 0.034. Having established such values, as for the baseline scenario explained in Section 2.8.2, the synthetic ground truth data is obtained from the output of the simulation model of experiment 1, and the HMM model is then estimated. The resultant state sequence predicted by the Viterbi algorithm at each time slot was then cross matched with the synthetic true states at each time slot, using a conventional (state-based) confusion matrix depicted in Figure 2.12.



UnNormalized confusion matrix

predicted states (columns)

true states (rows)	SR	/SR	S/R	/S/R
SR	507090	7930	6495	5570
/SR	1785	1912655	528	134705
S/R	20269	54032	2995	199169
/S/R	61568	175599	5460	792633

Normalized confusion matrix

predicted states (columns)

true states (rows)	SR	/SR	S/R	/S/R
SR	9.62E-01	1.50E-02	1.23E-02	1.06E-02
/SR	8.71E-04	9.33E-01	2.58E-04	6.57E-02
S/R	7.33E-02	1.95E-01	1.08E-02	7.20E-01
/S/R	5.95E-02	1.70E-01	5.27E-03	7.66E-01

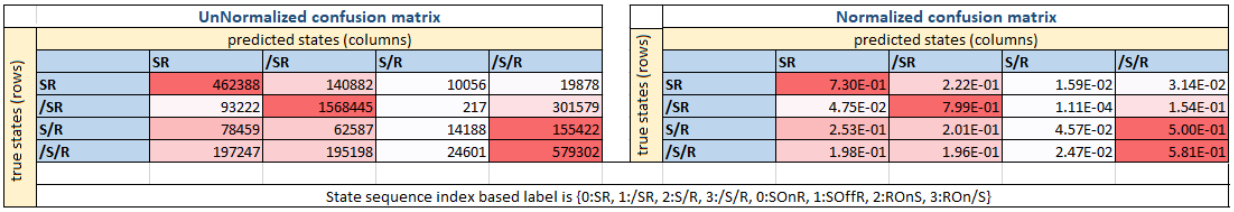
State sequence index based label is {0:SR, 1:/SR, 2:S/R, 3:/S/R, 0:SOnR, 1:SOffR, 2:ROnS, 3:ROn/S}

Figure 2.12: Confusion matrix created comparing synthetic ground truth data for the hidden state from Experiment 1 and contemporaneous model predicted hidden state. In this confusion matrix, columns correspond to predicted state and rows correspond to true state.

Here the results show a similar pattern to that of the baseline scenario, and the accuracy calculated from the confusion matrix is 82.68%, which is extremely close to that calculated for the baseline model (83%). As described in Section 2.8.2, the diagonal values indicate the number of time slots where the predicted label is same as that of the true label, here the sum of diagonal values of the unnormalized confusion matrix in Figure 2.12 above = (507090 + 1912655 + 2995 + 792633) = 3215373, and the total sum of all values of unnormalized confusion matrix = 3888483. The overall accuracy of HMM prediction is therefore = 3215373/3888483 = 0.8268 = 82.68% – a favourable value, despite the violation of HMM assumptions.

Experiment 2: Far higher simulation model transitions.

In Experiment 2, all 8 transition rates are set differently than that of the baseline scenario, as described in table in Figure 2.5 above. Here, all rate transitions between simulation model states are changed to a value 5 times that of the maximum value of the optimization search range. Specifically, the maximum value set for optimization for all 4 uncertain parameters of transition matrix is multiplied by 5 and then set here for the values of the corresponding simulation model transition parameters. The results obtained are depicted below, in the confusion matrix in Figure 2.13.



UnNormalized confusion matrix

predicted states (columns)

true states (rows)	SR	/SR	S/R	/S/R
SR	462388	140882	10056	19878
/SR	93222	1568445	217	301579
S/R	78459	62587	14188	155422
/S/R	197247	195198	24601	579302

Normalized confusion matrix

predicted states (columns)

true states (rows)	SR	/SR	S/R	/S/R
SR	7.30E-01	2.22E-01	1.59E-02	3.14E-02
/SR	4.75E-02	7.99E-01	1.11E-04	1.54E-01
S/R	2.53E-01	2.01E-01	4.57E-02	5.00E-01
/S/R	1.98E-01	1.96E-01	2.47E-02	5.81E-01

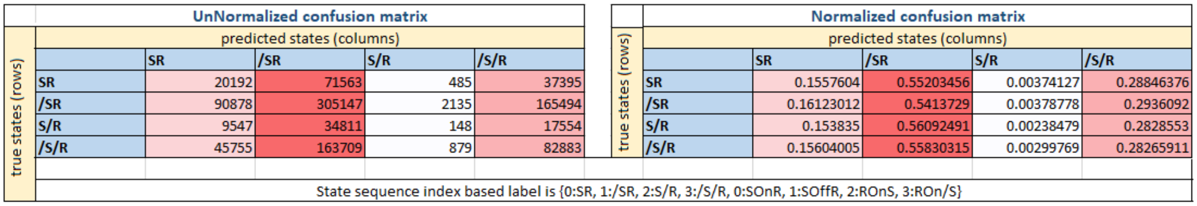
State sequence index based label is {0:SR, 1:/SR, 2:S/R, 3:/S/R, 0:SOnR, 1:SOffR, 2:ROnS, 3:ROn/S}

Figure 2.13: Confusion matrix created comparing synthetic ground truth data for the hidden state from Experiment 2 and contemporaneous model predicted hidden state. In this confusion matrix, columns correspond to predicted state and rows correspond to true state.

Here the results are notably worse than for the baseline scenario. The accuracy of the model calculated by dividing the sum of the diagonal values with the total sum of confusion matrix is 67.22% – notably lower value compared to the roughly 83% corresponding to the results of both the baseline and experiment 1.

Experiment 3: Broader parameter and residence time modifications

In this 3^{rd} experiment, I changed the rate transition between screen on and screen off states for simulation model experiment to timeout transitions, and modified all 8 of the associated parameter values. For that, firstly, 8 parameter values of the simulation model are modified in a way by setting them as the maximum limit set for the optimization for the corresponding 4 uncertain transition probabilities of the HMM model. Then the 4 transitions of the simulation experiment occurring between the screen turn on and turn off states during Ethica recording and non-recording states are modified so as to occur after a precise fixed time (constituting non-memoryless transitions), rather than according to a hazard rate. For converting the rate transition to a timeout transition, the timeout duration is set to the reciprocal of the rate values given for the associated parameter values, as described in Figure 2.5 for the parameters, namely $P\_Son\_Soff\_R$, $P\_Soff\_Son\_R$, $P\_Son\_Soff\_Roff$ and $P\_Soff\_Son\_Roff$. The resultant confusion matrix is depicted in Figure 2.14 below.



UnNormalized confusion matrix

		predicted states (columns)			
		SR	/SR	S/R	/S/R
true states (rows)	SR	20192	71563	485	37395
	/SR	90878	305147	2135	165494
	S/R	9547	34811	148	17554
	/S/R	45755	163709	879	82883

Normalized confusion matrix

		predicted states (columns)			
		SR	/SR	S/R	/S/R
true states (rows)	SR	0.1557604	0.55203456	0.00374127	0.28846376
	/SR	0.16123012	0.5413729	0.00378778	0.2936092
	S/R	0.153835	0.56092491	0.00238479	0.2828553
	/S/R	0.15604005	0.55830315	0.00299769	0.28265911

State sequence index based label is {0:SR, 1:/SR, 2:S/R, 3:/S/R, 0:SOnR, 1:SOffR, 2:ROnS, 3:ROn/S}

Figure 2.14: Confusion matrix created comparing synthetic ground truth data for hidden states from Experiment 3 and contemporaneous model predicted hidden state. In this confusion matrix, columns correspond to predicted states and rows correspond to true states.

The accuracy calculated from the confusion matrix in Figure 2.14 above by dividing the sum of diagonal values by the total sum of all values of the confusion matrix is only 38.94%, which is far lower than the accuracy of the above two experiments – a result highlighting the fragility of HMM results to pronounced multiple simultaneous deviations from HMM assumptions.

2.9 Conclusion

At an overall level, the unsupervised HMM implemented in this chapter appears to have successfully delivered on its goal of labeling the underlying hidden states with an acceptable level of accuracy. The results from cross checking with synthetic ground truth experiments suggest that during the Ethica recording states, the model was able to infer the latent evolution of screen state over time with satisfactory precision. Not surprisingly, in the case of Ethica non-recording states, the ability of the HMM to infer the screen state is substantially impaired, with the precision of model accuracy for screen off states lying in acceptable ranges, but with inferences for the screen on states exhibiting poor accuracy. This compromised accuracy reflects the fact that, in states in which Ethica is not recording data, there are no indications of screen state emissions recorded, and, over time, the model becomes increasingly uncertain as to whether the screen is on or off. Because the parameter estimates reflect the fact that the phone as a whole spends greater time with the screen off than on, the MLE-trained HMM favours predictions of screen off states by virtue of transition probabilities implying greater residence time in non-recording states in which the screen is off. This results in lower accuracy in predictions of screen on state in Ethica non-recording states. But overall, from the confusion matrix in Figure 2.10 and the precision-recall result in Figure 2.11, the use of synthetic ground truth data suggests that the model was successful in predicting hidden states and state sequences with an acceptable accuracy level.

It also bears emphasis that while the empirical data applied here was limited to a single study, the approach introduced above – making use of both HMMs and cross-validation using a simulation model – can be reapplied in similar studies involving screen_state sensor data. This capacity is particularly valuable in light of both the popularity of collecting screen state data in Ethica studies and the fact that inference of underlying screen state is a typical need of such studies, as virtually all such studies will share uncertainties associated with the data collection process and missing data points that require resolution for reliable understanding of screen time exposure on a participant-by-participant basis. Also, the re-usability of the approach in future studies will be unlikely to require changes in code or substantial extra effort due to the standard character of the Ethica data model for the relevant screen state and battery variables. This is likely to make the project helpful for other health researchers using the same platform. We hope that the contribution of the mechanisms presented here will help save money and time by allowing researchers to readily estimate study participants' daily screen time patterns. Amongst other uses, this will be helpful for examining the association between screen time, behavioural patterns and mental health, etc., and in understanding other aspects of health behaviours and exposures. The same patterns observed in the cross validation results using the original dataset and synthetic dataset further suggest considerable promise in readily cross validating machine learning algorithms using simulation models, including for measurands far removed from screen state. One limitation of this simulation model is that it does not consider the time-of-day effects shaping screen-time patterns. Such effects can be pronounced – for example, sleeping periods are expected to be characterized by long intervals between the screen turning on – such effects further seem likely to vary significantly across different people, both because of different times of the day spent sleeping, and – particularly considering youth – different patterns with respect to phone use while in bed. Hence as, a promising avenue of future work, the effectiveness of the current HMM should be evaluated using synthetic data produced from a simulation model exhibiting pronounced changes in screentime patterns across different times of day.

Chapter 3

Mood change patterns and influence of depression, irritability and connectedness on Suicidal ideation in psychiatric inpatients: A longitudinal smartphone based study

3.1 Introduction

Good mental health also plays a vital role alongside physical health for a person to live a healthy life. But sometimes – due to personal reasons, stressful lifestyle, financial crises, etc. – some people find it difficult to maintain this balance, leading to mental health conditions such as chronic anxiety, depression, etc. In some, mental health challenges lead to suicidal thoughts, and thus risk loss of life. But providing proper support at the right times can help people to regain balance and continue on to a successful life ahead.

Suicide is a leading cause of avoidable death in Canada [37]. According to Statistics Canada, from 2012 to 2016, intentional self-harm in the form of suicide ranked as the 9^{th} leading cause of Canadian deaths. This report also states that early detection of signs of suicidal behaviour and proper measures to help people in need can help reduce the burden of health losses due to suicides [38]. The literature further suggests that suicide is a complication of a psychiatric disorder, that most persons who attempt suicide have a psychiatric disorder, with more than 90% of suicide victims having a diagnosable psychiatric illness [39–44]. These studies also suggest that mood disorders are the most common psychiatric conditions associated with suicide or serious suicide attempts [39–45].

There are well-established literatures describing past studies conducted to analyze the influence of risk factors on suicidality. One such study indicates that identifying the risk factors can help to improve treatment and assessment of suicidality [46]. According to the Vantaa Depression Study (VDS), the risk of suicide attempts among patients with MDD (Major Depressive Disorder) during a major depressive episode is clearly higher when compared with the risk during a period of full remission [47]. That study further notes that reducing the time spent depressed is a highly credible measure for preventing future suicide attempts [47]. Another study conducted using a sample of 625 males in New York indicated that irritability and impulsivity are strongly associated with suicidal ideation [48]. Based on another study, adolescent self-rated irritability

($p < 0.001$) and depression ($p < 0.001$) were positively associated with adolescent suicidal ideation and are more strongly associated with suicidal ideation than other measured risk factors. By contrast, parent-rated adolescent irritability was not strongly associated with adolescent suicidal ideation [46]. Along with the association of depression and irritability level with a person's suicidal ideation, some other studies also point towards a lack of connectivity across the lifespan as a risk factor in suicidal behaviour [49]. Lack of connectedness was generally referred to in these studies as being constituted by a lack of social support, poor integration into a social network, or perceptions of social isolation [49, 50].

Taking all the above factors into consideration, in this study – led by psychiatrist Dr. Rudy Bowen of the University of Saskatchewan Department of Psychiatry but involving a larger team of researchers – four survey questions were formulated for smartphone-based data collection to provide a longitudinal record of the level and (especially) variability in suicidality and its influential risk factors in psychiatric inpatients. Specifically, self-reporting was conducted on four self-reported risk factors: Depression, and feelings of irritability, connectedness and suicidality. These are chosen to assess patients' mood fluctuations, mental health patterns, and also to better understand the influence of the other 3 risk factors on suicidal ideation.

Also, from the above mentioned studies, it was clear that self-reported measures from an index individual are distinct from – and likely offer greater accuracy than – those offered by people associated with that index person, such as parents or friends. Those findings point to the importance of collecting responses from the index person rather than from parents or other closely related contacts. Hence, in this study, the survey responses were collected from the patients themselves on a thrice-daily basis, and were prompted at fixed intervals of time. Such data collection supported reporting on a person's thoughts and mood changes over time. In addition to that, there were lines of evidence that frequent mood fluctuations play a role in depression and suicidality. Hence, in the major research paper for this study – lying outside the scope of this chapter – using the same dataset but also clinical data not examined here, an assessment on the impact of mood instability to predict suicidal thoughts among patients was performed by the research team.

The main objective of this chapter was to provide methods to support these major study results and the work of the broader research team. The work in this chapter therefore analyzed longitudinal data collected from psychiatric inpatients to support understanding of changes in patterns of their depression, irritability and connectedness responses over the study periods as main risk factors contributing towards mental illness and to support to understand their impact on suicidal behaviour.

3.1.1 Study setting

This was a longitudinal study employing de-identified data collected via smartphones from psychiatric inpatients while they were undergoing treatment in the Irene & Leslie Dubé Centre for Mental Health in Saskatoon, Saskatchewan, Canada. The author of this thesis was added to the project as a study team member for data analysis. The study was approved by the Behavioural Ethics Review Board (University of

Saskatchewan Behavioural Ethics Board protocol 15-201), with the study starting on 2015 December and ending by December 2017. The study duration for each patient was around 2 weeks, with variations depending on their admission and discharge date. 51 patients participated in the study, with patient consent, enrollment, phone deployment, and record keeping being managed by research assistants in the Department of Psychiatry. However, the start date and end dates of the study period for some participants were not recorded properly, yielding only incomplete resulting data. As part of the data cleaning for the work, 9 participants were therefore excluded from study, leaving data from 42 participants for further analysis in this chapter.

3.1.2 Data collection

Study participants included both females and males between 18 and 70 years of age who were admitted into the facility: Irene & Leslie Dubé Centre for Mental Health, with a mood disorder and suicidal behaviour/thoughts. There were two sources of data collection: First, the participants complete validated retrospective measures of depression, mood instability and suicidal thoughts under the supervision of research assistants on the first and last day of their studied period. Secondly, data was collected using a smartphone-based application called Ethica 1.2.2, installed on Android smartphones provided to the participating patients in the facility by the research assistants. Only data collected through smartphones were analyzed in this chapter.

Two types of data were collected using smartphones: Sensor data and survey data. Four distinct classes of on-phone sensors were used for this study: Location-based sensors (GPS and Wi-Fi), Environmental sensors (Ambient Temperature, Light, Proximity), those detecting participant-participant contact patterns (via Bluetooth), motion sensors (Accelerometer, Gravity, Gyroscope, Linear Acceleration, Magnetic Field, Orientation), and the battery sensor. Survey data was also collected from participants using a single survey questionnaire via thrice-daily surveys set to trigger at 9 am, 3 pm and 8 pm. On each presentation, the survey asked the same four questions, with the participant marking their response to each of the four using a horizontal visual analogue scale on a scale of 0 to 100 within the questionnaire. The four questions requested that the participant indicate their level of feeling related to each of depression, irritability, connectedness and suicidality. Question wording was as follows:

- How depressed do you feel right now?
- How irritable/angry do you feel right now?
- How close or connected to people do you feel right now?
- How suicidal are you right now?

References to survey responses throughout this chapter should be taken as referring to responses to the single survey questionnaire consisting of the above 4 questions. Participant answers to these questions were investigated in this study on the basis of their patterns and association between variables.

3.2 Methodology

3.2.1 Data pre-processing

Several stages of filtering logic were performed in this chapter to filter out certain types of data and to improve the trustworthiness of data to make it fit for further exploratory analysis. These steps were represented diagrammatically as a flowchart in Figure 3.1 below.

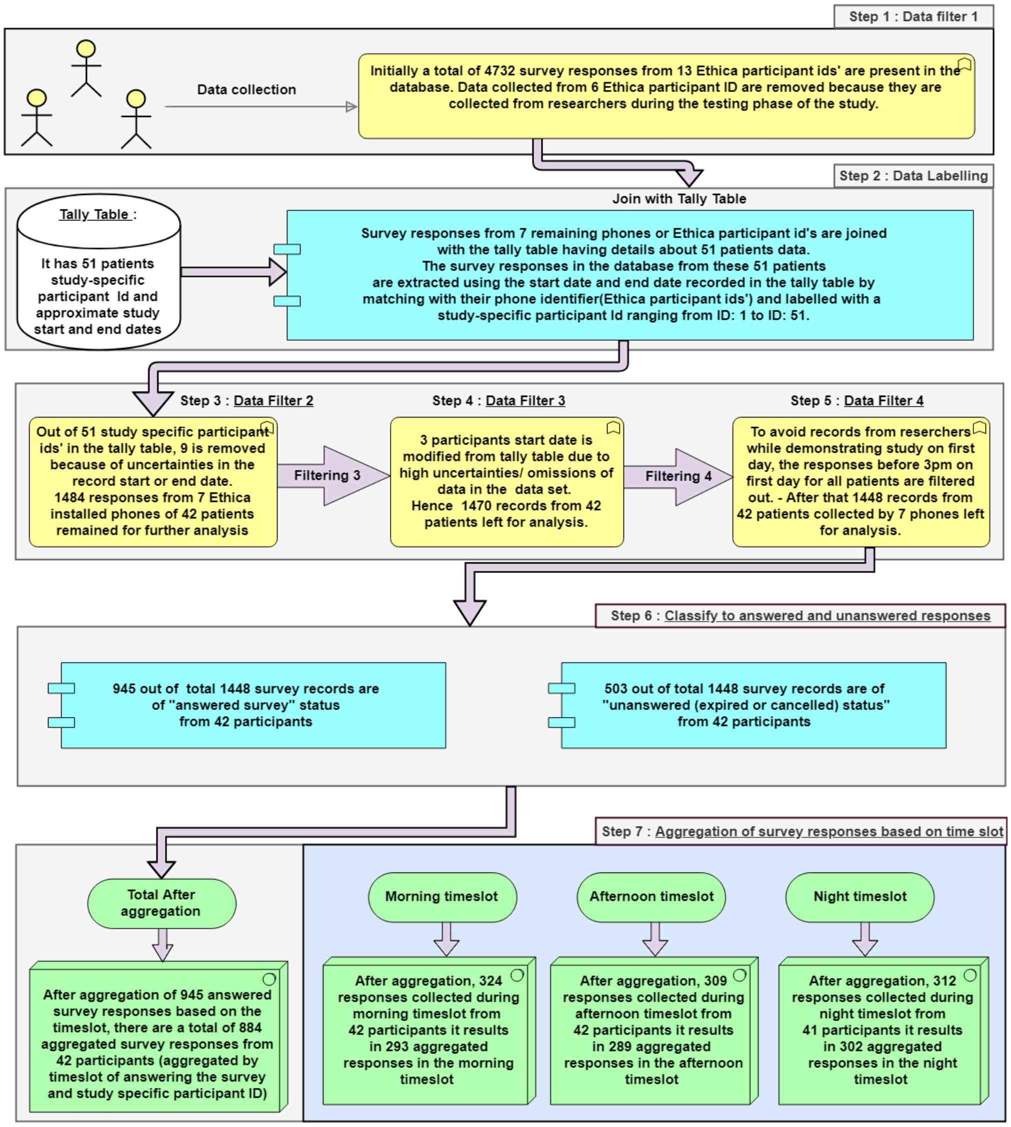

Step 1 : Data filter 1
Data collection
Initially a total of 4732 survey responses from 13 Ethica participant ids' are present in the database. Data collected from 6 Ethica participant ID are removed because they are collected from researchers during the testing phase of the study.
Step 2 : Data Labelling
Join with Tally Table
Tally Table :
It has 51 patients study-specific participant Id and approximate study start and end dates
Survey responses from 7 remaining phones or Ethica participant id's are joined with the tally table having details about 51 patients data. The survey responses in the database from these 51 patients are extracted using the start date and end date recorded in the tally table by matching with their phone identifier(Ethica participant ids') and labelled with a study-specific participant Id ranging from ID: 1 to ID: 51.
Step 3 : Data Filter 2
Out of 51 study specific participant ids' in the tally table, 9 is removed because of uncertainities in the record start or end date. 1484 responses from 7 Ethica installed phones of 42 patients remained for further analysis
Filtering 3
Step 4 : Data Filter 3
3 participants start date is modified from tally table due to high uncertainties/ omissions of data in the data set. Hence 1470 records from 42 patients left for analysis.
Filtering 4
Step 5 : Data Filter 4
To avoid records from reserchers while demonstrating study on first day, the responses before 3pm on first day for all patients are filtered out. - After that 1448 records from 42 patients collected by 7 phones left for analysis.
Step 6 : Classify to answered and unanswered responses
945 out of total 1448 survey records are of "answered survey" status from 42 participants
503 out of total 1448 survey records are of "unanswered (expired or cancelled) status" from 42 participants
Step 7 : Aggregation of survey responses based on time slot
Total After aggregation
After aggregation of 945 answered survey responses based on the timeslot, there are a total of 884 aggregated survey responses from 42 participants (aggregated by timeslot of answering the survey and study specific participant ID)
Morning timeslot
After aggregation, 324 responses collected during morning timeslot from 42 participants it results in 293 aggregated responses in the morning timeslot
Afternoon timeslot
After aggregation, 309 responses collected during afternoon timeslot from 42 participants it results in 289 aggregated responses in the afternoon timeslot
Night timeslot
After aggregation, 312 responses collected during night timeslot from 41 participants it results in 302 aggregated responses in the night timeslot

Figure 3.1: Flowchart of data pre-processing steps which includes filtering and aggregation of records

Step 1 to step 5 in the above Figure 3.1 are explained in the below filtering Section 3.2.2; the following steps (6 and 7) are explained in the operationalization Section 3.2.3.

3.2.2 Filtering

Filtering plays an important role in many data analysis processes, particularly in the context of data quality. The stages of filtering applied for this study are mentioned below.

Filtering step 1

Each patient participating in this study was provided with a single phone that they used throughout their study period; however, to conserve the study budget, the same phone was distributed to multiple individuals for different disjoint intervals of time. So as to support linkage with clinical data, the first analysis step was to identify the phone (as distinguished by a unique phone identifier) associated with a participant and to map that to a newly added unique study-specific participant ID. While identifying the mapping from a participant to the unique phone that they carried would be straightforward given typical data management practices for clinical studies, a persistent oversight by a research assistant involved meant that the retained paperwork was sometimes insufficient to directly determine which phone was paired with a subset of participants. As a result, this function mapping study-specific participant ID to phone identifier (Ethica participant ID) needed to be deduced through analysis.

There are in total 4732 survey responses from 13 Ethica participant ids' found in the study database, namely: 231, 238, 283, 471, 472, 474, 492, 493, 594, 609, 629, 630 and 809. But out of these 13, data from 6 – namely, 231, 238, 283, 594, 609 and 809 – were filtered out, because these phones were marked as the phones used by the research team for testing purposes only. The actual records from 7 phones – associated with Ethica participant ids 471, 472, 474, 492, 493, 629 and 630 – were marked as including data from patients participated in the study.

Filtering step 2

Initially, there were in total 51 study-specific participants details – such as start date, end date, and phone identifiers – present in the tally table. As the first step, these details present in the tally table regarding 51 patients were cross-matched with the survey responses collected from corresponding phone identifiers (Ethica participant ids). But in the case of 9 patients, uncertainty about their participation period was noticed in both the tally table and in the survey responses collected by Ethica. Data from these 9 participants were therefore excluded from the analysis, and the start date and end date for the remaining 42 participants were fixed as per the clinical records. These details were used to extract data from the 42 participants and used for further analysis. After performing this filter, there were a total of 1484 responses remaining in the database, provided 42 study-specific participant ids' collected using 7 mobile phones.

Filtering step 3

It was noticed that for 3 study-specific participant ids', after their first day of data collection, there were several days of non-activity (no data recorded by Ethica from their phones). Examples include the fact that one participant has no records from day 2 until day 6, another has no records from day 4 to day 14, and yet another has no records from day 2 to day 8. This non-activity can be due to several factors, such as their phone was turned off or Ethica deliberately terminated during those periods. It could also result from erroneously assigning the start date as being the patient facility admission date in the tally table due to researcher oversight in recording participant study start dates. Other explanations are also possible – such as those associated with treatment schedules, highly stressed mental conditions on initial days, etc. But in order to be scientifically conservative, I sought to eliminate the issues in the later stage of data analysis and to improve the quality of data for future analysis through filtering. Specifically, for those 3 participants, start dates for the analysis treated as different from the tally table by removing the missing data in the beginning; instead, the start date for the sake of the analysis was set to be the date at which the actual responses started being collected through the corresponding phones. This change in the start date for 3 participants removed 14 survey responses from the database, and left 1470 survey responses for further analysis.

The survey responses in the database included two types of responses, termed here "answered" and "unanswered".The unanswered responses included expired and cancelled survey responses. Out of 1470 survey responses collected from 7 distinct phones, there were 505 unanswered and 965 answered responses.

Filtering step 4

Based on information from the clinical team, it was determined that study enrollments were consistently happening in the morning time. From the exploratory examination of data patterns observed on the morning of the first day of participation and from the discussions with the clinical team, the possibility became strongly evident that the research assistant was employing the same phone given to patients for the demonstration to the participant of data collection instruments and their functionality. Hence, to avoid the risk of analysis inadvertently being contaminated by data drawn from responses made by the research assistant during the study demonstration session, following discussion with research team, logic was implemented across the dataset to remove data recorded prior to 3 pm on each participant's study entry date. This step filtered out 22 records from 1470 total responses – including both answered and unanswered surveys from the above step – after which 1448 records remained for further analysis.

After 4 filtering steps, there were in total 1448 survey responses remaining for analysis; out of this total, 503 were considered unanswered (expired/cancelled) and 945 were answered survey responses.

3.2.3 Operationalization

As the first step of operationalization, for each participant, the start date and end date of the study period were recorded. Then, for each survey response, the within-study day count for that response was calculated using the survey answered date and the start date for that participant, with the result being added as a new column to the participant's dataset called "dayinstudy"(start with a designation of the first day of participation as day 1). In the next step, the day of week and (integer) hour slot when the participant had answered the survey was added to the dataset. Then, based on the hour of answering surveys, a new column called "time slot" was added. This column has 3 possible values – designating the morning, afternoon and night. If the (integer) hour associated with the survey was answered fell in the (integer) range of 4 am and 1 pm, then it was marked as belonging to the morning slot; if that hour fell in the integer range between 1 pm and 6 pm, then it was marked as falling in the afternoon slot, and if the integer hour associated with the response fell in the integer range between 6 pm and 12 am, then it was marked as being in the night slot. It bears note that even though the survey timings are set as 9 am, 3 pm and 8 pm, some participants' responses are gathered shortly after that, due to the non-zero survey expiry time or technical glitches. Hence, this newly added time slot column helps in examining responses by differentiating into morning, afternoon and night responses. Out of 945 answered surveys, 324 surveys were answered during the morning slot, 309 surveys fell under the afternoon slot, and 312 surveys lie within the night time slot. The count of the responses in 3 different slots are almost the same, with the imbalances are primarily due to the difference in study start and completion timings between participants, and some technical glitches on a few days, which triggered more than 3 surveys.

After labelling the time slots, aggregation to get the average value of survey responses per time slot on each day was performed, by taking the average of 4 survey question responses after grouping the records using participant id, day within study and time slot. After this aggregation step, there are 1322 responses, expected based on participants' target study duration and the study plan for 3 survey responses per day. But there are missing responses for 438 time slots, and hence a total of 884 responses after aggregation (one per time slot) remained for analysis from a total of 42 participants. The quantity of answered surveys for 3 different time slots was described in the below section 3.3 (refer to table 3.2). And these aggregated 884 responses are used for the statistical analysis described in section 3.3.3 and for the linear regression in section 3.3.4.

3.3 Study results and discussion:

Several rounds of discussion had been undertaken with other researchers from the clinical team while executing the filtering steps, deciding the study periods associated with each patient from the records (steps described above), and to update the results accordingly. The dataset after performing the filtering steps detailed above was used to generate the final results presented in the chapter. The main results of this chapter explained in the current section 3.3 were obtained using 4 below approaches, namely:

- Quantitative approach - To study adherence patterns in aggregate, per participant and per study day basis.
- Exploratory approach - To understand the mood change patterns from 4 variables aggregated: overall, day wise and on a time slot basis.
- Statistical approach I - To assess differences between the distribution of samples across the 3 different time slots.
- Statistical approach II - To understand the influence of the 3 risk factors on suicidal ideation.
- Predictive analysis using a linear regression model - To predict suicidal ideation using the other 3 risk factors.

3.3.1 Quantitative analysis

Quantitative analysis was performed to secure an understanding of the amount of data available before proceeding with further analysis. This section characterizes the methods and results of the investigation of the participants' adherence patterns on a per-participant basis, per days of study basis, and aggregate basis.

Data set description The count, mean, standard deviation and median of the dataset is noted in Table 3.1 below.

Stats	Depression	Irritability	Connectedness	Suicidality
count	945.000	945.000	945.000	945.000
mean	54.583	32.541	37.190	32.547
standard deviation	31.887	28.700	30.875	31.803
median	56	29	32	35

Table 3.1: Mean and Standard deviation of the variables

A general description of the dataset – such as count of surveys answered, count of participants answering per time slot, etc. – are described in Table 3.2 below.

	Before Aggregation		After Aggregation	Both cases
Time slot	Received survey count	Answered survey count	Answered survey count	Answered participants distinct count
Morning	470	324	293 (33.14%)	42
Afternoon	495	309	289 (32.69%)	42
Night	483	312	302 (34.16%)	41
Total	1448	945	884	42

Table 3.2: General description of dataset per time slot per day

Table 3.2 above shows that, out of the total 1448 responses received by all participants, following 4 steps of filtering (section 3.2.2) and operationalization (section 3.2.3), a total of 945 answered surveys remained for analysis. These surveys were aggregated into time slots (involving 3 distinct time slots across different days), resulting in a total of 884 aggregated survey responses. The overall flow was also depicted in the flowchart depicted in Figure 3.1, as explained in the previous section. Table 3.2 above also describes the count of survey responses per time slot of day before and after aggregation by time slot. Table 3.2 shows that after aggregation of answered survey responses, out of the total 884 responses, 33.14% (293) were morning responses, 32.69% (289) were afternoon responses and 34.16% (302) were nighttime responses. These aggregated responses were used in the exploratory, statistical and predictive analyses explained in the results section, in results based on 6 figures namely: Table 3.3, Figure 3.5, Figure 3.8, Figure 3.9, Figure 3.10 and Figure 3.11.

However, the filtered dataset of total 1448 responses before aggregation was also used in the result section below for studying the adherence pattern of the participants in the initial stage of the study, which are explained in the below sections: section 3.3.1 and section 3.3.1. These 2 sections helped to understand the adherence pattern of all participants in a per-participant and per day within study basis, after filtering out the uncertain data. Also, this adherence graphs helped to understand the quantity and quality of data to make sure that the filtered dataset was fit for further exploratory/statistical analysis.

Survey response adherence analysis

This section explains 4 metrics aggregated based on participant study duration, the total count of surveys that they answered, the total count of days on which they answered at least 1 survey, and the fraction of surveys that they answered during their study period. In addition to that, this section also covers 2 other aggregate metrics on a per day of study basis – the fraction of all surveys answered by participants per day of (their involvement in the) study, and the fraction of participants answering at least one survey per day of (their involvement in the) study. For brevity, from this point forward, for a given participant, the term “per day of study” will be used as a shorthand for “per day of that participant’s involvement in the study”.

Analysis of 3 metrics: participant study duration, the total count of days on which they answered at least 1 survey and the total count of surveys answered The points below give an understanding of each participants' 3 metrics – their study duration, the total days on which they answered at least 1 survey, and the total count of surveys that they answered. It is divided into 2 sets of points. The first set discusses the main observations, which indicates the relation between the total count of days at which the participant answered at least one survey versus the total study duration for that participant. The second set of points discusses the main observations from the relation between the total count of surveys answered and the total study duration. While analysis conducted for this research included associated figures, such figures summarizing per-participant data are omitted from this chapter for privacy reasons.

The first set of observations are mentioned below:

- The study duration of participants ranges from a minimum of 2 days to a maximum of 25 days, while the total count of days on which surveys were answered varies over the range from 2 to 25.

- 47.619% of participants (20 out of total 42) had shown high adherence as judged by answering at least 1 survey on all days during their study period.

- However, from the balance of 52.38% of participants (22 out of total 42) who failed to answer surveys on all days, 23.809% of all participants (10 out of 42) had a minimum of 1 survey response on all days of their study period except their start date. Similarly, some participants have not answered any surveys on their last day – likely because of an early discharge time from the facility – which result in missing responses on either of the first or last day of study.

- From the rest of the 28.57% of participants (12 out of 42), 7.14% (3 participants) were observed as outliers on the graph A, with several days of missing data. But the remaining 21.43% (9 participants) answered at least one survey for far more than 50% of their study duration, and the count of days they had not answered was minimal when compared to their study duration.

- Overall, out of 42 total participants, 71.42% of participants (30 of total 42) answered at least 1 survey for more than 90% of their total study days – a notably high level of adherence. 21.43% of participants (9 out of 42 total) answered for more than 50% of their study duration and only 7.14%(3 participants) show low adherence among the total 42 participants. These findings indicate good study adherence in terms of participant survey answering behaviour within the study. The results also indicate active participation in the study in terms of participants opening the application at least 1 time on each day of the study.

Similarly, below are the main observations that indicate the relation between the total count of surveys answered and the total study duration.

- Total survey count answered per participants varies from a range of 3 to 55 surveys, and exhibits broad covariation with participant study duration.

- A total of 3 surveys were expected per day from each participant (as per the survey trigger schedule logic). Based on that calculation, the maximum count of surveys expected to be answered by each participant was 3 surveys per day that they spent in the study. However, in the case of 5 participants (included in the high compliance category), it was noticed that a total count of more than 3 surveys was answered on some days due to a technical glitch with an early version of Ethica employed; these participants are observed as the outliers. This difference in the count was handled in exploratory and statistical analysis steps by calculating the average of responses within the same time slot as explained in the operationalization section 3.2.3 above. It bears noting, however, that the presence of such outliers is also a testimonial to the exceptional level of adherence exhibited by those participants in the study.

- 4 participants were also noted as outliers, being distinguished by lower adherence in survey answering behaviour. The lesser adherence can be recognized based on the comparison between these participants' total surveys answered count and their expected survey answered count (calculated based on their total study duration). These participants are also distinguished in terms of exhibiting low adherence, with fewer days in which a minimum of 1 survey was answered when compared with their actual study duration.

- Overall, it is noticed that 34 participants (80.95% of total participants) had answered to $\geq$ 50% of their expected survey count (based on 3 surveys answered per day that they remained in the study).

To summarize, a good level of adherence was observed for the majority of participants, as judged in terms of 3 metrics – 1) Total study duration, 2) Total count of days answering a minimum of 1 survey and 3) Total count of surveys answered during their study period.

B. Fraction of surveys answered/survey response rate of participants The fraction of surveys answered by each participant was calculated by dividing the total number of surveys answered by each participant by their total count of surveys received. The total count of surveys received includes the count of answered, expired and cancelled surveys.

The major observations are noted below.

- It was noted that 83.33% of the total of 42 participants (35 out of total 42) have answered at least 50% of the total surveys they received during their study period.

- Among the 35 participants exhibiting a response rate greater than 0.5, 15 demonstrated greater adherence yet, as judged by answering $\geq$ 75% of the total surveys received.

- Among the 7 participants who answered fewer than half the surveys issued to them, 71.42 % (5 participants) have answered $\geq$ 25% of the total received surveys, with only 2 participants answering fewer than one in four.

Overall, the findings demonstrated that the majority of the participants (83.33%) exhibited a moderate to a high level of adherence to survey answering behaviour by answering the majority (and often the large majority) of the total surveys that they received.

Fraction of surveys answered aggregated by days in study

In this section, the changes in the fraction of surveys answered (refer Figure 3.2) and fraction of participants answering (refer Figure 3.3) over study days are analyzed according to the days that the participant has spent in the study, with the result being characterized by the below graphs in Figure 3.2 and Figure 3.3.

Fraction of surveys answered by all participants (by day in study) As mentioned in the paragraph above, the study duration for different participants varies between 2 days and 42 days. As per the study duration recorded in this graph, both the average and median count of days all participants participated in the study is 11.5 days. In Figure 3.2 plots – aggregated across all participants – the fraction of surveys answered for each successive count of days that the participant has spent within the study. This figure thus depicts how the fraction of surveys answered by all participants change over their study day 1 (the first day that they have spent in the study) to day 25.

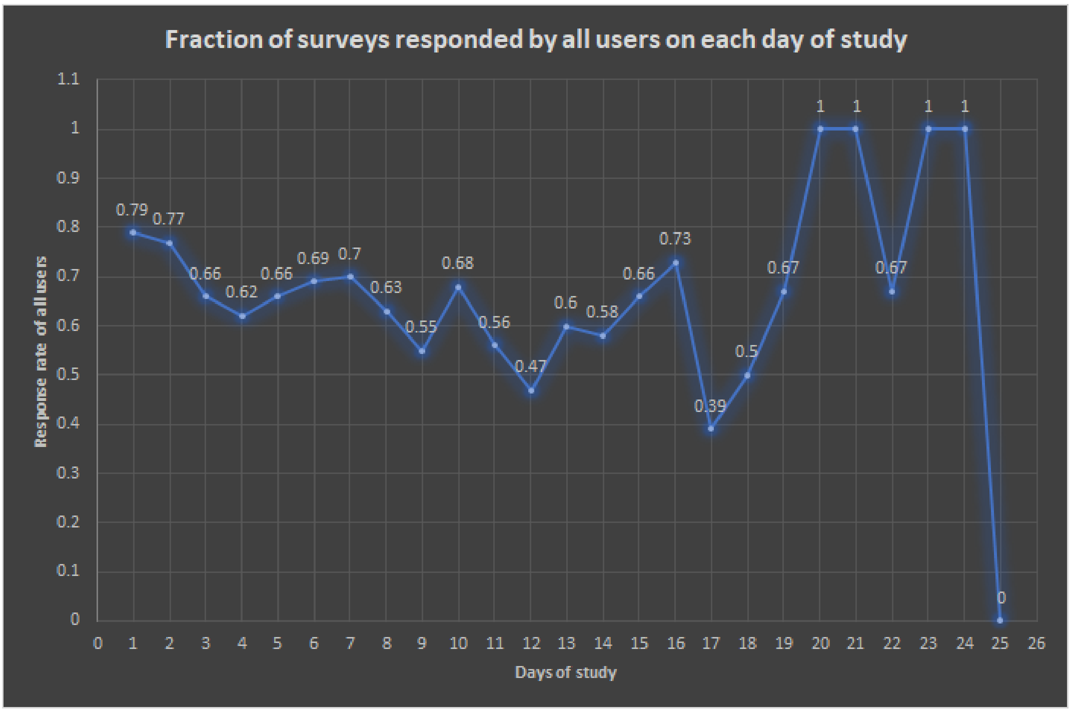


Figure 3.2: Fraction of surveys answered per day by all participants (x-axis: day in study and y-axis: fraction)

Below are the main observations from Figure 3.2:

- Considered over by all users, the fraction of surveys answered shows a decreasing trend over study days.

- One major reason for the fluctuation in the fraction observed is because of the varied study duration of different participants.

- Up to day 11, the overall fraction of surveys answered is always greater than 0.50. Given that the mean study duration was 11.5 days, this indicates that a majority of surveys received by all participants are answered up to approximately the mean study duration.

- Less than a 0.50 response rate was observed only in 3 days – day 12, day 17 and day 25. The decline in day 17 likely reflects the fact that only 5 participants have a duration greater than 17 days, and out of that, the last study day for 4 participants is on day 18.

- Only one participant remains in the study for more than 18 days, and that person answered all surveys received from day 18 to day 24 with a response rate greater than or equal to 0.67. But on the last day, no surveys were answered by that sole remaining participant – yielding the zero value observed on day 25 in the graph.

- Overall, the figure suggests a solid and generally relatively robust level of per-day adherence to survey responses with growing time in the study, albeit one subject to a gently decreasing adherence level over time.

Fraction of participants answering a minimum of 1 survey (by day in study) Figure 3.3 below analyzed how the fraction of participants answering a minimum of one survey evolved over time, starting from their first day to the last day in the study. This fraction is calculated by dividing the total count of participants who answered a minimum of 1 survey on each day of their participation in the study by the total count of participants remained to that day of their participation in the study.

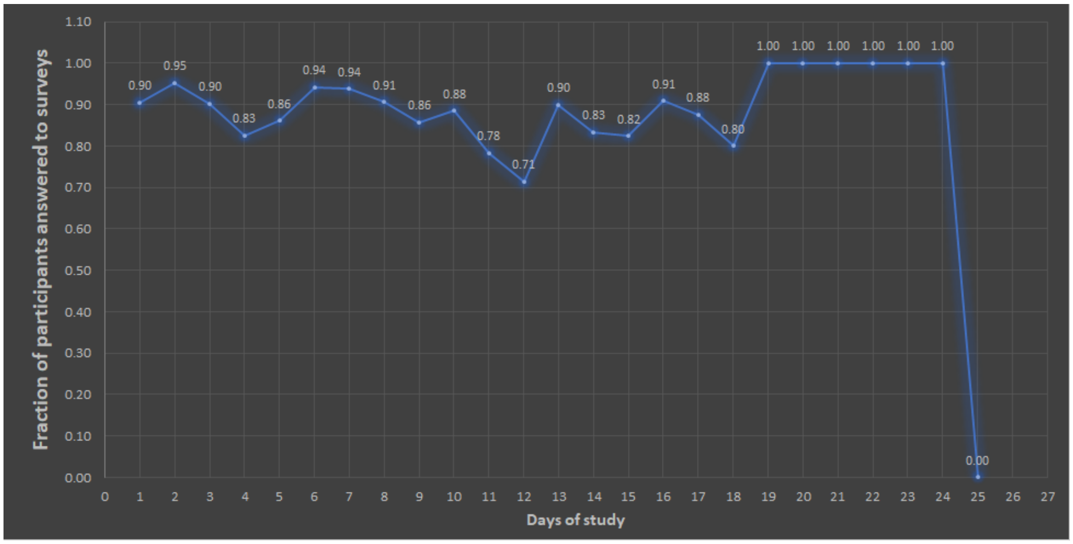


Figure 3.3: Fraction of participants answering at least one survey per day, by day within study (x-axis: days in study and y-axis: fraction)

Below are the main observations with regard to Figure 3.3.

- It is noted that on all days in the study except day 25, more than 70% of participants remaining in the study answered at least 1 survey on each of their study days.

- In the initial 10 days, more than 80% of participants remaining in the study answered a minimum of 1 survey, but on days 11 and 12 there is a precipitous reduction in the answering behaviour, suffering a decline by 17% in absolute terms (from 88% to 71%). But on the next day (the 13th day of their time in the study) it rises suddenly from 0.71 to 0.90 – indicating that 90% of participants present in the study at that time on the 13^{th} day have answered a minimum of 1 survey.

- Small fluctuations are noticed in the exhibited answering behaviour, which may reflect interruptions due to mental conditions or treatment schedules. But no big fluctuations or drops are noticed, and the overall pattern looks promising in the sense that a minimum of 1 response related to mental health condition is collected from almost all participants on almost all days of their participation.

- As noted above, just 1 participant remains in the study for more than 18 days. From days 18 to 24, the graph exhibits an anomalously perfect level of adherence – reflecting the fact that such data points represent only a single participant's responses, where that participant reliably answers a minimum of 1 survey on each such day, with the notable exception his last day in the study (day 25). A sudden drop from fraction of 1.0 to 0.0 is observed on the last day because of no response from that sole remaining participant on the last day.

To summarize, the above three subsections in Section 3.3.1 relate to adherence, and demonstrate sufficiently high level of adherence to make the data suitable for use in further analysis on an aggregate basis. However, since the response rate and duration differ across participants, only an aggregate analysis is performed in this chapter after grouping all participants responses over 3 different time slots during the day and days in the study.

3.3.2 Exploratory analysis

Central tendency measures: average of responses calculated after grouping responses per day and per time slot

To understand the changing pattern of 4 variables per time slot and also per person, the below table Table 3.3 and Figure 3.4 are used.

Average and median of 4 responses from all participants per time slot: The average and median per time slot of the four responses from all participants are showing in Table 3.3 below.

Measure	Morning time slot	Afternoon time slot	Night time slot
Average of Suicidality	31.25	31.75	34.68
Average of Depression	53.24	53.90	56.65
Average of Irritability	31.54	33.17	32.96
Average of Connectedness	34.39	37.83	39.47
Median of Suicidality	30	30	39
Median of Depression	52	53	59
Median of Irritability	24	30	26
Median of Connectedness	27	30	29.5

Table 3.3: Average and median of 4 responses per time slot for all participants

Average of responses per study day: Over days spent in the study, the variation of the average of the four responses from all participants are showing in Figure 3.4.

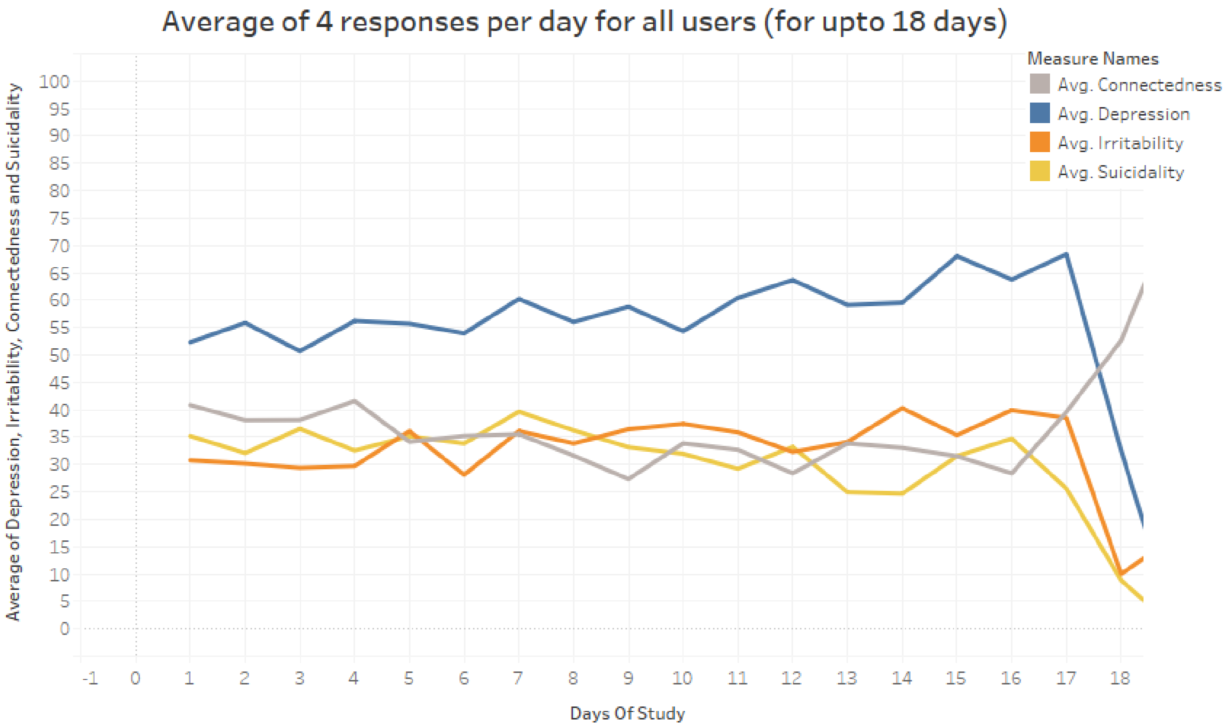


Figure 3.4: Average of 4 responses per day in study for all participants (x-axis: day in study and y-axis: average)

The average (mean) study duration of all participants is 11 days, but Figure 3.4 only shows up to day 18 because after the 18th day, only a single participant remains in the study. The main observations from the above graph are as described below:

- The mean and median study duration for all participants is 11.0 days; hence, the observation points given below primarily focused on the study days up to 15^{th}.

- The average of four variables calculated from all users over their days in the study is plotted in Figure 3.4 above.

- When compared to the other 3 variables, depression stands out in terms of exhibiting larger values, and in terms of a rising trend.

- Suicidality and irritability exhibit almost the same range for the average response throughout the study participant duration. With the exception of the 18th day of participation – when only 1 participant remained in the study – the average connectedness value also shares an almost identical range.

- There is no big time dependent variation (fluctuation) observed in the average of the 4 variables, such as cyclic variation.

- No upward or downward trend was observed in the case of average of irritability and suicidality responses, but a slight downward trend was observed in the case of connectedness when considering only up to day 16 with most users present in the study. By contrast, there was a modest – but potentially troubling – rise in depression values over the course of participants' time in the study.

- There is big drop observed in day 17 and day 18 for depression, irritability and suicidality and a big upward trend is observed for connectedness on that day; these seem to be of questionable significance because of the decline in the count of participants on those days when compared to previous days.

Average of responses per study day and 3 time slots: Over study days, the variation of the average of four responses in 3 different time-of-day slots from all participants are showing in Figure 3.5 below.

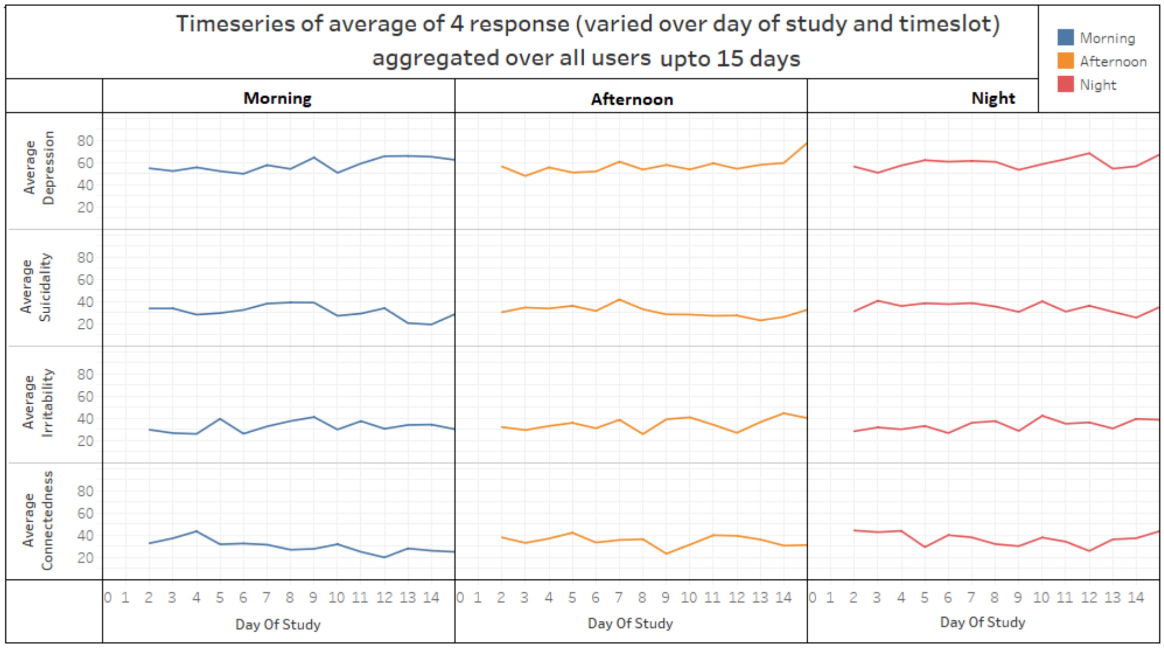


Figure 3.5: Average over all responding participants of 4 reported variables per time-of-day slot and study days (x-axis: days in study and y-axis: average)

In the time series plot shown in Figure 3.5, all responses from all participants are divided into 3 time-of-day slots as mentioned in the operationalization section (Section 3.2.3) above, aggregated by days in study and time-of-day slots, namely morning, afternoon and night, for all users. Only responses up to 15^{th} day in study is used here to plot this graph to reduce problems with small sample size challenges, reflecting the fact that for the majority of participants, duration is up to 15 days, while mean duration is 11.5 days. Also, within this figure, the responses from first day are removed; this reflects the fact that on the first day, the majority of participants responses are removed with a 3 pm filter mentioned in filtering section (ref section 3.2.2), so as to improve the quality of data. As a result, the majority lack morning or afternoon responses after this filtering step. Below are the main observations from Figure 3.5:

- All 3 time-of-day slots exhibit a higher level of depression response compared to the other 3 responses.
- **Depression**: No big fluctuations or clear trends are noticed in the depression range from 3 different time-of-day slots, except a slight increase noticed in the morning and afternoon responses. The range remains between 50 and 70 in all days from 2 to 15.
- **Suicidality**: The range for 3 time-of-day slots over the days in the study is between 20 and 45. When compared to ranges in the morning and afternoon suicidal responses, the values reported in the night slot is slightly higher.
- **Irritability**: A range between 25 and 45 is noticed across the 3 time-of-day slots. No trends are

noticed; while night time readings exhibit a slightly increasing trend with fluctuations, no statistical significance tests have been conducted to evaluate the statistical reliability of this modest trend.

- **Connectedness**: The ranges for 3 responses remain between 45 and 20 over days.

For further analyzing the responses between time-of-day slots from the 4 variables in detail, a boxplot is included below in the next section (Section 3.3.2), and a KS test is performed in Section 3.3.3.

Histogram of 4 Responses (all time slots)

The four responses to the different questions from all participants (without aggregation) are depicted in the histogram below in Figure 3.6. This histogram provides a sense of the distribution of 945 responses received from all users about their reported depression, irritability, connectedness and suicidality level.

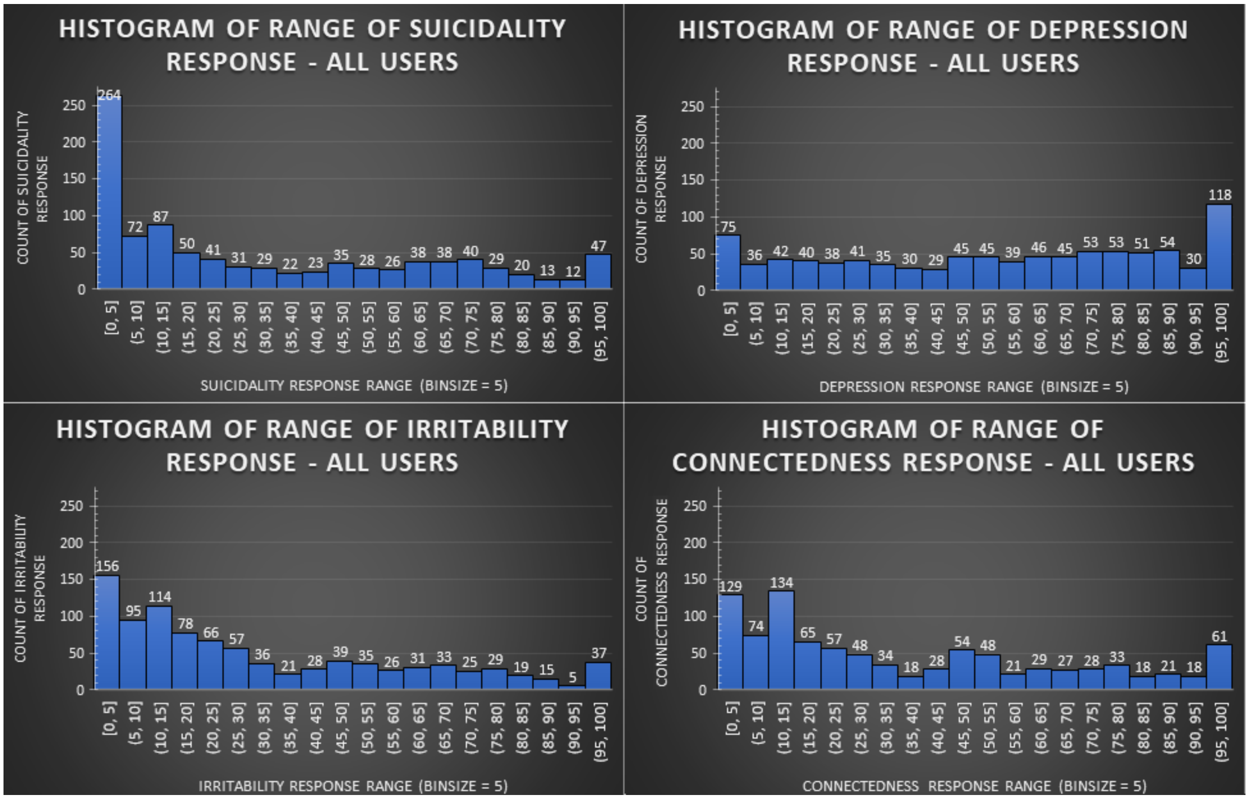


Figure 3.6: Histogram of 4 responses for all participants without aggregation within time slots: 945 responses total and bin size = 5 (x-axis: visual analogue scale reading, y-axis: count of responses)

The main observations from the histograms in Figure 3.6 are noted below.

- The histograms of responses from all 4 questions related to depression, irritability, connectedness and suicidality is highly dispersed, multimodal and asymmetric.

- The histogram from 3 responses – suicidality, irritability and connectedness – is skewed right and exhibits a mean and median less than the midpoint of the scale, whereas the histogram of depression is skewed left, with a mean and median greater than the midpoint of the scale. This suggests a higher range of depression responses when compared to the other 3 variables.

- The histogram of self-reported suicidality located at the top left of Figure 3.6 indicates that 514 responses out of 945 (54.39%) of responses lie in the range 0 to 25, with a markedly large count of reports between 0 and 5. And only 121 responses (12.80% of 945 responses) reported a level of suicidality greater than 75. The right skewed nature of the histogram indicates the lower range of suicidal levels observed in participants.

- The histogram of irritability exhibits some similar patterns to those seen for suicidality, with more responses in the lower ranges of 0 to 25, indicating the lower irritable feeling of participants; however, the two differ in terms of the fact that the histogram for irritability is far less zero-heavy than that for suicidality.

- In the case of connectedness, there are 459 responses (48.57% of overall 945 responses) indicating a low feeling of connectedness (one falling in a range between 0 to 25). It bears emphasis that this cannot necessarily be seen as an indication of social isolation, but is instead an indication of how less socially connected the participants feel regardless of how many they meet or interact within that day.

To conclude, all of these 4 responses demonstrate the non-symmetric nature of the dataset corresponding to depression, irritability, connectedness and suicidality. The results suggest higher levels of reported depression when compared to their irritability and suicidality responses, and indicates a less connected feeling among the participants.

Dispersion measures

In this section, boxplots are used to understand patterns of variability in each response, and to find the differences in the distribution of 4 responses from all participants.

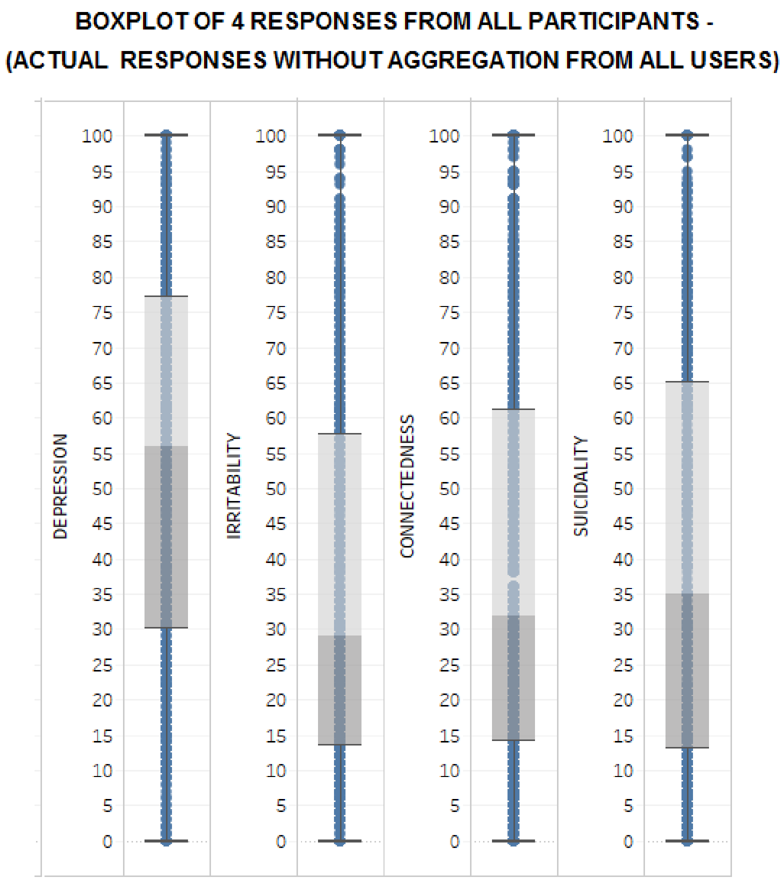


Figure 3.7: Box plot of 4 responses from all participants

Variable	Q1	Q2(median)	Q3	IQR	Range	Shape of dataset
Suicidality	13	35	65	52	0 to 100	skewed right
Depression	30	56	77	47	0 to 100	skewed left
Irritability	13.5	29	57.5	44	0 to 100	skewed right
Connectedness	14	32	61	47	0 to 100	skewed right

Table 3.4: Observed pattern of all responses range obtained from all participants using boxplot

Distribution of all responses across all participants Below are the main observations from the 4 measurands on an overall basis.

- The values of all four of the 4 measurands occupy a range between 0 and 100 without any outliers, lends the possibility of 4 responses having a similar distribution and almost same level of dispersion. The Q1-Q3 interquartile range was similar across the 4 measurands.

- The Q1-Q3 interquartile range for irritability, connectedness and suicidality exhibit high overlap with one another.

- While the median line of depression lies notably higher than for the other responses, it overlapped with the Q1-Q3 range (the "box") (Q1 to Q3 range) of the other 3 measurands. The generally higher values associated with participants' responses with respect to feelings of depression suggest that the participants generally express stronger feelings of depression than the other 3 measurands.

- As per Figure 3.7 and Table 3.4, the extent of the interquartile range of the 4 measurands is very similar, with that of irritability responses being slightly smaller when compared to other 3.

- As observed in the case of above histogram Figure 3.6 and time series Figure 3.4, the higher value of self-reported feelings of depression is also evident in box plot Figure 3.7 below, where reported feelings of depression are associated with a higher median and higher position of the Q1 and Q3 range when compared with the other 3 variables of same data overall range. Also, the median of the reported depression levels was 56 – almost double that of median (29) of Irritability response.

Distribution of responses per time-of-day slot for all participants : Similar to the above, but stratified by time of day, the response range for all participants per time-of-day slot is analyzed in Figure 3.8, and the results from these 3 time-of-day slots are summarized in Table 3.5 below.

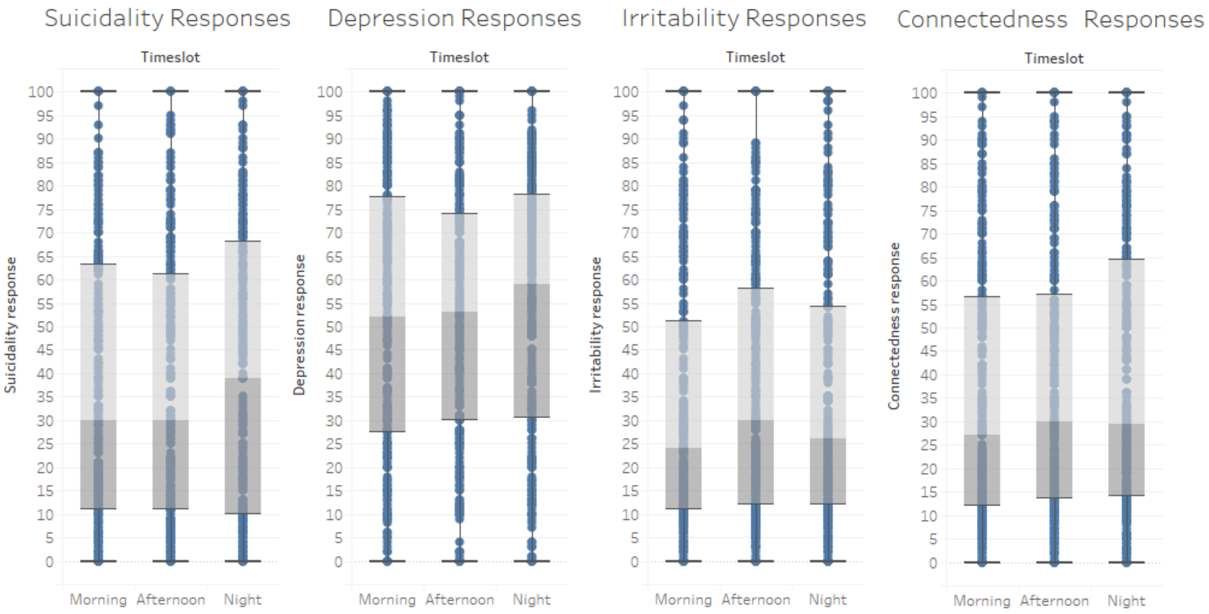


Figure 3.8: Box plot of 4 measurands from all participants on 3 different time-of-day slots (morning, afternoon and night)

Morning slot Responses	Q1	Q2(median)	Q3	IQR	Range	Shape of dataset
Suicidality	11	30	63	52	0 to 100	skewed right
Depression	27.5	52	77.5	50	0 to 100	skewed left
Irritability	11	24	51	40	0 to 100	skewed right
Connectedness	12	27	56.5	44.5	0 to 100	skewed right
Afternoon slot Responses	Q1	Q2(median)	Q3	IQR	Range	Shape of dataset
Suicidality	11	30	61	50	0 to 100	skewed right
Depression	30	53	74	44	0 to 100	skewed left
Irritability	12	30	58	46	0 to 100	skewed right
Connectedness	13.5	30	57	43.5	0 to 100	skewed right
Night slot Responses	Q1	Q2(median)	Q3	IQR	Range	Shape of dataset
Suicidality	10	39	68	58	0 to 100	skewed right
Depression	30.5	59	78	47.5	0 to 100	skewed left
Irritability	12	26	54	42	0 to 100	skewed right
Connectedness	14	29.5	64.5	50.5	0 to 100	skewed right

Table 3.5: Descriptive statistics regarding the 4 measurands across from all participants for each time-of-day slot

Below are the observations from the time-of-day slot based analysis. The statistical significance of any differences commented upon is evaluated in the next section.

- The median and range of suicidality and depression at night exhibit somewhat higher values when compared with the corresponding medians of the morning and afternoon responses. This suggests that participants may show comparatively higher depression and suicidal feelings at night when compared to morning and afternoon.

- In the case of irritability, the median shows somewhat higher values in the afternoon than in the morning and at night.

- In the case of connectedness, the medians are relatively similar, with slightly lower median connectedness in the morning.

- As above and across time-of-day slots, participants reported somewhat stronger feelings of depression than for the other 3 responses.

3.3.3 Statistical analysis

Statistical analysis I: Kolmogorov–Smirnov test (KS test)

The nonparametric Kolmogorov–Smirnov (KS) test was used to test (separately for each of the 4 measurands, and each unordered pair of time-of-day slots for that measurand) whether we can with confidence reject the null hypothesis that values for each of the times of day considered are drawn from same distribution – for example, that the values for depression in the morning and night follow the same distribution. Within these tests, I elected to apply a 0.05 significance level. The results are shown in tables 3.6, 3.7, 3.8, 3.9. An illustration of empirical cumulative distribution functions (ECDF) associated with each measurand for each time of day – which form the basis for the KS test – is illustrated in Figure 3.9.

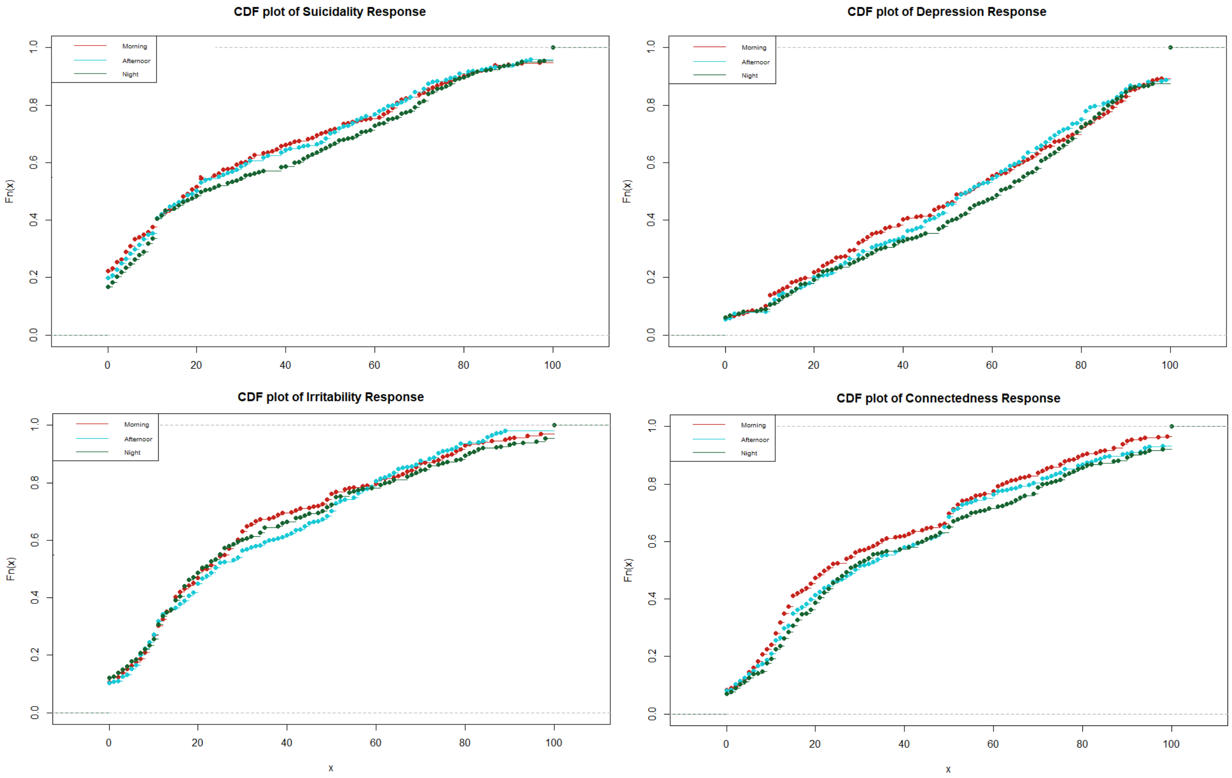


Figure 3.9: ECDF plot to compare summary statistics of time-of-day based responses: Suicidality (top left), Depression (top right), Irritability (bottom left) and connectedness (bottom right)

The empirical cumulative distribution function plots for each of the 4 measurands and (via colour) time of day are shown in Figure 3.9 above. Green colour represents night time responses, blue the afternoon responses and red denotes morning responses; the left-heavy character of the suicidality, irritability, and connectedness is evident through the associated ECDFs – regardless of time of day – while the highly dispersed. By contrast, the ECDFs for depression – regardless of time of day – suggest an empirical distribution that is close to uniform over most of its range, with a pronounced mass at the upper extreme.

Here, through, separately for each of the 4 measurands, and for each unordered pair of time-of-day slots for that measurand, I use the KS two sample test to evaluate if we can reject the hypothesis that the two variables (e.g., depression in the morning, depression at night) are drawn from the same underlying distribution. This test makes use of the cumulative distributions of 2 data sets[51]. Table 3.6 shows the KS test results for suicidal responses from each unordered pair of the 3 different times of the day. Similarly, tables 3.7, 3.8, 3.9 show the corresponding KS test results for depression, irritability, and connectedness. The tables include K-S test results in terms of both D statistics and p-values. The D statistic (also called the K-S statistic) characterizes the maximum value (supremum) of the absolute difference in the empirical distribution functions of the two samples. The smaller – and closer to zero – the D value was, the greater the probability that the two samples are drawn from the same distribution.

time slot	After Noon	Night
Morning	D = 0.035599, p-value = 0.9881	D = 0.080128, p-value = 0.2592
After Noon	"Not Applicable"	D = 0.060711, p-value = 0.6163

Table 3.6: Suicidality : Kolmogorov - Smirnov Test Result

In the case of the suicidality responses, the p-value for all KS test results performed between the 3 times of day exceeds the significance level of 0.05. And the D values are showing as 0.03, 0.08 and 0.06, which are notably close to zero. Hence we accept the null hypothesis that the samples used for each of the 2-sample KS tests involved – comparing suicidality responses in each unordered pair of morning, afternoon and night – were drawn from the same distribution.

time slot	After Noon	Night
Morning	D = 0.061429, p-value = 0.5894	D = 0.083808, p-value = 0.2142
After Noon	"Not Applicable"	D = 0.073407,p-value = 0.3729

Table 3.7: Depression: Kolmogorov - Smirnov Test Result

In the case of the depression response, the p-value for all KS test results performed between unordered pairs of 3 time slots also exceeds the significance level of 0.05. The D values – 0.06, 0.08 and 0.07 – are further notably close to zero. In accordance with the chosen significance level, I accepted the null hypothesis that the samples used for 2-sample KS tests between each unordered pair of times of day were drawn from the same distribution.

time slot	After Noon	Night
Morning	D = 0.090315, p-value = 0.1515	D = 0.054487, p-value = 0.7329
After Noon	“Not Applicable”	D = 0.090315, p-value = 0.718

Table 3.8: Irritability: Kolmogorov - Smirnov Test Result

Likewise, in the case of irritability responses, the p-value for all KS test results performed between all unordered pairs of times of day exceeds the significance level of 0.05. The corresponding D values – 0.09, 0.05 and 0.09 – also remain close to zero. In accordance with chosen significance level, here I also accepted the null hypotheses associated with each test.

time slot	After Noon	Night
Morning	D = 0.066013, p-value = 0.4959	D = 0.1028, p-value = 0.0695
After Noon	“Not Applicable”	D = 0.05234, p-value = 0.7887

Table 3.9: Connectedness: Kolmogorov - Smirnov Test Result

Finally, in the case of connectedness responses, the p-value for all KS test results performed between the 3 unordered pairs of time slots all exceed the chosen significance level of 0.05. However, the P-value between morning and night is notable as being (barely) borderline significant at 0.0695; the corresponding value of the D-statistic is 0.1, which suggests that while the null hypothesis cannot be rejected at the chosen level of significance, there may be some difference between morning and night responses. By contrast, for the other 2 test results – between morning-afternoon and afternoon-night –the p-values are far larger and D statistic was closer to zero, suggesting a clear need to accept the associated null hypotheses.

To conclude, no statistically significant differences are noticed while comparing responses across time-of-day slots for each of the 4 variables, and barely borderline significance was only encountered for one test. Despite the visual differences noted above, for each of the 4 measurements, it is not possible to confidently posit any differences in underlying distributions across different times of day at the chosen significance level.

Statistical analysis II: Correlation between variables and influence of 3 risk factors on suicidal ideation

Pearson correlation statistics between the four variables are examined in this section; Figure 3.10 also presents with accompanying scatter plot matrices created using the R program base function pairs(). The x-axis and y-axis of that scatterplot matrix feature the 4 measurands in the order of depression, irritability, connectedness and suicidality. Diagonals in that scatterplot matrix display histograms of responses with bin size 10.

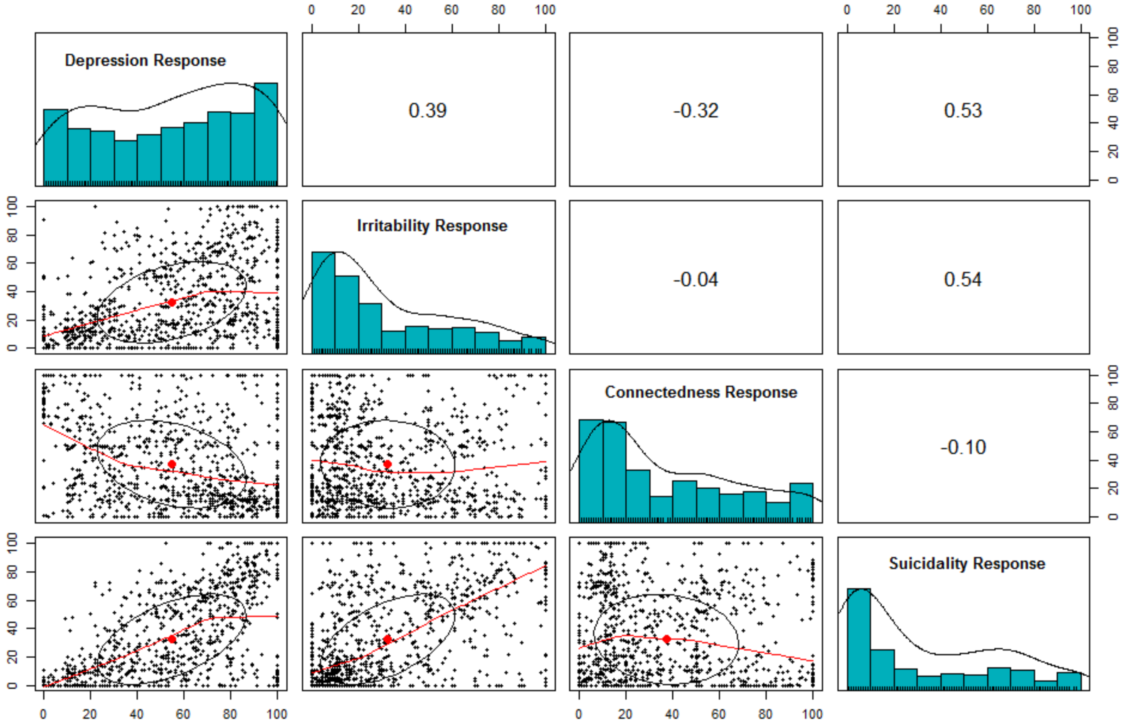


Figure 3.10: Scatterplot matrix of responses for 4 measurands from all participants

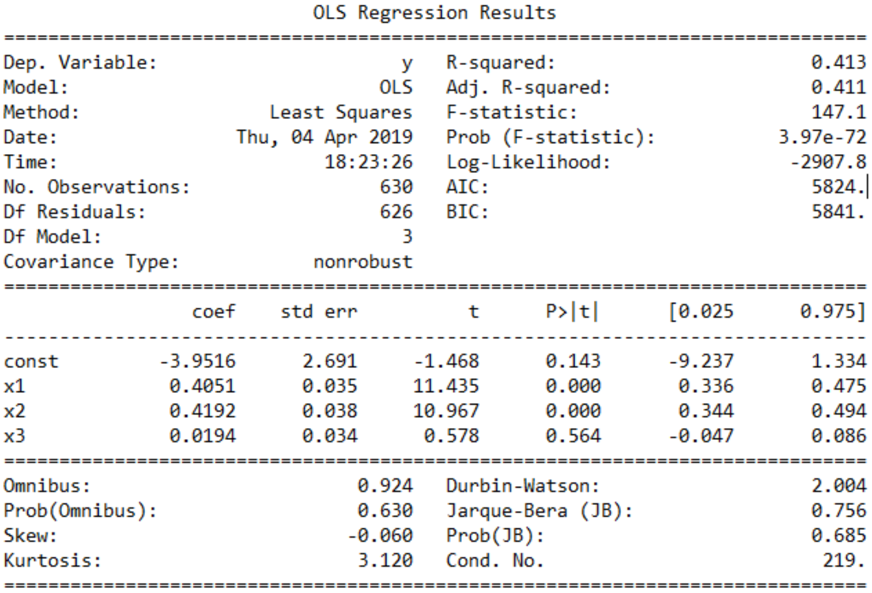
Some observations about the relationship between variables as per Figure 3.10 are made below. These comments need to be interpreted with great caution given the fact that the statistical significance of the apparent associations has not been demonstrated.

- A moderate positive linear correlation of 0.54 exist between irritability and suicidality, a moderate positive linear correlation of 0.53 exists between depression and suicidality. Also, a positive linear correlation of 0.39 was observed between depression and irritability.
- A negative linear correlation of -0.32 was showing between connectedness and depression and a negative weak nonlinear relation of -0.10 was showing between connectedness and suicidality.
- Connectedness and irritability appear to show a nonlinear relationship with a negative correlation of -0.04.
- Some clusters near zero values are noticed in the case of the suicidality vs. irritability scatter plot and suicidality vs. connectedness.
- An apparent positive association was observed between suicidality and 2 other risk factors and a weak negative association was observed with connectedness. This suggests some possible associations of the risk factors with suicidal ideation.

3.3.4 Predictive analysis: Linear regression

Linear regression was performed on the dataset after aggregating the responses over day and time of day for all participants, with suicidality as the dependent variable (Y values) and the 3 variables of depression, irritability and connectedness as the independent variable (X values). Rather than using traditional statistical methodologies such as stepwise regression, preceding the multi-variate regression with screening of covariates based on univariate analysis and potential for confounding, the approach explored here followed machine learning practice, involving cross-validation.

Altogether, the dataset has 884 responses, which were divided as a single division into 30% of data as a testing set and 70% as the training set. The Scikit-Learn method train_test_split() was used for undertaking this partition and statsmodels.api.OLS library was used for linear regression in python. The R^2 score and Root mean square error (RMSE) value for the model was recorded in Table 3.10 below for comparison. Ordinary least square (OLS) method in linear regression was used to create the model which tries to minimize the sum of square error between the real data point Y and predicted data point Y_p [52]. Here, the first OLS model was created, using depression, irritability and connectedness as covariates(continuous predictor variable): x1, x2, and x3, respectively. The term "const" is the offset(constant term) and the beta values are showing under "coef" for all covariates in the result Figure 3.11 below.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.413
Model:                            OLS   Adj. R-squared:                  0.411
Method:                 Least Squares   F-statistic:                     147.1
Date:                Thu, 04 Apr 2019   Prob (F-statistic):           3.97e-72
Time:                        18:23:26   Log-Likelihood:                -2907.8
No. Observations:                 630   AIC:                             5824.
Df Residuals:                     626   BIC:                             5841.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.9516      2.691     -1.468      0.143      -9.237       1.334
x1             0.4051      0.035     11.435      0.000       0.336       0.475
x2             0.4192      0.038     10.967      0.000       0.344       0.494
x3             0.0194      0.034      0.578      0.564      -0.047       0.086
==============================================================================
Omnibus:                        0.924   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.630   Jarque-Bera (JB):                0.756
Skew:                          -0.060   Prob(JB):                        0.685
Kurtosis:                       3.120   Cond. No.                         219.
==============================================================================
```

Figure 3.11: Result of Ordinary Least Squares

Then the relationship between different univariate and multivariate combinations and suicidality as an outcome was also calculated; results are summarized in Table 3.10.

Regression Type	Independent variables	Dependent variable	RMSE value	R^2 score	Adj. R^2 score	Prob (F-test)	AIC	BIC
Multiple linear	Depression, Irritability, Connectedness	Suicidality	24.24	41.30%	41.10%	3.97e-72	5824.	5841.
Multiple linear	Depression, Irritability	Suicidality	24.29	41.30%	41.10%	2.79e-73	5822.	5835.
Multiple linear	Depression, Connectedness	Suicidality	27.19	30.10%	29.80%	2.00e-49	5932.	5946.
Multiple linear	Irritability, Connectedness	Suicidality	26.37	29.10%	28.90%	1.59e-47	5941.	5955.
Simple linear	Depression	Suicidality	27.35	29.80%	29.70%	3.29e-50	5933.	5942.
Simple linear	Irritability	Suicidality	26.22	27.90%	27.80%	1.71e-46	5950.	5959.
Simple linear	Connectedness	Suicidality	31.64	1.80%	1.60%	0.000741	6144.	6153.

Table 3.10: Linear Regression result

The R^2 value of multiple linear regression with suicidality as the dependent variable and the 3 other measurands – depression, irritability and connectedness – as the independent variables was 41.30% which indicates that the model was able to explain 41.30% of the variance using the model, demonstrating a moderate degree of capacity to predict the outcome. Even though the RMSE (Root mean square error) value of the model was 24.24 and R^2 value was moderate, our focus here was to find the relationship between variables rather than prediction, hence the moderate R^2 value was acceptable [53]. Moreover, there are influences of many other factors on suicidality other than the 3 variables considered here, making it surprising if these 3 variables were to explain a high percentage variation in suicidality. To better understand the results, beyond the multiple linear regression with 3 risk factors as the feature set, the regression analysis was repeated with fewer features (by considering pairs of covariates) and also performed at a univariate level. The results are present in Table 3.10. The R^2 value remains as 41.30 in the case of the variable including all covariates as well as that dropping the connectedness covariate; for all other models examined, the resulting R^2 value below that for the first scenario with 3 independent variables [53]. For the univariate models, models including depression and irritability as the sole covariate demonstrate almost the same R^2 value; by contrast, the univariate modelling including connectedness as the sole covariate exhibits a very low value R^2 value of 1.80%.

3.4 Weakness

One weakness of this study related to missing data. Some participants did not answer surveys during some days, and hence their responses on those particular days are missing. Another weakness observed in the data used for this study was the inconsistency observed in the count of surveys received by 7 participants in a few days. In detail, seven participants answered more than 3 surveys on some of their study days. The expected survey count per day for all participants was 3 surveys spanning 3 different times of day, but these 7 participants have received more surveys than planned, to a maximum of 8 and 9 times on a single day (two participants whose identifiers are omitted here for privacy, but which received such surveys on their study days 2 and 6). Another limitation is the omission of analyses to assess the statistical significance of trends over time, and nonlinear models such as those built from neural networks which are not analyzed in this chapter. Given the significant challenge of missing data, the use of data imputation techniques could also provide one approach for trying to limit the effects of missing data on future analytics. Also, in the multiple linear regression, N-fold cross validation was not used; addition of that methodology constitutes a high implementation priority for future work to publish elements of the analysis covered in this chapter. Another concern is related to data quantity, and particularly the relatively small sample size of this study, which strongly limits the generalizability of these results to other populations. While there was a significant number of longitudinal data points gathered per person within each of the four central self-reported measures considered here, with the average study duration of participants being just 11 days, many temporal patterns were incompletely revealed in the time series data. If data were to be collected for a larger duration – for example, 1 month continuously – then the trends and prediction of time series could be more effectively performed. Also for a future work, continuing the study even after patients are discharged from the facility (say for a period of 2 weeks) would be helpful to compare the pattern of responses reported from the facility and outside the facility. Also, the author of this thesis had permission only to access and analyze the data collected by smartphones using "Ethica" software, and lacked access to the clinical data related to patient demographics such as their age, gender, the reason for hospitalization (mental-illness condition), etc.. Analyses based on such patient characteristics are therefore not performed in this thesis chapter. Also, the small sample size of this study – including just 42 patients after filtering – is insufficient to support strong conclusions about the psychiatric inpatient population as a whole. Taken all of these factors into consideration, the findings from this chapter exhibit limited generalizability. However, the analytical methodologies implemented in this chapter can be adapted for larger future versions of related studies using the "Ethica" platform to arrive at stronger conclusions.

3.5 Conclusion

To conclude, the main observation from the analysis are summarized below:

- As per the first paragraph, an overwhelming fraction of participants (39 participants – 92.85% of the total) answered a minimum of 1 survey for more than 50% of their study duration, with 20 participants (47.61%) exhibiting high adherence, as judged by answering a minimum of 1 survey on all days of their study duration.

- As per the adherence analysis in Section 3.3.1, the large majority of participants (35 participants – 83.33% of the total participants) answered more than 50% of the total surveys received by them during the study period. Among those, 15 participants (35.71%) answered to more than 75% of the total surveys that they have received.

- As per the adherence graph in Figure 3.2, the fraction of total surveys answered by all participants decreases over participants' time in the study, but up to the median duration (here, day 11), most surveys were still being answered.

- As per the adherence graph in Figure 3.3, it was noticed that with the special case of the last day, more than 70% of participants remaining in the study on any given day in the study have answered a minimum of 1 survey.

- As per the histogram in Figure 3.6, the distributions associated with all 4 measurands are non-symmetrical. Suicidality, irritability and connectedness are right skewed, while depression was left skewed, exhibiting higher levels of feelings of depression reported by participants during the study period when compared to other responses.

- As per the time series plot in Figure 3.4, no pronounced time dependent or cyclical fluctuations are noticed in the mean value of the 4 measurands over days. Increasing study duration in the future would be more effective to analyze this time dependent nature of responses.

- As per the time-of-day based analysis from time series plotted using the aggregated average on Figure 3.5, no big trends are noticed per time of day on the basis of the time series analysis.

- From the box plot shown in Figure 3.7 and time-of-day based box plot in Figure 3.8, higher ranges of depression were clear from the median and plots, when compared with the 3 other measurands. From the overall responses plotted in Figure 3.7, the Q1 to Q3 range for 3 responses - depression, irritability and connectedness – were highly overlapped. From the time slot based box-plot in Figure 3.8, a higher median for depression and suicidal level in night time slot, irritability in the afternoon time slot and a comparatively lower connectedness median in morning time slot from the participants were noted; however, such observations must be interpreted with great caution due to the hypothesis tests below.

- Histogram (Figure 3.6) and box plots (Figure 3.7 and Figure 3.8) and time series (Figure 3.4) suggested that participant depression level tends to have higher ranges even though their irritability and suicidality were observed to be lower.

- The KS test results summarized in 4 tables – Table 3.6, Table 3.7, Table 3.8 and Table 3.9 – suggest that it is not safe to assume that there are differences between the underlying distribution applying at different times of day for a given measurand at a 0.05 significance level. Most such comparisons clearly suggested accepting the null significance; marginally borderline results ($p = 0.0695$) were only observed within the case of the ks-test between connectedness night and morning responses.

- As per the scatter plot matrix in Figure 3.10, while the statistical significance of the trends requires great caution and invites critical scrutiny, a moderate level of positive correlation was observed between suicidality and 2 other risk factors – irritability and depression. Also – and again requiring strong caution and more rigorous hypothesis testing – a moderate level correlation was observed between depression and irritability variables, and a weak negative correlation was observed between connectedness and 2 other variables namely depression and suicidality.

- Linear regression results conducted according to machine learning practice shows a moderate level of association between suicidal ideation (as the dependent variable) and depression and irritability. While evaluating this model N-fold cross validation was not performed for the which need to be implemented as future work.

To summarize, the study design was successful in collecting responses from patients over time, and a similar study design for the participants which can continue the data collection even after treatment period would be helpful in understanding their behavioural patterns during and after their treatment period, which can be considered as an attractive avenue for future work. Also, an increased study duration will help to understand the patterns more clearly and make the accuracy of prediction stronger with added data points. While statistical significance testing suggests no big differences in the distribution of responses between time slots for a given variable, it was observed that depression responses tend to have higher levels than for the other measurands. Rigorous statistical testing is required to evaluate apparent covariations among the responses across measurands, but results using simple linear regression do suggest a moderate capacity to predict self-reported suicidality responses on the basis of responses regarding feelings of depression and irritability.

Chapter 4

Feasibility of a Mobile Health Monitoring Application to Capture Individual-Level Behaviours in HIV+ Persons

4.1 Introduction

Saskatchewan has the highest rate of HIV among Canadian provinces. As a communicable and historically lethal disease, it is highly important to find ways to control HIV spread. There are various risk factors influential in the spread of HIV. While the most important risk groups are injection drug users (IDU), men who have sex with men (MSM), female sex workers (FSW), in the era of Highly Active Antiretroviral Therapy (HAART), one of the foremost risks for spread lies in risks of transmission elevated by lapses in medication adherence. Often people living with HIV are struggling with complex medical needs, socioeconomic challenges, mental health issues, social needs, etc., and a systematic way of investigating barriers to medication adherence will be helpful in understanding the generative mechanisms underlying a central risk factor for the further spread of HIV. Given the distinct character and life circumstances often associated with the risk groups noted above, it is important that studies seeking to understand such barriers include individuals drawn from a diversity of backgrounds.

This chapter analyzes the preliminary study adherence results from a smartphone based study conducted jointly with and geographically in the Regina Qu'Appelle Health Region (a part of what is now amalgamated as the provincial Saskatchewan Health Authority). This study continuously monitored HIV patients to understand difficulties preventing them from maintaining adherence to medication guidelines and other protective behaviours, which additionally play a major role in shaping risk of HIV transmission.

Past studies have suggested that circumstances associated with patient mental health, physical health, treatment patterns, and other factors can be studied in detail with the help of self-reported data, subject to the reporting burden being appropriate for participant circumstances. So as to allow for easier capturing of patient context (e.g., using photos, geotagging), and self-reporting (e.g., via option for audio recording of responses), this work applies the mobile based data collection tool Ethica introduced above 1.2.2 to collect participant data. This chapter discusses analyses undertaken with the collected data through the specific lens of feasibility assessment – reflecting the central uncertainties faced with respect to whether the reporting regimen employed would be acceptable across the diverse patient populations. To address such feasibility, such analyses place a particular focus on study adherence.

4.1.1 Background & literature review

HIV (Human Immunodeficiency Virus) is a virus that attacks the human immune system. If left untreated for several years, it can lead to progressive and (absent aggressive treatment) eventually lethal damage to the immune system [54]. HIV has imposed a heavy burden on Saskatchewan (SK) in recent decades. As per reports in 2017, while the highest number (n=935) and proportion (38.9%) of reported HIV cases was in Ontario, the highest provincial HIV diagnosis rate is in Saskatchewan, which is 15.5 per 100,000 population – about 2.4 times higher than the rate of 6.5 per 100,000 for Canada as a whole [55, 56]. The rate per 100,000 in SK decreased during 2013 and 2014 to 11.4 and 9.8 [55], but that favourable trend was reversed, with a rise to 13.9 in 2015, and 15.5 in 2017 – rates similar to those in 2008 to 2012 [55]. Most of the newly diagnosed in 2017 were from Regina, Saskatoon, and Prince Albert, but there was also an increase in the proportion of cases from rural areas in 2010 and this proportion has fluctuated since then [55]. Laboratory testing for HIV had increased by 71% over the decade leading up to the 2017 report [57], and 53% of the individuals diagnosed with HIV from 2008 to 2017 remained alive [57]. Various factors have played an important role in the spread of HIV in Saskatchewan. As per the reports, the primary risk factors noticed amongst newly diagnosed people in Saskatchewan are as follows: 67% reported injection drug use (IDU), 8% were men who reported having sex with men (MSM), 20% likely acquired the infection via heterosexual sex, and 5% were individuals associated with other risk factors [55].

Several studies have been performed to understand the influence of various factors on the spread of HIV epidemics in Canada and U.S. and to take measures to control the spread of epidemics. Many of them pointed that the HIV-related stigma is an important barrier that discourages patients from social interactions, disclosing their HIV status to family and friends, or even prevents them from getting tested or taking medications etc., which in turn results in poorer clinical outcomes, elevated risk of mental health challenges such as depression, etc., [58–62]. It can also adversely impacts the proper utilization of care facilities, adherence to medication and the quality of life of HIV patients [58, 63, 64]. This stigma affects not only the lives of people living with HIV/AIDS but it also affects others who are at risk of HIV infection [62]. The continuous improvement in HIV treatment resulted in a direct delay in the disease progression of HIV patients [65–67]. Antiretroviral therapy (ART) helps to turn HIV condition from a fatal stage into a manageable stage and thus helps the people living with HIV (PLWHIV) to live longer [64, 65, 68]. Studies focus on high-income countries shows that the HIV infected patients of age 20 years and receiving highly active antiretroviral therapy (HAART), have a life expectancy of about two-third of the general population [65, 69]. Hence, conjoint efforts by clinicians and patients are needed to challenge the deep-rooted practices and behaviours existing in the health care system and to improve the quality of care [70]. Patients know their condition and the impact of treatment and disease on their lives better than clinicians and know how better designed services could improve their situations [70, 71]. Hence involving patients in research and getting their feedback while designing and implementing a study to control diseases would be helpful to understanding the situation better and improving the interest of patients to continue with their treatments. This encouragement of patient and public engagement in the research studies have been rapidly increasing in the past decade [70, 72, 73].

While people living with HIV in SK come from diverse backgrounds, important subsets of those individuals struggle with social and health needs and risk behaviours such as drug usage, sex work, etc. A clear understanding regarding the barriers that prevent patients from maintaining a healthy medication regimen, care-seeking and protective behaviours may help put in place changes and supports which help reduce the spread of HIV in Saskatchewan. Design of an appropriate study within this area will secure great value through involvement of social workers, clinical teams and – most importantly – patients who literally play the key role in a way that they are the only people who can let other team members know about their situation and needs, and who can provide guidance as to data collection regimens that are likely to be appropriate for the broader HIV patient population. Such patient partners offer knowledge and expertise that can provide insights about the needs and issues faced by other people who are not able to communicate on their own behalf [74]. The involvement of those with lived experience in the design and the conduct of research thus constitutes an important step to realize the full potential of diverse studies. As per the CIHR Strategy for Patient-Oriented Research (SPOR), the major goal of patient-oriented research is to achieve benefits that matter to patients by including them as an important partner in health care, and thereby improve their health by providing the right treatment at right time, improved access to the health care system, and to contribute to improved cost effectiveness of the health care system, etc. [74]. Research in patient-oriented health methods have suggested needs for involvement of those with lived experience not only in terms of participation, but also to help improve the formulation of study design, decision making, and quality by providing feedback on study operation and interpretation.

This study sought to involve existing HIV patients in designing an initial pilot study to assess the feasibility of a future broader study of barriers to medication adherence, and to inform an understanding of aspects of study design that pose risks to participant involvement, either as a whole or within certain classes of patients. For the feasibility study described here, 15 patients from Regina and Saskatoon were involved in study planning, design, and decision making during study operation.

4.2 System description & methodology

4.2.1 Study design

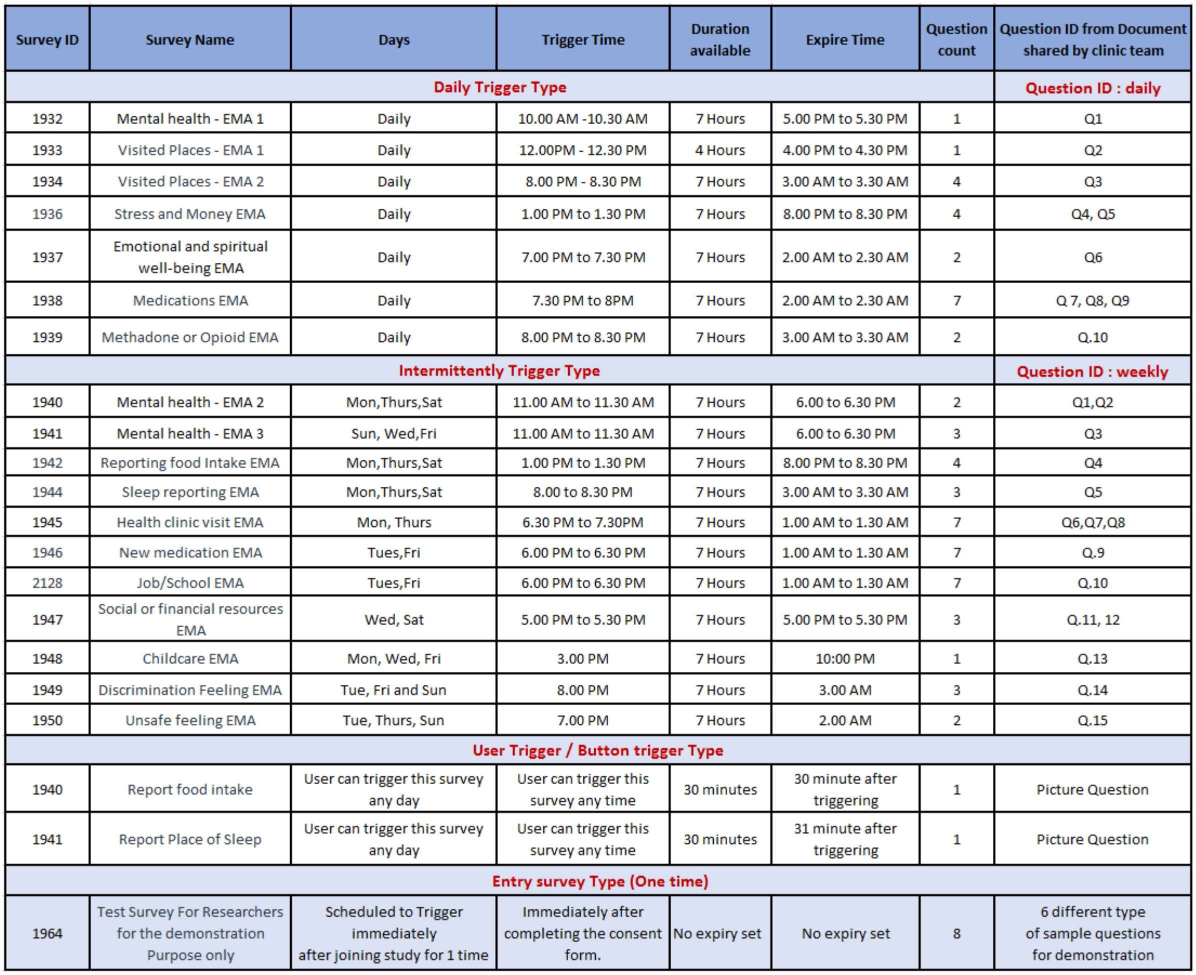
The feasibility study was set up using the smartphone based application named Ethica, described in Section 1.2.2. Several rounds of discussions were conducted before finalizing elements of the study design, including study duration, sensors and the character, frequency and content of questionnaires used for surveys, which were developed alongside patient advisors and community-based representatives. As noted above, and reflecting the study's commitment to patient orientation, feedback from HIV patients who are part of the research team and other patients played a vital role in the study designs. This study was approved by the Regina Qu'Appelle Health Region human ethics review board (RQHR file REB-17-47), with the stipulation from the RQHR ethics review board that analysis of the data collected by the study was to be limited to assessing adherence, rather than to draw substantive health insight concerning barriers to HIV treatment. While this restriction on analysis scope substantially limited the breadth of insights that could be secured through the study, the principal investigators (Drs. Wong and Osgood) decided that sufficient scientific value would likely still be delivered to merit conducting the study. This study was additionally approved by the University of Saskatchewan Behavioural Ethics Review Board (UofS Beh 17-179). The author of this thesis was specified as a student working on this project in the REB file associated with this study.

Fifteen PLWHIV from southern Saskatchewan were recruited as participants in the study within the Infectious Diseases Clinic of Regina General Hospital. The study design emphasized recruitment of a broad set of participants; the fact that some participants were homeless was discussed in study team meetings in which the author participated, particularly because it had bearing on study planning and operation. Before recruitment, patients were provided with a consent form informing them of study goals and procedures, the sources of data to be collected in the study, and regarding the option to "snooze" (pause) data collection completely for an hour. They were further informed about the option of disabling sensor data such as GPS at any time.

Participants who lacked an unlimited data plan received a pre-paid data plan to use with their personal or study-provided smartphones. Ten smartphones were provided to participants who were in need of one. Compensation was commensurate with adherence, with participants being remunerated $0.15 for each EMA completed, to a maximum of $50 per 30-day cycle. Using the participant adherence section of the web-based Ethica dashboard, the study coordinator reviewed participant adherence information once per week to determine the extent to which participants were completing EMAs. That coordinator further attempted to make contact with the participant if there was a sudden decrease in the number of EMAs completed, so as to help ensure that there were no technical problems at hand, and that the participant's phone remained available and in working condition.

The thesis author was responsible for the specification, configuration and refinement of the study in Ethica, with that design evolving in significant ways across five iterations of co-design. For each participant, the study was conducted for a duration of six months. The study employed a combination of twenty survey instruments – includes photos and audio recordings – and six sensors. This study employed a later and more evolved form of Ethica than that employed in Chapter 3, which further allowed for geotagging of surveys; in this study, all surveys were geotagged.

These instruments were set up with different timings to frequently monitor patients' mental health, adherence to medicine, food intake, social interactions, environmental conditions, etc., with each such survey being specified by the thesis author. A table describing the list of surveys set up in Ethica for this study, their associated triggering logic type, count of questions in each survey, etc., is described in Figure 4.1 below.



Survey ID	Survey Name	Days	Trigger Time	Duration available	Expire Time	Question count	Question ID from Document shared by clinic team
Daily Trigger Type							Question ID : daily
1932	Mental health - EMA 1	Daily	10.00 AM -10.30 AM	7 Hours	5.00 PM to 5.30 PM	1	Q1
1933	Visited Places - EMA 1	Daily	12.00PM - 12.30 PM	4 Hours	4.00 PM to 4.30 PM	1	Q2
1934	Visited Places - EMA 2	Daily	8.00 PM - 8.30 PM	7 Hours	3.00 AM to 3.30 AM	4	Q3
1936	Stress and Money EMA	Daily	1.00 PM to 1.30 PM	7 Hours	8.00 PM to 8.30 PM	4	Q4, Q5
1937	Emotional and spiritual well-being EMA	Daily	7.00 PM to 7.30 PM	7 Hours	2.00 AM to 2.30 AM	2	Q6
1938	Medications EMA	Daily	7.30 PM to 8PM	7 Hours	2.00 AM to 2.30 AM	7	Q 7, Q8, Q9
1939	Methadone or Opioid EMA	Daily	8.00 PM to 8.30 PM	7 Hours	3.00 AM to 3.30 AM	2	Q.10
Intermittently Trigger Type							Question ID : weekly
1940	Mental health - EMA 2	Mon,Thurs,Sat	11.00 AM to 11.30 AM	7 Hours	6.00 to 6.30 PM	2	Q1,Q2
1941	Mental health - EMA 3	Sun, Wed,Fri	11.00 AM to 11.30 AM	7 Hours	6.00 to 6.30 PM	3	Q3
1942	Reporting food Intake EMA	Mon,Thurs,Sat	1.00 PM to 1.30 PM	7 Hours	8.00 PM to 8.30 PM	4	Q4
1944	Sleep reporting EMA	Mon,Thurs,Sat	8.00 to 8.30 PM	7 Hours	3.00 AM to 3.30 AM	3	Q5
1945	Health clinic visit EMA	Mon, Thurs	6.30 PM to 7.30PM	7 Hours	1.00 AM to 1.30 AM	7	Q6,Q7,Q8
1946	New medication EMA	Tues,Fri	6.00 PM to 6.30 PM	7 Hours	1.00 AM to 1.30 AM	7	Q.9
2128	Job/School EMA	Tues,Fri	6.00 PM to 6.30 PM	7 Hours	1.00 AM to 1.30 AM	7	Q.10
1947	Social or financial resources EMA	Wed, Sat	5.00 PM to 5.30 PM	7 Hours	5.00 PM to 5.30 PM	3	Q.11, 12
1948	Childcare EMA	Mon, Wed, Fri	3.00 PM	7 Hours	10:00 PM	1	Q.13
1949	Discrimination Feeling EMA	Tue, Fri and Sun	8.00 PM	7 Hours	3.00 AM	3	Q.14
1950	Unsafe feeling EMA	Tue, Thurs, Sun	7.00 PM	7 Hours	2.00 AM	2	Q.15
User Trigger / Button trigger Type							
1940	Report food intake	User can trigger this survey any day	User can trigger this survey any time	30 minutes	30 minute after triggering	1	Picture Question
1941	Report Place of Sleep	User can trigger this survey any day	User can trigger this survey any time	30 minutes	31 minute after triggering	1	Picture Question
Entry survey Type (One time)							
1964	Test Survey For Researchers for the demonstration Purpose only	Scheduled to Trigger immediately after joining study for 1 time	Immediately after completing the consent form.	No expiry set	No expiry set	8	6 different type of sample questions for demonstration

Figure 4.1: Table describing the triggering logic and other details of surveys set up in the study

As shown in Figure 4.1 above, there were two primary types of surveys set up in Ethica for this study. The first type served as ecological momentary assessments (EMAs); these were set up to be administered at specific intervals (scheduled trigger logic). The second type was "self-triggered surveys", which could be triggered by the participant through a button click in the Ethica smartphone application on their phone. The EMAs involved 2 sets of surveys – "daily surveys" and "weekly surveys". These surveys were scheduled to repeatedly trigger at specific times until the end date of the study. There were in total seven surveys programmed to trigger daily, and eleven surveys programmed to trigger intermittently on two or three days of each week, until completion of a participant's time in the study. The survey trigger time and expiry time for each of these surveys were different and they were set up during the study design phase of study through the Ethica website. The trigger time and expiry time for each survey are mentioned in respective columns

"Trigger time" and "Expiry time" of Figure 4.1 above.

While frequently triggered surveys can provide insight into participant patterns, for some aspects of participant behaviour and context – such as indicating food intake, the purchasing of food, place of sleep, etc. – it was more convenient if the participants could elect to report it at a time of their choosing. The study, therefore, employed the second type of survey triggering logic in the form of "self-triggered surveys" or "participant-triggered surveys". Such surveys can be triggered and answered by the participant anytime using a press of a button in the Ethica interface for the study. The capacity to trigger such surveys aided the participant in reporting responses at any time, by clicking a button in the interface and answering the resulting questions. There are in total two self triggered surveys set up in this study – 1) Report food intake and 2) Report place of sleep. As shown in Figure 4.1, with the exception of one survey, all of the other seventeen surveys utilizing scheduled trigger type logic have a fixed expiry time of seven hours; by contrast, in reflection of the more temporally immediate nature of the information being collected, self triggered surveys have an expiry time of 30 minutes. If not submitted after the specified expiry duration, surveys will expire.

There are in total six mobile sensors used in the study, which were focused on location data, motion data and battery level of the phones. The location sensors used were GPS and Wi-Fi; motion sensors were accelerometer, linear acceleration and orientation. Finally – and particularly to provide an understanding as to the degree to which lack of access to reliable charging options might serve as a barrier to study adherence – the phone battery level and charging status were captured using the battery sensor.

GPS records the location of the device roughly every 5 minutes; however, participants can always exercise the option of disabling the GPS data collection, or can "snooze" it for 1 hour, so as to avoid data collection during that time period (see Section 4.3.4). Wi-Fi was another location based sensor, and records the Wi-Fi signals in the surrounding environment of smartphones; this sensor can be used in the future to aid in inference of whether a participant is indoor or outdoors at a given time, and can also give a sense as to the degree to which the participants enjoy access to the internet.

4.2.2 Data collection

De-identified data were collected using Ethica. The patients were recruited by research assistants at the Infectious Diseases Clinic at the Regina General Hospital, Regina, SK, with enrollment taking place following briefing on the study, and only if and when consent was offered by a candidate. Recruitment emphasized diversity in risk factors, so as to understand the variety of feasibility barriers associated with populations bearing such distinct risk factors. Specifically, recruitment efforts were made to enroll those associated with five distinct types of risk factors: intravenous drug users (IVDU), men who have sex with men (MSM), people from countries where HIV is endemic, heterosexual, and those who were believed to have experienced a transfusion of blood carrying HIV+. Fifteen individuals were recruited, with representation of each of the five above risk factors.

The restrictions stipulated by the Human Ethics Review Board proscribed analysis of the survey contents and GPS or sensor data. Hence, only counts of answered surveys are analyzed here, and no other details concerning the data were analyzed.

Participant demographics and Participation Outcomes As per the demographics of participants – such as their risk factors, study duration, phone status, overall count of surveys answered, etc., out of 15 total participants recruited for the study, there were 7 intravenous drug users (IVDU), 3 who reported being MSM, 2 from endemic countries, 2 with Heterosexual HIV partners and 1 infected through transfusion of blood carrying HIV+. It was believed by the study team that all of the MSM would have been infected with HIV from their partner, and each individual within the study belonged to a single one of these categories. For confidentiality reasons, no further details on participant demographics are provided here.

Ten phones were distributed to participants in need of phones; out of those, 4 phones were lost by study participants during their study period, and 1 person lost the sim card that allows the phone to connect to call voice and data networks, rather than the phone itself; loss of that sim card disqualified that participant from continuing in the study. For 7 participants, data were collected for the entire study duration of approximately 6 months. The remaining 4 participants had data provided over a calendar time span of between 126 and 170 days. The adherence patterns of these participants in terms of survey answering behaviour are presented in the results discussed in Section 4.3.

4.2.3 Filtering

Data filtering is an important step for many studies associated with data. Prior to the study launch, some members of the research team had used Ethica to join the Ethica study to test the interface and survey schedule patterns. Data from these test pseudo-participants was consequently stored in the Ethica databases. Hence, as a pre-processing step, data not originating from genuine study participants were filtered out based on user_id and study start and end dates. Ethica has a unique ID (termed the user_id) associated with every study participant 1.2.2, and is an anonymous identifier not associated with any identifying information regarding participants. However, as suggested by clinical team, as an additional layer of security, a different unique label lying in the range of 1 to 15 was used as the user_id in participant-specific results shared with the research team. However, for privacy reasons, the text of this chapter omits mention of participant identifiers. This chapter was focused only on deriving feasibility metrics for the study, and hence only quantitative analysis was performed to understand the quantity and availability of data. No other filters related to response rate or data quality were applied.

4.2.4 Operationalization

To understand the quality of data extracted from the Cassandra database, an aggregation was performed. The nominal study duration of all participants is fixed as six months; however, their study start dates varied. Moreover, some participants either left the study early or lost the phone or sim card during the study (see

Section 4.2.2), terminating their study involvement. As a result, both the study start dates and the time at which effective study involvement terminated differed across participants. The study duration was calculated based on start date reported by the clinical team and the time participant ended the Ethica study. The end date was determined based on the last time at which Ethica recorded a battery data point from that participant.

With participants being associated with different study duration, start dates and end dates, analysis to understand all participants' independent patterns in a single time series plot was challenging, and an aggregate level understanding was sought to better understand adherence patterns. To reveal these patterns of study adherence in detail across participants in terms of their time of the study – such as the 1^{st} day in the study, 2^{nd} day, etc. – additional labels were added to all records collected from participants. Data processing pipelines using Scala and Apache Spark described in Section 1.2.3 were implemented to undertake all the data processing steps in an effective and efficient way. The record_time column associated with each record represents the timestamp at which particular data point was recorded, and the record_time at which the first Ethica data point was recorded was used to compute and label as on which day in the participant's time in the study a particular data row was collected. This study day label ranges from day 1 to day 183, based on each participant's actual study start and end dates. In reflection of concerns that study adherence may flag with a participant's cumulative time in the study – due to loss of novelty, competing priorities, need to free up space on their phone, or survey fatigue – this label helps in aggregating the responses per-days-in-study for all users and aids understanding of the adherence patterns across the amount of time that participants have spent in the study. The questions investigated concerning such adherence related to fraction of participants answering surveys on successive days of their time in the study, and the fraction of surveys answered per day-in-study were addressed by aggregating across this days-in-study label and participant id. More specifically, as reported in figures in Section 4.3 below using the days in study label, two metrics were applied:

- Fraction of total surveys answered by all participants out of the total surveys they all received on each successive day of their time in the study (for a total of 6 months).
- Fraction of participants answering a minimum of 1 survey on a given day of their participation in the study.

The next step of the analysis was focused on assessing which users exhibited good adherence in terms of survey responses. The previous step created labels associated with study days aided understanding of the temporal adherence patterns across the participants population. But an understanding at an individual level for each of the participants was also important, particularly given the diversity of risk factors associated with participants mentioned in Section 4.2.2. Within this section, I aggregated based on participant id and (where required) the study day label, in order to explore at a high level whether any particular risk factor is associated with lower adherence. For each participant below, the following measures were calculated after aggregation; while graphs were produced for the research team for each measure below, such graphs were omitted from this thesis due to privacy concerns. But some observations are noted below for each of these measures.

- Total count of surveys answered per participant, as a whole and stratified to distinguish between scheduled trigger (EMA) type and self (button) triggered survey type
- The fraction of total surveys answered throughout entire study period
- The fraction of days in the study during which a given participant answered a minimum of one survey according to each of two metrics: Based on the expected study duration of 183 days and (separately) based on the actual study duration calculated using that participant's start and end dates.
- The average count of surveys answered daily by each participant, also according to each of two particular metrics: Average count of surveys answered within 24 hours of receiving notification that a response is required, and (separately) average count of surveys answered without the condition with respect to response time.
- The fraction of days on which GPS was recorded by the participant, also according to two specific metrics: Based on the expected study duration of 183 days and (separately) based on the actual study duration calculated using that participant's start and end dates

In the case of GPS data, only the count of GPS traces recorded on each day of the study ranging from day 1 to the study end day number for each participant was analyzed. The count of GPS traces per individual varies each day, and hence, for each participant, a threshold of 50 data points per day was set to label whether GPS data was well represented in a study day or not for that participant. If the count of GPS records on a given study day was below 50 for a participant, that day was marked as day with sparse GPS data. After an operationalization step, the resulting files were extracted, and results are plotted as below.

4.3 Results & Discussion

Study adherence – as operationalized in terms of survey answering behaviour and GPS data quantity – is explained in this section. Adherence analysis was performed as per the aggregation steps mentioned in the operationalization Section 4.2.4, on the basis of days in the study and participant id.

4.3.1 Study feasibility criteria

A pre-specified feasibility criterion was stipulated during the study planning process prior to study launch. To prove study feasibility, it was necessary to determine if the criteria proposed at the beginning were met or not. There were in total five feasibility criteria established:

- 1. Recruitment requires no more than 5 researcher hours of recruitment time per recruited participant.

- 2. At least 50% of participants provided with phones possessed those same phones in working condition at the completion of the study.

- 3. A study withdrawal rate of below 50%.

- 4. For at least 50% of participants, location data being recorded on average at least 8% of the time.

- 5. At least 50% of participants completed at least 20% of EMAs within 1 day of issuance.

All of these 5 criteria had to be met to demonstrate study feasibility. The first criterion was met during the recruitment phase itself, at which no more than 5 researcher hours were spent to recruit each participant. The other 4 criteria were analyzed and are discussed in greater detail in the sections below.

4.3.2 Per participant survey data adherence

Survey types

As noted in Figure 4.1, the surveys were classified into 2 types a) EMAs triggered on a Time-based schedule and b) Self-triggered surveys. Analysis results for such survey are included below. While analysis took place using de-identified data using participant numeric identifiers, for privacy reasons, all particular participant identifiers are omitted for the below.

From the analysis, it is noted that the participants demonstrated a limited tendency to self trigger and answer the self triggered EMAs, except for two participants, who answered 90 and 134 surveys out of the total 131 and 147 self triggered by them, respectively. Except 3 participants all other 12 participants had self triggered surveys for only ≤ 25 times during their entire study duration. But regardless of that, all participants answered more than 50% of the self triggered surveys except 2 participants with no self triggered surveys answered. It was noticed that 4 participants lost phones provided by the study team and that a single participant had lost their sim card. Out of these 5 participants, 4 provided fewer responses to survey and demonstrated low adherence as they lost phone during their initial days in the study. By contrast, one user demonstrated good adherence until the day the phone was lost. This user had answered 90 self triggered surveys – the 2^{nd} highest count of self triggered surveys answered amongst participants. That participant was further notable for having answered 701 surveys out of 1391 received – a response rate of 0.50. It was further noted that 7 out of 15 participants had completed the entire expected study duration of 183 days. Four others remained in the study for more than a 4 month time period and contributed a correspondingly high quantity of survey data. Overall, 11 participants exhibited higher study adherence out of 15 recruited by participating in the study for more than 68% of total planned study duration (6 months). Findings reported in the results section below omit mentions of participant demographics and participant numeric identifiers due to privacy concerns. A detailed characterization of the survey response rate is discussed in the paragraph below.

Survey response rate

The survey response rate was calculated by dividing the total number of surveys answered by each participant by their total count of surveys received or self triggered. The value for the denominator was the total count of all surveys which were issued to participants and have the status of answered, expired, or cancelled. As per feasibility criteria 5, a 24 hour filter was applied before conducting the analysis and filtered out all responses received more than 24 hours following the survey issuance. But this filter did not make much impact on the results, because the surveys set up in Ethica are of 7 hours expiry time (as configured through the Ethica interface), and hence surveys expired after 7 hours of non-response. i.e., in detail, out of 21306 total surveys received by all participants, 14817 surveys were answered and out of that only 15 surveys (0.1% of total answered) from 5 participants remained active for more than 24 hours, which is a minimal fraction(0.07%) of total surveys issued, and one that exerted a minimal impact on the fraction calculated.

The bar chart created using the survey response rate per participant is excluded due to privacy concerns. But the main observations from the survey response rate analysis is added below, which shows the level of adherence shown by participants while in the study. This criterion is notable because different participants took part in the study for different periods of time. To draw two extreme examples, two participants had only a study duration of 3 days and 1 day; while these two participants have answered 25% and 49% of the received surveys, respectively, this needs to be considered in light of the fact that they have only participated for a brief duration.

- As per the analysis, feasibility criteria 5 – which stipulated that in a successful study, at least 50% of participants should complete at least 20% of EMAs within the day of issuance – was met. Specifically, it was found that 12 out of 15 (80%) of participants completed at least 20% of EMAs within 24 hours of receiving it.

- Also, it was noticed that 11 participants (73.33%) demonstrated significantly greater adherence yet by completing $\geq$ 40% of surveys.

The fraction of days on which at least 1 survey was answered

The results from this subsection enumerate for all participants the fraction of days on which at least 1 survey was answered – a per participant variant of the overall survey response rate calculated in the previous subsection of Section 4.3.2. The motivation behind this criterion involving answering at least a single survey was that it gives an indication of whether a person has opened or interacted with application on study days at all, and thereby some indication as to how actively the person participated in the study. Two major metrics are calculated and explained in points below. But the bar chart created on a per-participant basis using the fraction of days at which a minimum of one survey was answered are excluded from inclusion in the thesis document due to privacy concerns. Firstly, the metric A (hereafter denoted FDS metric A) represents over what fraction of the days in the total expected study duration a given participant answered a minimum of

1 survey. In this case, the fraction was calculated by fixing the denominator as the total expected study duration of 183 days, which was the maximum duration for which a participant could remain in the study. By contrast, the second metric – termed FDS metric B – represents, for a given participant, what fraction of that participant's actual study duration they answered a minimum of 1 survey. For this metric, the actual study duration for a participant reflects their join and end dates (as mentioned in Section 4.2.4). In short, for FDS metric B, we are calculating how adherent each participant was during their actual study duration, whereas in FDS metric A, we calculated how adherent participants when take in light of their planned study duration. These 2 metrics are valuable for understanding the adherence rate for each participant at an overall level. Below are the main observations:

- Out of the 3 participants with a lower overall survey response rate of <20% and who therefore failed to meet the 5^{th} feasibility criteria at an individual, one participant exhibited an overall survey response rate of just 0.16, but maintained a fraction of 0.66 for the FDS metric A calculated using 183 days, and a fraction of 0.71 for FDS metric B calculated using the original study duration. This highlights the fact that even though that person did not satisfy the per-person analogue to the 5^{th} criteria on account of answering only 16% of the overall survey received, that participant may be associated with high participation when judged in terms of answering at least 1 survey per day –in this case, the participant answered at least 1 surveys on 71% of their actual days in the study, and on 66% of the total study expected days.

- When applying the FDS metric A, it was found that 73.33% of total participants (11 out of 15) answered at least a single survey for more than half of the days of their expected study duration (183 days).

- Out of the 7 participants who remained in the study for the entire 183 days of participation – 6 participants answered at least 1 survey on 90% of their study duration (in accordance with FDS metric B).

- The adherence of 4 participants exhibited a marked difference as judged with metrics A and B, because they lost the phone in the initial days of the study, and have data only up to a maximum of ≤ 13 days. Other than these 4 participants, all the other 11 participants have answered at least a single survey on more than 50% of the 183 expected days.

Average daily count of surveys answered

The main observations from the calculated average count of surveys that each user answered on their study days are added as points below. But the resultant bar chart was excluded from the results section due to privacy concerns. Two variants of the average counts were calculated. For ADC metric 1, they are calculated after filtering out all surveys answered more than a day after issuance. Secondly, for ADC metric 2, the average was calculated based on surveys regardless of the delay with which they were answered. In both cases, the average was calculated by dividing total surveys answered by the total days in the study for each participant. Only slight variations are noticed between the two metrics. Below are the main observations:

- ADC metric 1: The average daily count of surveys answered across all users (after filtering out surveys answered after 24 hours of receiving it) was 7.80 surveys per day.

- ADC metric 2: Average daily count of surveys answered without applying the timeliness filter was 7.81 surveys per day. Only a slight difference was noticed, because, across all users, only 15 surveys were answered over 24 hours late.

- In the case of 11 users demonstrating higher adherence, the average count was 8.42 (ADC metric 1) and 8.41 (ADC metric 2), respectively.

4.3.3 Survey data adherence: Per day in study

Survey adherence per day in study

Figure 4.2 shows the overall fraction of surveys responded to by participants by their day in the study, as computed over all participants. The day in the study (calculated after aggregating each user's responses based on their record_time as described in the operationalization Section 4.2.4) was plotted on the x-axis, and the aggregated survey response rate was plotted on the y-axis. The survey response rate in Figure 4.2 was calculated in fashion similar to that used for the previous subsection: *Survey response rate* of Section 4.3.2. But the previous subsection considered the per-participant fraction aggregated over all days that a particular participant had been in the study. By contrast, Figure 4.2 considers instead the fraction of surveys answered per day in study, aggregating overall participants. Hence, in this case, for each study day (measured relative to the study start date for a participant), the total count of surveys answered by any participant on that day of their study participation are divided by the total surveys that participants received on that study day of their participation (as determined by the sum of answered, expired and cancelled surveys on that day).

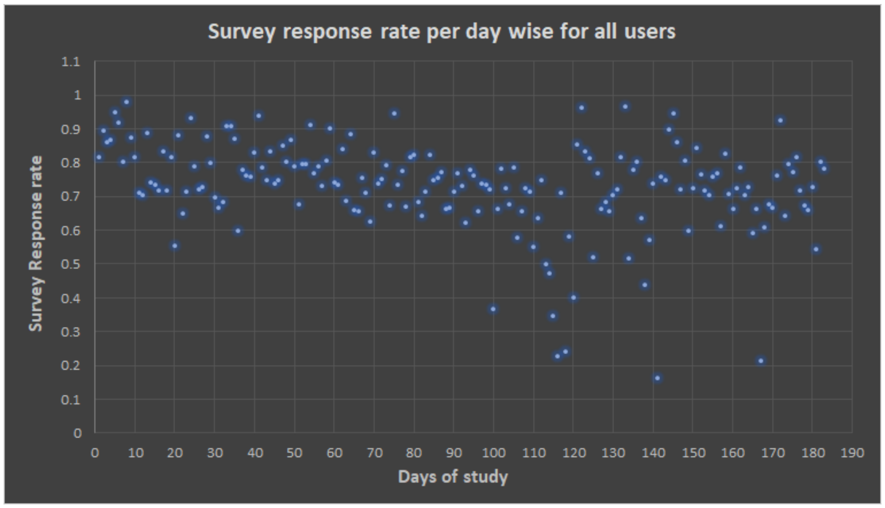


Figure 4.2: Fraction of surveys responded to (y-axis) per day in the study (x-axis), aggregated across all participants

Hence, in Figure 4.2, the change in pattern of survey response over time-in-study was displayed. Below are the main observations:

- Survey response rates from all participants within the 1^{st} two months of participation (until day 61) were higher than response rates for the following 4 months. Over the first 2 months, the response rate lay between 0.7 and 1.0, with a few outliers noticed on fewer than 10 days in the study.
- On the 3^{rd} month – starting from day 60 until day 110 –when considering data aggregated from all participants, a decreasing trend in survey response rate was noticed.
- After 110 days, the response rate dropped and showed markedly lower – and decreasing – trend until day 120.
- From the beginning of the 5^{th} month on day 121, the response rate resumed a higher level, and remained in a range greater than 0.5, except on 2 or 3 days. This pattern remained until the end of the 5^{th} month, on the 150^{th} day.
- From the beginning of the 6^{th} month – starting from day 151 – until the last (183^{rd}) day, the response rate remained in a range between 0.6 and 1.0 on all but 3 days.
- To summarize, a decreasing response rate trend was noticed in the 3^{rd} and 4^{th} months. Then, from the beginning of the 5^{th} month, it showed an upward trend and remained in a range above 0.6 for the remainder of the study period. It could be that a follow up by the research team played a role in shaping this pattern, as participants had visits scheduled with the clinic during the study period.

To conclude, the overall response rate – aggregated over all participants – remained at a value greater than 0.5 on all but 10 days – an exception constituting less than 6% of the entire study duration.

The fraction of Participants answering at least one survey

Similar to Figure 4.2 above that describes the change over time in the fraction of overall survey responses across all participants, this section analyzes the evolution of the fraction of participants answering a minimum of one survey on each successive day of participants' involvement in the study. Figure 4.3 characterizes this evolution in graphical form.

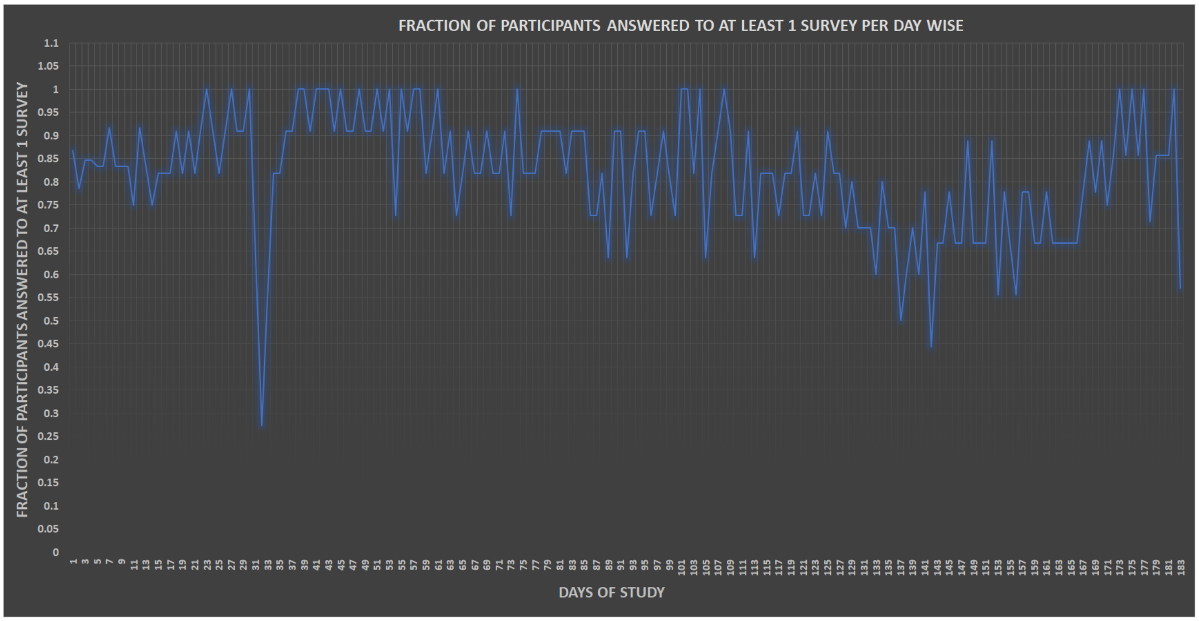


Figure 4.3: Fraction of participants answering at least one survey per day, by day in study

Below are the main observations with regards to Figure 4.3:

- For the 1^{st} month, the fraction of participants answering at least one survey per day remained in a range of ≥ 0.75.
- A sudden drop to around 0.27 was noticed at day 32 and 33; after a cross check with the Ethica survey logs, it was understood that a technical glitch occurred on that day. Because of that glitch, out of 11 participants remaining in the study on day 31, only 3 participants received surveys; that is, 8 out of 11 participants did not receive any surveys through Ethica. This glitch precipitated a sudden drop on day 31 of the fraction of participants answering a minimum of 1 survey. By contrast – as shown in Figure 4.2 – the survey response rate was not much affected and remained at 0.68. This reflects the fact that for this alternative metric – as shown in Figure 4.2 – we calculated the fraction of surveys responded to out of those *received by* participants. Out of the total surveys received by those 3 participants who received any surveys, 68% were answered. A similar condition applied on day 33: 4 participants out of the 11 who remained in the study did not receive any surveys; by contrast, 7 participants received surveys; as per Figure 4.2, the response rate 0.90 was amongst those 7 participants receiving surveys on day 33.

- After day 33 and for the balance of the 2^{nd} month until day 60, the fraction of participants answering at least a single survey on each successive day of participation remained above 0.8, with only one day showing a drop to 0.72.

- From the start of 3^{rd} month (day 61) until towards the end of 3^{rd} month (day 90), the response remained in the range of 0.7 to 1, except for a slight dip to 0.69 on day 89.

- Subsequent months exhibit more variable – and generally lower – adherence, as judged by this measure.

- Overall, the graph exhibits moderate to high level of adherence patterns for all participants, ranging from 0.5 to 1 on all days except 2 days – of which one day (31) exhibited a technical glitch. Hence, with the exception of a single day, on all days exhibiting normal system operation exhibited greater than or equal to 50% of the participants remaining in the study at that time answering at least one survey. Also, on 152 days out of 183 (83.52% of the total study days), greater than 70% of participants answered at least one survey per successive day of participation in the study.

In one of the quantitative analysis approaches implemented in this chapter, a qualitative understanding related to the quantitative data collection provided an explanation concerning an observed anomaly in the data. Specifically, in the observations of Figure 4.3, an anomaly was noted on day 32 and day 33 in terms of a precipitous drop in the fraction on the number of participants answered on those days. Discussion revealed that this was due to a technical glitch with the Ethica system; this underlying cause was known by the research team only because it was a technical failure reported by the Ethica team and the patients on those particular days. It bears noting that this study lacked any formal incorporation of qualitative data collection, such as those involving patient narratives or perceptions; the author believes it likely that such types of data collection could have served to deepen an understanding of many of the patterns noted with respect to adherence.

4.3.4 Sensor data adherence: Per participant

The fraction of days that GPS was recorded for each participant

The 4^{th} feasibility criteria – which stipulated that for at least 50% of participants, location data should be recorded on average at least 8% of the time – was evaluated in this section, both in aggregate and in terms of its per-participant components. By default, GPS data was recorded every 5 minutes for all consenting participants. But if a participant did not want to share their location, Android and iOS smartphones provide the user an option to disable phone location services for the device as a whole. Such permissions can also be declined specifically with respect to Ethica. Participants can further elect to grant permission only when the Ethica app is open. In addition, with the press of a button in the Ethica app interface, participants can also request to “snooze” Ethica’s sensor data collection process for 1 hour. Finally, the permission granted at the beginning of the study can be revoked anytime by the participant during their study period by terminating the application or by turning off their device's GPS access. In all such cases, Ethica will not collect GPS data and will inform the participant via a notification. The research team will be informed of the absence of such GPS data in the web-based Ethica dashboard 1.2.2.

This research calculated – for each participant – the fraction of days on which at least 50 records of GPS data were collected for that participant. The fraction was calculated for each participant according to two different metrics. FDG metric 1: This was determined for a participant by dividing the total count of days on which that quantity of GPS information was recorded by that participant's actual study participation duration. (It bears recollection that the study duration was calculated as mentioned above in operationalization Section 4.2.4, by calculating the difference between start date and the date at which the last record for that participant was collected via Ethica from the phone). On the other hand, FDG metric 2 was an alternative metric that considered not their actual but planned duration, by dividing the total count of such GPS-collection day by the expected (fixed) participant duration of 183 days. While produced as part of the research, the resulting chart is omitted from this thesis for privacy reasons.

Below are the main observations notable from GPS sensor data analyzed in this section:

- FDG metric 1 demonstrated that 14 out of 15 participants met the criteria of having location data recorded on average at least 8% of the study time. This criterion was focused only on the days participants are engaged in the study, rather than the planned duration of 183 days, hence, 4 participants who left the study early, but had recorded GPS data for $\geq 0.8\%$ of time, were included amongst those 14 judged to have exhibited acceptable levels of location data.

- Results for FDG metric 2 shows that if we consider the planned study duration of 183 days in the denominator, then 73.33% of participants (11 out of 15) have provided GPS data for at least 8% of the planned study duration – a threshold which still satisfies the feasibility criteria.

- Also, analysis with FDG metric 1 demonstrated that 86.86% of total participants (13 out of 15) had GPS data recorded for more than 40% of their participation duration. By contrast, FDG metric 2 demonstrated that 66.67% of total participants (10 out of 15) recorded GPS data for more than 40% of their planned study duration. In the case of FDG metric 2, three participants with few days of participation fail to meet that the individual-level variant of the feasibility criterion on account of that short participation, even though they demonstrated greater adherence in terms of providing location data during the periods of study participation.

- As quantified by FDG metric 1, 60% of total participants (9 out of 15) maintained exceptional adherence, with GPS recorded for $\geq 90\%$ of their participated duration. By contrast, as measured by FDG metric 2, 40% of total participants (6 out of 15) had GPS recorded for $\geq 90\%$ of their planned study duration – a qualification that constitutes exceptional adherence in terms of remaining in the study for a longer period and providing GPS data for more than 90% of those days.

To conclude, a high level of adherence is noted in terms of GPS data from all participants, irrespective of short or long study participation periods. This suggests that – subject to approval from the appropriate ethics review boards – high-resolution geographic data on participant location and mobility patterns would likely be available for scientific enquiry into barriers to HIV care seeking and medication adherence in a future larger-scale study. The results further demonstrate a successful passing of all feasibility criteria with respect to GPS data availability.

4.4 Conclusion

All of the above results discussed in this Section 4.3 suggest that the study was successful in terms of collecting data with the support of participants and researchers and that several fundamental aspects of data quality are sufficiently favourable to motivate a larger future study. More specifically, the results demonstrate that all 5 feasibility criteria mentioned in Section 4.3.1 are met.

The 1^{st} criteria – involving recruitment time – was met upon recruitment, as no more than 5 researcher hours were required per participant.

The 2^{nd} criteria concerned study phones. As discussed in the study demographics Section 4.2.2, 10 persons were provided with a smartphone and data plan. Out of such participants, 3 lost their phones within days of beginning the study, and 1 lost their phone after 126 days. Altogether, 4 smartphones were lost, and 1 sim card was lost. As a result, 6 still possessed study-provided phones in working order at study completion, fulfilling the 2^{nd} criteria.

The 3^{rd} criteria concerned whether the study withdrawal rate was below 50%. As demonstrated in the subsection of *Participant demographics and Participation Outcomes* (see Section 4.2.2), 7 participants (46.6% of the total) completed the entire study duration of 183 days (6 months), and thus successfully completed the study. So a strict interpretation of the withdrawal rate criteria was almost met. But an important point to note concerns two participants who nominally withdraw from the study, but have study duration of 170 and 169 days, respectively, having left the study only 12 and 13 days before their study end date: Were either of these participants to be considered as satisfying the criterion independently or in combination, the withdrawal rate criterion would be met.

The 4^{th} criteria concerned the provision of location data. The quantity of GPS data collected over time was analyzed for cross checking this criteria. As judged by their per-participant study duration, it was found that for 93.33% of participants (14 of the 15 participants) location data was being recorded on average at least 8% of their study participation duration, and hence such criteria were met. This was calculated for a given participant by keeping denominator as the count of days of actual study participation by that participant and using as the numerator for how many days that participant had at least a minimum threshold of location data (>50 data points). But if we consider the total study duration of 183 days in the denominator, then 73.33% of participants (11 out of 15) have provided GPS data for at least 8% of the expected study duration.

The criterion is thus met for either interpretation of "study participation duration".

The 5^{th} criteria were with regards to the EMA completion rate. The criteria stipulated that at least 50% of participants completed at least 20% of EMAs within the day of issuance. It was found that 12 out of 15 contributing 80% of participants completed at least 20% of EMAs within 1 day of issuance, successfully meeting – and substantially exceeding – that criterion.

The major vision of this chapter (Chapter 4) is to assess the feasibility for a larger study seeking to understand the barriers that prevent patients from maintaining proper adherence to medications and to support patients in following a healthy lifestyle. The current results suggest that scaling up this HIV study may yield sufficient involvement by participants to perform more sophisticated analytics. Strong involvement by patients and researchers may reveal insights into factors shaping medication adherence that could help to promote and elevate such adherence. It is hoped that such changes inspired by the findings of a larger study could better support HIV patients in adhering to their treatment regimes, and to thereby control the spread of HIV in the province. The evidence base gathered through a larger study could also help to link treatment results to mental health outcomes, and thus enhance the capacity for researchers to characterize patterns at both the individual- and population-level. While the work performed for this chapter was limited to qualitative analysis, both the process of conducting such work and the findings suggest that incorporation of a qualitative element into the larger study may offer additional insights, both in terms of the substantive findings concerning factors shaping medication adherence, and in terms of patient involvement within the study.

Chapter 5

Conclusion

The count of mobile phone users and the number of smartphone applications are increasing day by day. This inflow of new applications expands the possibilities of using smartphones in areas other than mobile phones' original purpose of ensuring uninterrupted communication opportunities – areas such as entertainment, social networking, e-learning, navigation, business, e-commerce, health and fitness, etc. The availability of such a broad set of applications offers convenience for smartphone users by allowing them to depend only on a single device instead of multiple such devices for diverse day to day needs. In order to improve the capacity of smartphones to support these additional functions, it has become standard to incorporate additional sensors into smartphones – sensors, such as an accelerometer, GPS, Wi-Fi, phone proximity, gyroscope, screen state sensor, etc. [75]. Such sensors support a growing number and diversity of applications, which attract users, and have resulted in a situation in which smartphones are no longer a luxury, but are instead increasingly important for modern lifestyle and needs. This has further resulted in an increasing trend of smartphone usage among all age groups – especially among teenagers – and has increasingly influenced the lifestyle and behavioural patterns of the population, with widespread use even being seen within low socioeconomic status populations.

This increase in smartphone usage also opens up possibilities for using smartphones as data collection tools – tools that not only serve as powerful vehicles for eliciting self reporting, but which further employ device sensors for data collection purposes. Beyond supporting application functionality in day-to-day use, such sensors can serve as a powerful tool for research studies involving consenting participants. Among participants in such studies, smartphones can track activity patterns, behavioural patterns, exposures, etc. using location data, screen usage data, physical activity data, data on contact patterns, etc. collected by the large battery of on-device sensors. A central advantage of using smartphones as behavioural research tools lies in such phones' capability to record near-continuous data from study participants, and the consequent ability to monitor their behavioural patterns in an ecological context – involving both physical context and electronic context – rather than in a laboratory or confined area. The high volume, velocity and variety of the sensor data involved are additional advantages that strengthen the accuracy of results and aid researchers in applying several data models and arriving at conclusions. All of such features has supported the increasing popularity of smartphones as data collection tools, especially in the field of health studies.

In Chapter 2, the machine learning approach of unsupervised Hidden Markov Modeling was implemented to label the underlying hidden states associated with the screen state sensor data generation process; testing with synthetic ground truth data suggests that the approach is likely quite accurate. This chapter provides an example of the need for – and use of – machine learning techniques for basic understanding of big data in the field of health research and for studying the behavioural patterns of the population, in handling or overcoming the issues associated with the level and quality of collected raw datasets or big data collected by smartphones. The case studies further demonstrate the power of such advanced methodologies to reveal meaningful human health exposures – here, screen time – from raw datasets.

In Chapter 3, quantitative, exploratory and statistical analysis was performed on data collected from smartphones to understand mental health patterns – including variability in reported feelings – and the co-occurrence of 3 risk factors on suicidal ideation. This study applies several data analytic methods that can be applied for similar studies to understand participant adherence patterns, data distribution, relationship between variables, prediction of response variables, etc. It also covers several visualization techniques to help understand the adherence patterns of participants over their time in the study. While the thesis chapter focused on results aggregated over participants, it bears emphasis that the research underlying it secured considerable insight into the patterns through per-participant figures that were omitted from the thesis document for privacy reasons. This chapter further covers the widespread step of pre-processing involved in data analysis projects, so as to improve the quality of data in terms of filtering steps and operationalization. Such adherence and quality improvement steps could further be adapted for future studies handling similar datasets.

In Chapter 4, a quantitative study of survey and GPS sensor data was performed evaluate whether this study that sought to capture individual-level behaviour in HIV participants achieved criteria demonstrating the feasibility of the approach. A further exploratory analysis was not performed in this project due to restrictions stipulated by the Regina Qu'Appelle Health Region Research Ethics Board. Prior to study launch, the study team established 5 criteria that a successful study must meet to be considered as demonstrating feasibility of the approach. The study analysis clearly indicates that the study met all these criteria, and in most cases greatly exceeded them. This study was notable success when judged in terms of adherence, data quality and data quantity. This suggests that the approach will be suitable for exploration at a larger scale in future studies investigating aspects of participant mental health, activity patterns, sedentary behaviour, adherence to medications, and barriers preventing participants from following treatment regimes, etc.

To summarize, this thesis covers several data analytics and machine learning approaches that can be used to handle big data collected by smartphones to understand participant behaviour, exposures and adherence patterns and to address major issues affecting data quality. Three different datasets involving distinct populations collected from three corresponding case studies were discussed here to explain the challenges faced by the health researchers and the ways to tackle these challenges. The relatively small sample sizes within each of these studies – and particularly those involving suicidal ideation and people living with HIV – limit the generalizability of the findings advanced here to larger populations. But overall, the approaches implemented in this thesis were successful and helped the three collaborating health research teams to secure valuable insights from the big data collected using smartphones. The outcomes of this approach will be used by some of the health researchers in the future projects, to compare with various other metrics associated with same participants calculated from the study or collected directly from the study. Also most of the methods implemented in the thesis can be used in further projects using the same mobile data collection platform (Ethica), and hence are reusable.

References

[1] F. Alexander. *Data is everywhere and that's a good thing.* IBM Business Analytics Blog, Oct, 2016. Accessed on: Apr. 30, 2019. [Online]. Available: https://www.ibm.com/blogs/business-analytics/data-is-everywhere/.

[2] "*What is Health Data Science? | CBDRH - Centre for Big Data Research in Health*," 2019. Cbdrh.med.unsw.edu.au. Accessed on: Apr. 30, 2019. [Online]. Available: https://cbdrh.med.unsw.edu.au/what-health-data-science.

[3] "*The Four V's of Big Data*," 2019. IBM Big Data & Analytics Hub. Accessed on: Apr. 30, 2019. [Online]. Available: https://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[4] "*Big Data needs Big analytics*," 2012. Software World, Gale General OneFile. Accessed on: Apr. 30, 2019. [Online]. Available: http://link.galegroup.com.cyber.usask.ca/apps/doc/A307920860/ITOF?u=usaskmain&sid=ITOF&xid=a44b98b0.

[5] D. L. Knowles, K. G. Stanley, and N. D. Osgood, "A field-validated architecture for the collection of health-relevant behavioural data," in *2014 IEEE International Conference on Healthcare Informatics.* IEEE, 2014, pp. 79–88.

[6] L. Fragoso, T. Paul, F. Vadan, K. G. Stanley, S. Bell, and N. D. Osgood, "Intrinsic dimensionality of human behavioral activity data," *PLoS ONE*, vol. 14, no. 6, p. e0218966, 2019.

[7] K. Stanley, S. Bell, L. K. Kreuger, P. Bhowmik, N. Shojaati, A. Elliott, and N. D. Osgood, "Opportunistic natural experiments using digital telemetry: a transit disruption case study," *International Journal of Geographical Information Science*, vol. 30, no. 9, pp. 1853–1872, 2016.

[8] Y. Qin, W. Qian, N. Shojaati, and N. D. Osgood, "Identifying smoking from smartphone sensor data and multivariate hidden Markov models," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation.* Springer, 2017, pp. 230–235.

[9] P. Seitzinger, N. Osgood, W. Martin, J. Tataryn, and C. Waldner, "Compliance Rates, Advantages, and Drawbacks of a Smartphone-Based Method of Collecting Food History and Foodborne Illness Data," *Journal of Food Protection*, vol. 82, no. 6, pp. 1061–1070, 2019.

[10] I. Andone, K. Blaszkiewicz, M. Böhmer, and A. Markowetz, "Impact of location-based games on phone usage and movement: A case study on Pokémon GO," in *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services.* ACM, 2017, p. 102.

[11] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, May 2006. [Online]. Available: https://doi.org/10.1007/s00779-005-0046-3

[12] M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, J. S. Brownstein, and A. Vespignani, "Digital Epidemiology," *PLoS Computational Biology*, vol. 8, no. 7, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3406005/pdf/

[13] M. Hashemian, D. Knowles, J. Calver, W. Qian, M. C. Bullock, S. Bell, R. Mandryk, N. D. Osgood, and K. G. Stanley, "iEpi: An end to end solution for collecting, conditioning and utilizing epidemiologically relevant data," *Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 06 2012.

[14] M. Hashemian. *Ethica: Using data to make a better future*. Ethicadata.com. 2019. Accessed on: Apr. 30, 2019. [Online]. Available: https://www.ethicadata.com/about.

[15] N. D. Osgood. *Ethica/iEpi: A Robust & Versatile Smartphone-Based Epidemiological Data Collection System*. usask.ca. 2019. Accessed on: Apr. 30, 2019. [Online]. Available: https://www.cs.usask.ca/~osgood/iEpi/iEpi.html.

[16] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, Aug. 2012. [Online]. Available: http://dx.doi.org/10.14778/2367502.2367572

[17] "*Apache Spark^{TM} - Lightning-fast unified analytics engine*," Spark.apache.org. Accessed on: Apr. 30, 2019. [Online]. Available: https://spark.apache.org/.

[18] D. Singh. *What is the Difference Between Hadoop and Spark*. Datasciencecentral.com, Apr. 2019. Accessed on: Apr. 30, 2019. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/what-is-the-difference-between-hadoop-and-spark.

[19] "*The Scala Programming Language*," scalacenter, Accessed on: May. 3, 2019. [Online]. Available: https://www.scala-lang.org/.

[20] "*Differences Between Python vs Scala*," www.educba.com, Accessed on: May 3, 2019. [Online]. Available: https://www.educba.com/python-vs-scala/.

[21] A. R. Lauricella, E. Wartella, and V. J. Rideout, "Young children′s screen time: The complex role of parent and child factors," *Journal of Applied Developmental Psychology*, vol. 36, pp. 11–17, 2015.

[22] B. Kim, "The popularity of gamification in the mobile and social era," *Library Technology Reports*, vol. 51, no. 2, pp. 5–9, 2015.

[23] G. Lissak, "Adverse physiological and psychological effects of screen time on children and adolescents: Literature review and case study," *Environmental research*, vol. 164, pp. 149–157, 2018.

[24] Y. L. R. Chassiakos, J. Radesky, D. Christakis, M. A. Moreno, C. Cross, C. O. Communications, and Media, "Children and adolescents and digital media," *Pediatrics*, vol. 138, no. 5, p. e20162593, 2016.

[25] M. Hashemian. *Common Data Sources*. Ethicadata.com. 2019. Accessed on: Apr. 9, 2019. [Online]. Available: https://community.ethicadata.com/hc/en-us/articles/236225788-Common-Data-Sources#article-body_toc_2_25.

[26] W. Zucchini, *Hidden Markov models for time series an introduction using R*, ser. Monographs on statistics and applied probability 110. CRC Press, 2009, vol. 150.

[27] G. Pulford, J. C. Gallant, R. Kennedy, and S. H. Chung, "Evaluation and Estimation of Various Markov Models with Applications to Membrane Channel Kinetics," *Biometrical Journal BIOM J*, vol. 37, pp. 39–63, 01 1995.

[28] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[29] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition*, 3rd ed. web.stanford.edu, 2018, ch. A, pp. 465 – 479. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

[30] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-scale Bound-constrained Optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, Dec. 1997. [Online]. Available: http://doi.acm.org/10.1145/279232.279236

[31] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995. [Online]. Available: https://doi.org/10.1137/0916069

[32] X. Li, M. Parizeau, and R. Plamondon, "Training hidden markov models with multiple observations-a combinatorial method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 371–377, 2000.

[33] "*Exporting data to MS Excel workbook*," anylogic.com, Accessed on: Jul. 12, 2019. [Online]. Available: https://help.anylogic.com/index.jsp?topic=%2Fcom.anylogic.help%2Fhtml%2Fconnectivity%2FExport_Excel.html.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] D. L. Olson and D. Delen, *Advanced data mining techniques.* Springer Science & Business Media, 2008.

[36] Y. Sasaki, "The truth of the F-measure," *Teach Tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[37] "*Health at a Glance: Suicide rates: An overview*," Statistics Canada Catalogue no. 82-624-X. Accessed on: Jun. 29, 2018. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/82-624-x/2012001/article/11696-eng.htm.

[38] "*Leading causes of death, total population, by age group*," Statistics Canada. Table 13-10-0394-01. Accessed on: Jun. 29, 2018. [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039401.

[39] E. Robins, G. E. Murphy, R. H. Wilkinson Jr, S. Gassner, and J. Kayes, "Some clinical considerations in the prevention of suicide based on a study of 134 successful suicides," *American Journal of Public Health and the Nations Health*, vol. 49, no. 7, pp. 888–899, 1959.

[40] T. L. Dorpat and H. S. Ripley, "A study of suicide in the Seattle area," *Comprehensive Psychiatry*, 1960.

[41] C. L. Rich, R. C. Fowler, L. A. Fogarty, and D. Young, "San Diego suicide study: III. Relationships between diagnoses and stressors," *Archives of General Psychiatry*, vol. 45, no. 6, pp. 589–592, 1988.

[42] E. Isometsa, M. Henriksson, M. Marttunen, M. Heikkinen, H. Aro, K. Kuoppasalmi, and J. Lonnqvist, "Mental disorders in young and middle aged men who commit suicide," *British Medical Journal*, vol. 310, no. 6991, pp. 1366–1367, 1995.

[43] B. Barraclough, J. Bunch, B. Nelson, and P. Sainsbury, "A hundred cases of suicide: clinical aspects," *The British Journal of Psychiatry*, vol. 125, no. 587, pp. 355–373, 1974.

[44] J. J. Mann, "A current perspective of suicide and attempted suicide," *Annals of Internal Medicine*, vol. 136, no. 4, pp. 302–311, 2002.

[45] A. L. Beautrais, P. R. Joyce, R. T. Mulder, D. M. Fergusson, B. J. Deavoll, and S. K. Nightingale, "Prevalence and comorbidity of mental disorders in persons making serious suicide attempts: a case-control study." *American Journal of Psychiatry*, vol. 153, no. 8, pp. 1009–14, 1996.

[46] E. A. Frazier, R. T. Liu, M. Massing-Schaffer, J. Hunt, J. Wolff, and A. Spirito, "Adolescent but not parent report of irritability is related to suicidal ideation in psychiatrically hospitalized adolescents," *Archives of Suicide Research*, vol. 20, no. 2, pp. 280–289, 2016.

[47] T. Sokero, T. Melartin, H. Rytsala, and U. Leskela, "Suicidal ideation and attempts among psychiatric patients with major depressive disorder," *The Journal of Clinical Psychiatry*, vol. 64, no. 9, p. 1094, 2003. [Online]. Available: http://search.proquest.com/docview/208806140/

[48] K. R. Conner, S. Meldrum, W. F. Wieczorek, P. R. Duberstein, and J. W. Welte, "The association of irritability and impulsivity with suicidal ideation among 15-to 20-year-old males," *Suicide and Life-Threatening Behavior*, vol. 34, no. 4, pp. 363–373, 2004.

[49] S. S. Daniel and D. B. Goldston, "Hopelessness and lack of connectedness to others as risk factors for suicidal behavior across the lifespan: Implications for cognitive-behavioral treatment," *Cognitive and Behavioral Practice*, vol. 19, no. 2, pp. 288–300, 2012.

[50] J. W. Kaminski, R. W. Puddy, D. M. Hall, S. Y. Cashman, A. E. Crosby, and L. A. Ortega, "The relative influence of different domains of social connectedness on self-directed violence in adolescence," *Journal of Youth and Adolescence*, vol. 39, no. 5, pp. 460–473, 2010.

[51] "*GraphPad Statistics Guide : Interpreting results: Kolmogorov-Smirnov test*," graphpad.com. Accessed on: Apr. 20, 2019. [Online]. Available: https://www.graphpad.com/guides/prism/7/statistics/interpreting_results_kolmogorov-smirnov_test.htm?toc=0&printWindow.

[52] "*Ordinary Least Squares*," onclick360.com. 2019. Accessed on: Apr. 4, 2019. [Online]. Available: https://onclick360.com/ordinary-least-squares/.

[53] K. G. Martin. *Assessing the Fit of Regression Models*. theanalysisfactor.com. Accessed on: Apr. 9, 2019. [Online]. Available: https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/.

[54] "*What are HIV and AIDS*," avert.org. 2019. Accessed on: Apr. 24, 2019. [Online]. Available: https://www.avert.org/about-hiv-aids/what-hiv-aids.

[55] "*HIV & AIDS in Saskatchewan (2017)*," publications.saskatchewan.ca. Accessed on: Apr. 24, 2019. [Online]. Available: http://publications.gov.sk.ca/documents/13/108031-2017-HIV-AIDS-in-Saskatchewan-infographic.pdf.

[56] "*HIV in Canada—Surveillance Report, 2017*," www.canada.ca. Accessed on: Apr. 24, 2019. [Online]. Available: https://www.canada.ca/en/public-health/services/reports-publications/canada-communicable-disease-report-ccdr/monthly-issue/2018-44/issue-12-december-6-2018/article-3-hiv-in-canada-2017.html.

[57] "*HIV Prevention and Control Report 2017*," publications.saskatchewan.ca. Accessed on: Apr. 24, 2019. [Online]. Available: http://publications.gov.sk.ca/documents/13/108029-2017-Saskatchewan-HIV-Prevention-and-Control-Report.pdf.

[58] D. Tzemis, J. I. Forrest, C. M. Puskas, W. Zhang, T. R. Orchard, A. K. Palmer, C. W. McInnes, K. A. Fernades, J. S. Montaner, and R. S. Hogg, "Identifying self-perceived HIV-related stigma in a population accessing antiretroviral therapy," *AIDS care*, vol. 25, no. 1, pp. 95–102, 2013.

[59] B. L. Fife and E. R. Wright, "The dimensionality of stigma: A comparison of its impact on the self of persons with HIV/AIDS and cancer," *Journal of Health and Social Behavior*, pp. 50–67, 2000.

[60] T. G. Heckman, E. S. Anderson, K. J. Sikkema, A. Kochman, S. C. Kalichman, and T. Anderson, "Emotional distress in nonmetropolitan persons living with HIV disease enrolled in a telephone-delivered, coping improvement group intervention." *Health Psychology*, vol. 23, no. 1, p. 94, 2004.

[61] E. Kang, B. D. Rapkin, and C. DeAlmeida, "Are psychological consequences of stigma enduring or transitory? A longitudinal study of HIV stigma and distress among Asians and Pacific Islanders living with HIV illness," *AIDS Patient Care & STDs*, vol. 20, no. 10, pp. 712–723, 2006.

[62] J. N. Sayles, G. W. Ryan, J. S. Silver, C. A. Sarkisian, and W. E. Cunningham, "Experiences of social stigma and implications for healthcare among a diverse population of HIV positive adults," *Journal of Urban Health*, vol. 84, no. 6, p. 814, 2007.

[63] K. R. Schafer, H. Albrecht, R. Dillingham, R. S. Hogg, D. Jaworsky, K. Kasper, M. Loutfy, L. J. MacKenzie, K. A. McManus, K. A. K. Oursler, S. D. Rhodes, H. Samji, S. Skinner, C. J. Sun, S. Weissman, and M. E. Ohl, "The continuum of HIV care in rural communities in the United States and Canada: what is known and future research directions," *Journal of Acquired Immune Deficiency Syndromes (1999)*, vol. 75, no. 1, p. 35, 2017.

[64] S. M. Sweeney and P. A. Vanable, "The association of HIV-related stigma to HIV medication adherence: a systematic review and synthesis of the literature," *AIDS and Behavior*, vol. 20, no. 1, pp. 29–50, 2016.

[65] B. Nosyk, J. Min, V. D. Lima, B. Yip, R. S. Hogg, and J. S. Montaner, "HIV−1 disease progression during highly active antiretroviral therapy: an application using population-level data in British Columbia: 1996–2011," *Journal of Acquired Immune Deficiency Syndromes (1999)*, vol. 63, no. 5, p. 653, 2013.

[66] C. C. J. Carpenter, M. A. Fischl, S. M. Hammer, M. S. Hirsch, D. M. Jacobsen, D. A. Katzenstein, J. S. G. Montaner, D. D. Richman, M. S. Saag, R. T. Schooley, M. A. Thompson, S. Vella, P. G. Yeni, and P. A. Volberding, "Antiretroviral Therapy for HIV Infection in 1996: Recommendations of an International Panel," *The Journal of the American Medical Association*, vol. 276, no. 2, pp. 146–154, 07 1996. [Online]. Available: https://doi.org/10.1001/jama.1996.03540020068031

[67] M. A. Thompson, J. A. Aberg, J. F. Hoy, A. Telenti, C. Benson, P. Cahn, J. J. Eron, H. F. Günthard, S. M. Hammer, P. Reiss, D. D. Richman, G. Rizzardini, D. L. Thomas, D. M. Jacobsen, and P. A. Volberding, "Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society–USA panel," *The Journal of the American Medical Association*, vol. 308, no. 4, pp. 387–402, 2012.

[68] K. M. Harrison, R. Song, and X. Zhang, "Life expectancy after HIV diagnosis based on national HIV surveillance data from 25 states, United States," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 53, no. 1, pp. 124–130, 2010.

[69] H. R. Antiretroviral Therapy Cohort Collaboration, S. J. Lima V, B. M. Grabar S, D. M. A. Bonarek M, G. M. Esteve A, J. A. Harris R, L. F. Hayden A, M. M. Mocroft A, W. J. Staszewski S, K. M. van Sighem A, E. M. Guest J, and M. M., "Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies." *The Lancet*, vol. 372, no. 9635, pp. 293–299, 2008.

[70] T. Richards, V. M. Montori, F. Godlee, P. Lapsley, and D. Paul, "Let the patient revolution begin," *British Medical Journal*, vol. 346, no. 7908, p. 7, 2013. [Online]. Available: http://search.proquest.com/docview/1367590055/

[71] K. Young, "Doctors′ understanding of rheumatoid disease does not align with patients′ experiences," *British Medical Journal*, vol. 346, p. f2901, 2013.

[72] A. Boivin, A. Lesperance, F. Gauvin, V. Dumez, A. C. Macaulay, P. Lehoux, and J. Abelson, "Patient and public engagement in research and health system decision making: A systematic review of evaluation tools," *Health Expectations*, vol. 21, no. 6, pp. 1075–1084, 2018.

[73] J. Boote, R. Wong, and A. Booth, "'Talking the talk or walking the walk': A bibliometric review of the literature on public involvement in health research published between 1995 and 2009," *Health Expectations*, vol. 18, no. 1, pp. 44–57, 2015.

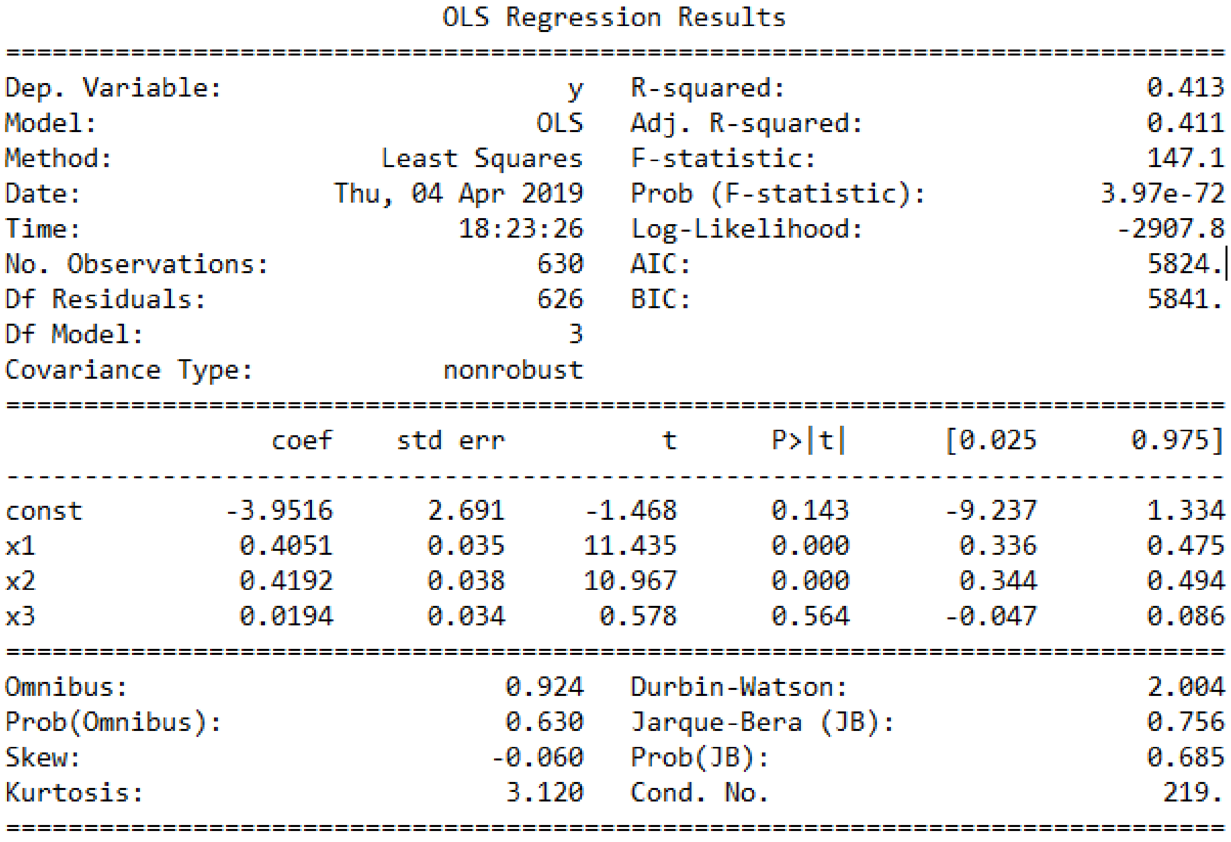
[74] "*Strategy for Patient-Oriented Research - Patient Engagement Framework - CIHR*," cihr-irsc.gc.ca. 2019. Accessed on: Aug. 30, 2019. [Online]. Available: http://www.cihr-irsc.gc.ca/e/48413.html.

[75] D. Nield. *All the Sensors in Your Smartphone and How They Work.* Gizmodo.com. 2017. Accessed on: Mar. 28, 2019. [Online]. Available: https://gizmodo.com/all-the-sensors-in-your-smartphone-and-how-they-work-1797121002.

Appendix A

Additional graphs of Chapter 3

A.0.1 Result of linear regression with 3 variables as independent and suicidality as dependent variable

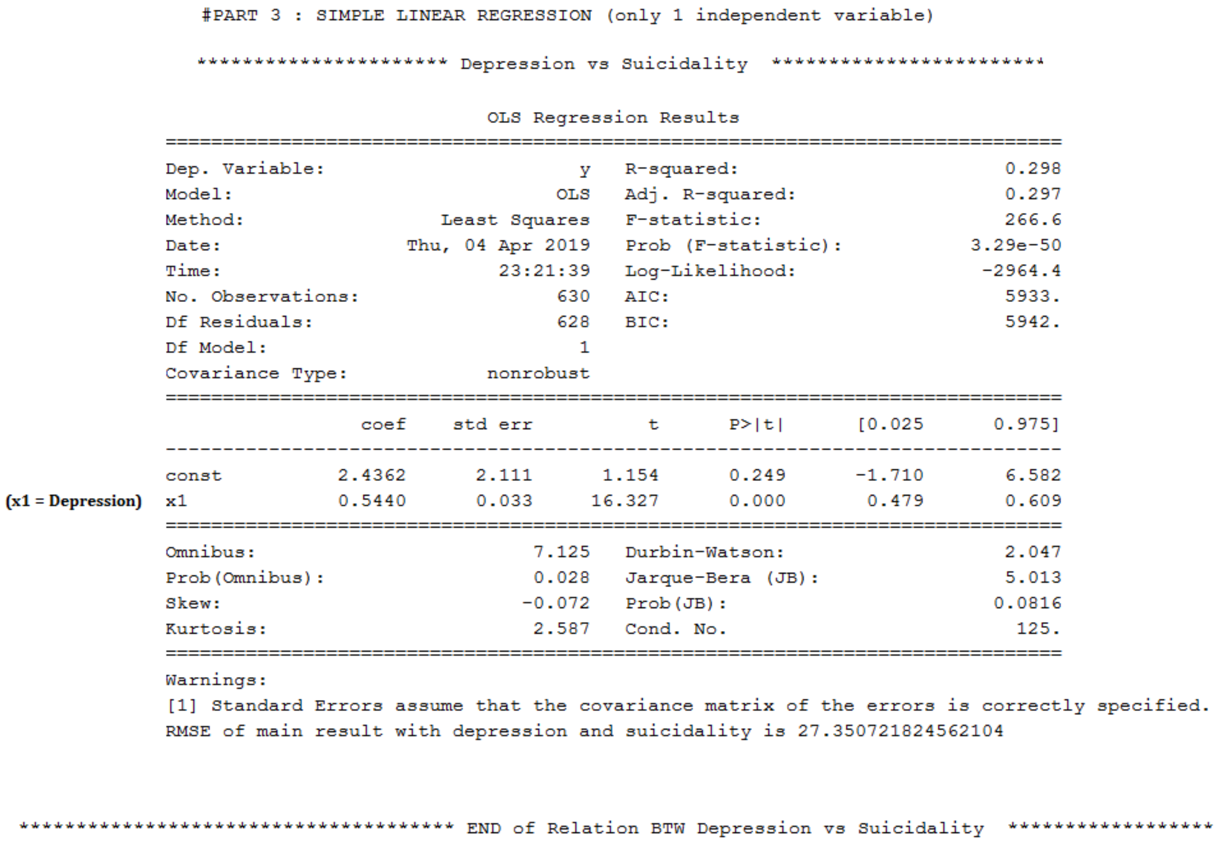


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.413
Model:                            OLS   Adj. R-squared:                  0.411
Method:                 Least Squares   F-statistic:                     147.1
Date:                Thu, 04 Apr 2019   Prob (F-statistic):           3.97e-72
Time:                        18:23:26   Log-Likelihood:                -2907.8
No. Observations:                 630   AIC:                             5824.
Df Residuals:                     626   BIC:                             5841.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.9516      2.691     -1.468      0.143      -9.237       1.334
x1             0.4051      0.035     11.435      0.000       0.336       0.475
x2             0.4192      0.038     10.967      0.000       0.344       0.494
x3             0.0194      0.034      0.578      0.564      -0.047       0.086
==============================================================================
Omnibus:                        0.924   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.630   Jarque-Bera (JB):                0.756
Skew:                          -0.060   Prob(JB):                        0.685
Kurtosis:                       3.120   Cond. No.                         219.
==============================================================================
```

Figure A.1: Multiple linear regression with Depression, Irritability and connectedness (3 independent variables) VS suicidality(response variable)

************************************ End of 3 variable relation ************************************

A.0.2 Simple linear regression (depression vs. suicidality)

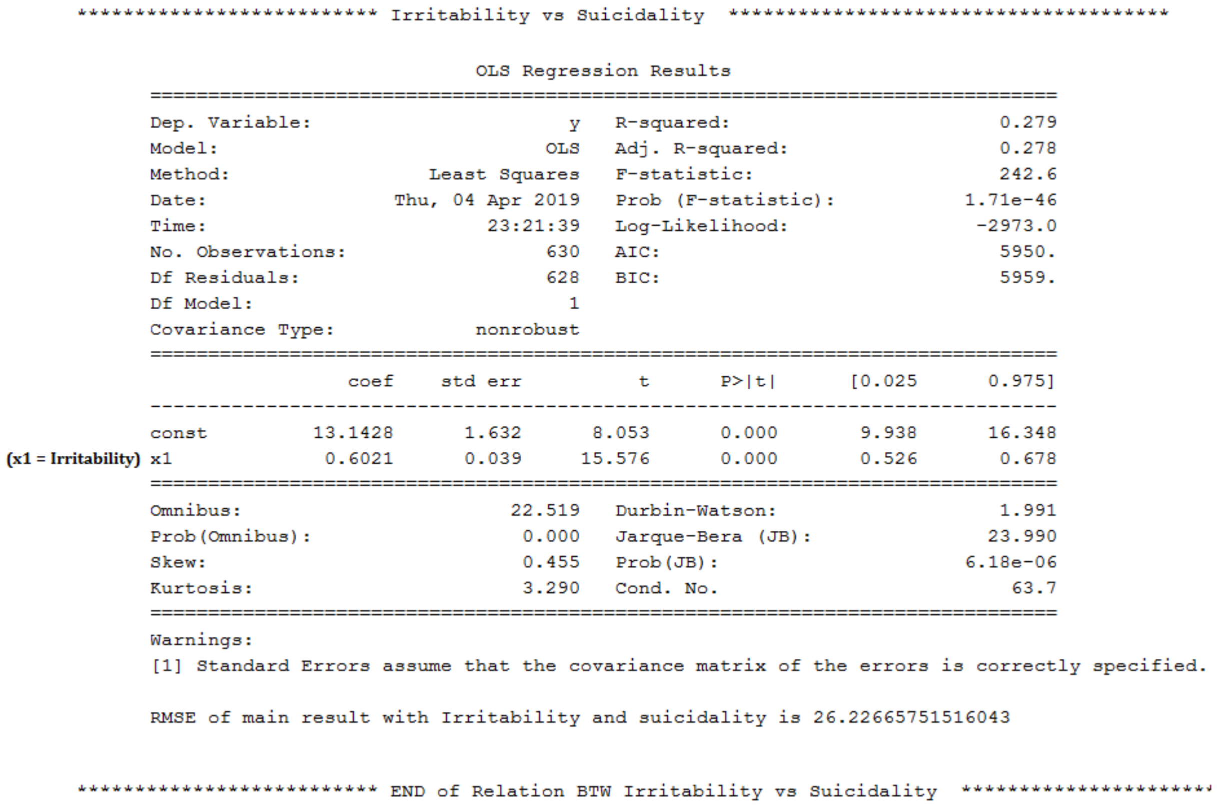


```
#PART 3 : SIMPLE LINEAR REGRESSION (only 1 independent variable)

********************** Depression vs Suicidality  ************************

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.298
Model:                            OLS   Adj. R-squared:                  0.297
Method:                 Least Squares   F-statistic:                     266.6
Date:                Thu, 04 Apr 2019   Prob (F-statistic):           3.29e-50
Time:                        23:21:39   Log-Likelihood:                -2964.4
No. Observations:                 630   AIC:                             5933.
Df Residuals:                     628   BIC:                             5942.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.4362      2.111      1.154      0.249      -1.710       6.582
x1             0.5440      0.033     16.327      0.000       0.479       0.609
==============================================================================
Omnibus:                        7.125   Durbin-Watson:                   2.047
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                5.013
Skew:                          -0.072   Prob(JB):                       0.0816
Kurtosis:                       2.587   Cond. No.                         125.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
RMSE of main result with depression and suicidality is 27.350721824562104


************************************** END of Relation BTW Depression vs Suicidality  *****************
```

Figure A.2: Simple Linear Depression vs. Suicidality

*********************************** End of Relation between depression vs suicidality **************************

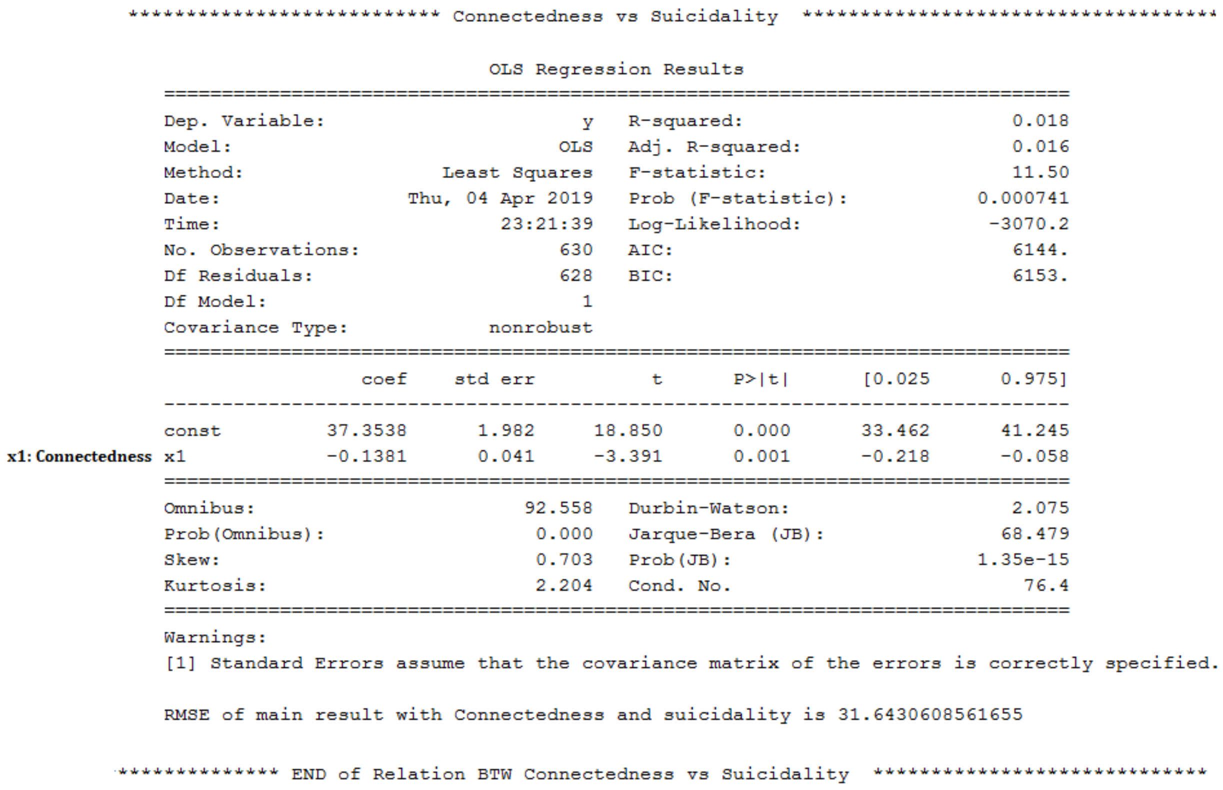
A.0.3 Simple linear regression (irritability vs. suicidality)



```
              ************************** Irritability vs Suicidality  **************************************

                                          OLS Regression Results
                    ==============================================================================
                    Dep. Variable:                      y   R-squared:                       0.279
                    Model:                            OLS   Adj. R-squared:                  0.278
                    Method:                 Least Squares   F-statistic:                     242.6
                    Date:                Thu, 04 Apr 2019   Prob (F-statistic):           1.71e-46
                    Time:                        23:21:39   Log-Likelihood:                -2973.0
                    No. Observations:                 630   AIC:                             5950.
                    Df Residuals:                     628   BIC:                             5959.
                    Df Model:                           1
                    Covariance Type:            nonrobust
                    ==============================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
                    ------------------------------------------------------------------------------
                    const         13.1428      1.632      8.053      0.000       9.938      16.348
(x1 = Irritability) x1             0.6021      0.039     15.576      0.000       0.526       0.678
                    ==============================================================================
                    Omnibus:                       22.519   Durbin-Watson:                   1.991
                    Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.990
                    Skew:                           0.455   Prob(JB):                     6.18e-06
                    Kurtosis:                       3.290   Cond. No.                         63.7
                    ==============================================================================
                    Warnings:
                    [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

                    RMSE of main result with Irritability and suicidality is 26.22665751516043


              ************************** END of Relation BTW Irritability vs Suicidality  **********************
```

Figure A.3: Simple linear: irritability vs. suicidality

A.0.4 Simple linear regression (connectedness vs. suicidality)



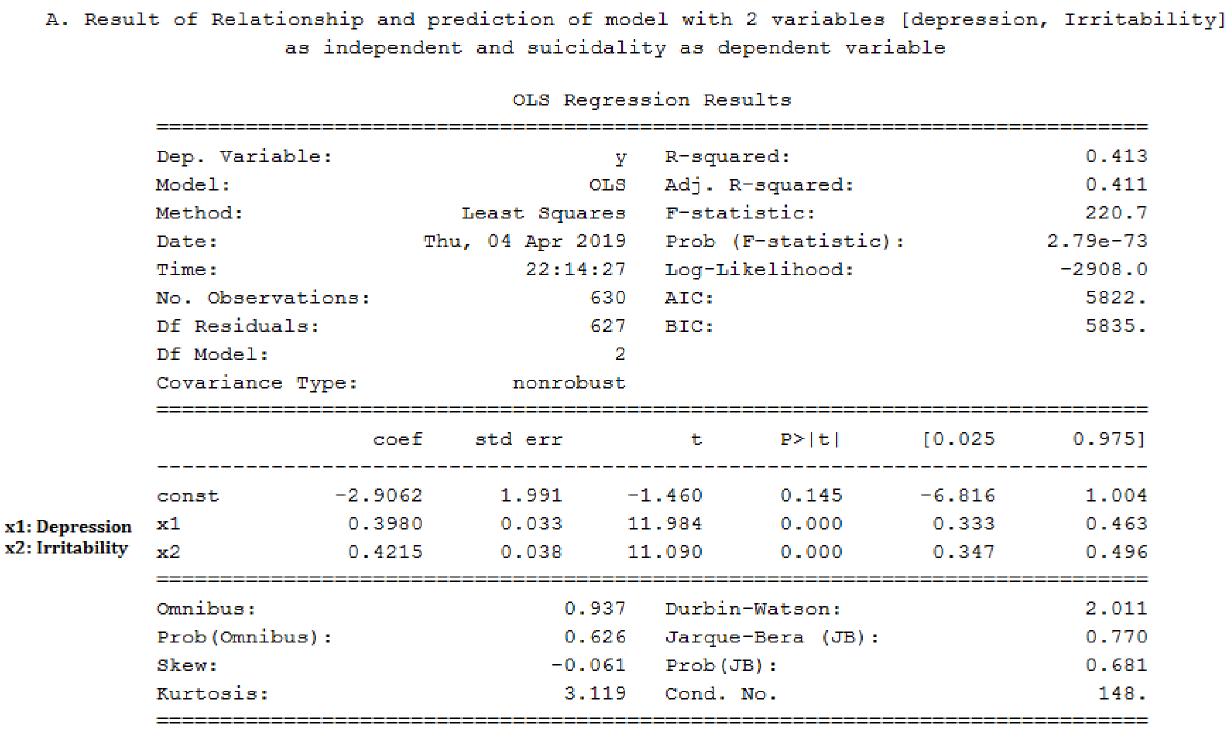
```
*************************** Connectedness vs Suicidality  ************************************

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.018
Model:                            OLS   Adj. R-squared:                  0.016
Method:                 Least Squares   F-statistic:                     11.50
Date:                Thu, 04 Apr 2019   Prob (F-statistic):           0.000741
Time:                        23:21:39   Log-Likelihood:                -3070.2
No. Observations:                 630   AIC:                             6144.
Df Residuals:                     628   BIC:                             6153.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         37.3538      1.982     18.850      0.000      33.462      41.245
x1            -0.1381      0.041     -3.391      0.001      -0.218      -0.058
==============================================================================
Omnibus:                       92.558   Durbin-Watson:                   2.075
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               68.479
Skew:                           0.703   Prob(JB):                     1.35e-15
Kurtosis:                       2.204   Cond. No.                         76.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

RMSE of main result with Connectedness and suicidality is 31.6430608561655

************** END of Relation BTW Connectedness vs Suicidality  *****************************
```

Figure A.4: Simple linear connectedness vs. suicidality

Part 3:

Multiple linear regression with 2 independent variables and predicting suicidality

A.0.5 Multiple linear regression with 2 independent variables [depression, irritability] as independent and suicidality as dependent variable

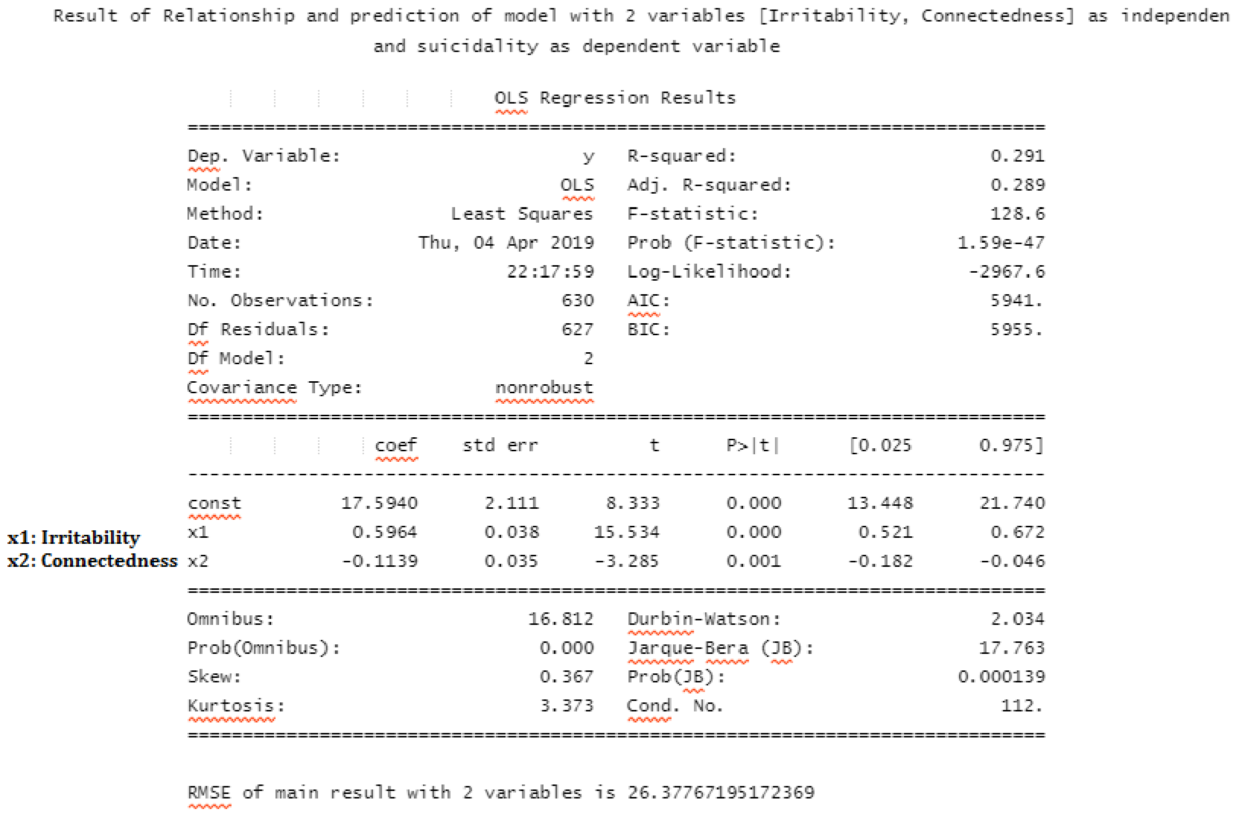


```
A. Result of Relationship and prediction of model with 2 variables [depression, Irritability]
                   as independent and suicidality as dependent variable

                                OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.413
Model:                            OLS   Adj. R-squared:                  0.411
Method:                 Least Squares   F-statistic:                     220.7
Date:                Thu, 04 Apr 2019   Prob (F-statistic):           2.79e-73
Time:                        22:14:27   Log-Likelihood:                -2908.0
No. Observations:                 630   AIC:                             5822.
Df Residuals:                     627   BIC:                             5835.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.9062      1.991     -1.460      0.145      -6.816       1.004
x1             0.3980      0.033     11.984      0.000       0.333       0.463
x2             0.4215      0.038     11.090      0.000       0.347       0.496
==============================================================================
Omnibus:                        0.937   Durbin-Watson:                   2.011
Prob(Omnibus):                  0.626   Jarque-Bera (JB):                0.770
Skew:                          -0.061   Prob(JB):                        0.681
Kurtosis:                       3.119   Cond. No.                         148.
==============================================================================
```

Figure A.5: Multiple linear: depression and irritability vs. suicidality

RMSE of main result with 2 variables is 24.295802743406625

A.0.6 Multiple linear regression with 2 independent variables [irritability, connectedness] as independent and suicidality as dependent variable

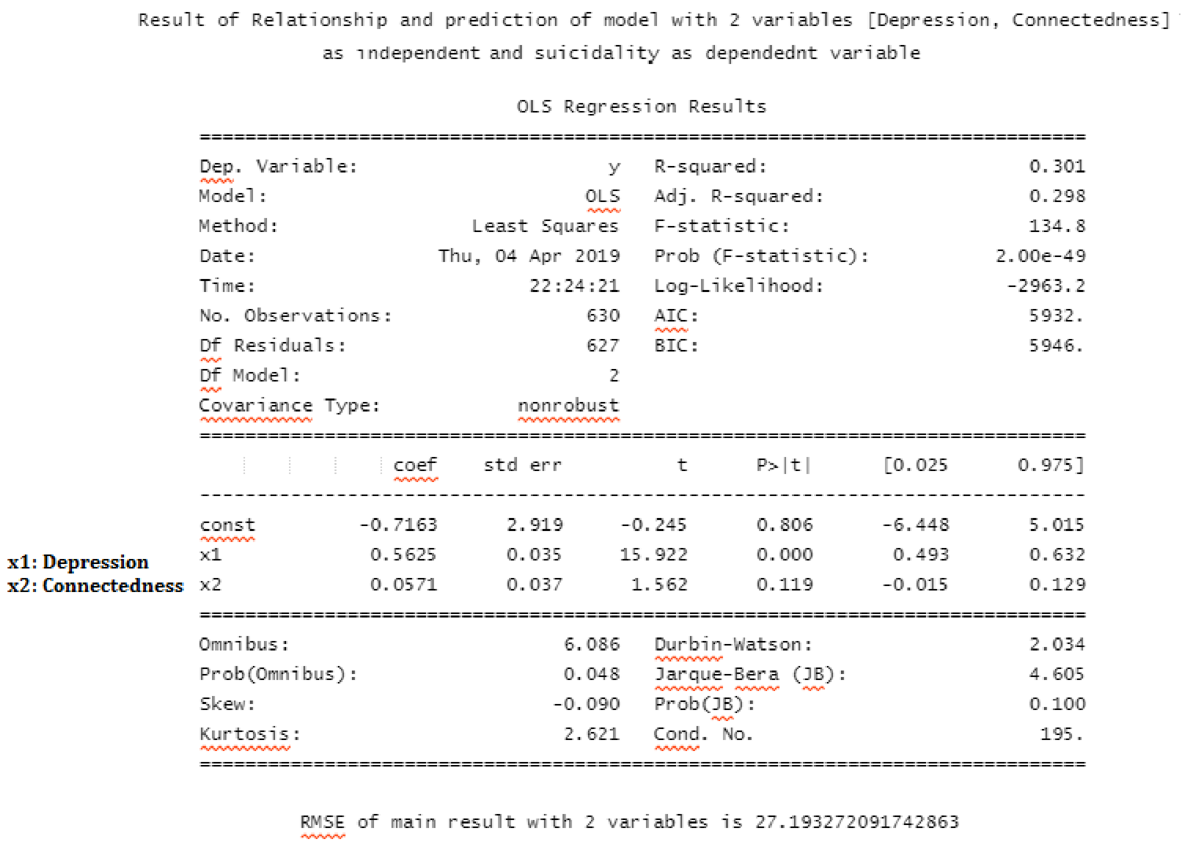


```
Result of Relationship and prediction of model with 2 variables [Irritability, Connectedness] as independen
                          and suicidality as dependent variable

                              OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.291
Model:                            OLS   Adj. R-squared:                  0.289
Method:                 Least Squares   F-statistic:                     128.6
Date:                Thu, 04 Apr 2019   Prob (F-statistic):           1.59e-47
Time:                        22:17:59   Log-Likelihood:                -2967.6
No. Observations:                 630   AIC:                             5941.
Df Residuals:                     627   BIC:                             5955.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         17.5940      2.111      8.333      0.000      13.448      21.740
x1             0.5964      0.038     15.534      0.000       0.521       0.672
x2            -0.1139      0.035     -3.285      0.001      -0.182      -0.046
==============================================================================
Omnibus:                       16.812   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.763
Skew:                           0.367   Prob(JB):                     0.000139
Kurtosis:                       3.373   Cond. No.                         112.
==============================================================================

RMSE of main result with 2 variables is 26.37767195172369
```

Figure A.6: Multiple linear regression: Irritability and connectedness vs. suicidality

RMSE of main result with 2 variables is 26.37767195172369

A.0.7 Multiple linear regression with 2 independent variables [depression, connectedness] as independent and suicidality as dependent variable



```
Result of Relationship and prediction of model with 2 variables [Depression, Connectedness]
               as independent and suicidality as dependednt variable

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.301
Model:                            OLS   Adj. R-squared:                  0.298
Method:                 Least Squares   F-statistic:                     134.8
Date:                Thu, 04 Apr 2019   Prob (F-statistic):           2.00e-49
Time:                        22:24:21   Log-Likelihood:                -2963.2
No. Observations:                 630   AIC:                             5932.
Df Residuals:                     627   BIC:                             5946.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.7163      2.919     -0.245      0.806      -6.448       5.015
x1             0.5625      0.035     15.922      0.000       0.493       0.632
x2             0.0571      0.037      1.562      0.119      -0.015       0.129
==============================================================================
Omnibus:                        6.086   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.048   Jarque-Bera (JB):                4.605
Skew:                          -0.090   Prob(JB):                        0.100
Kurtosis:                       2.621   Cond. No.                         195.
==============================================================================

              RMSE of main result with 2 variables is 27.193272091742863
```

x1: Depression
x2: Connectedness

Figure A.7: Multiple linear regression: depression and connectedness vs. suicidality

A.0.8 Count of surveys answered (day wise and hour wise)

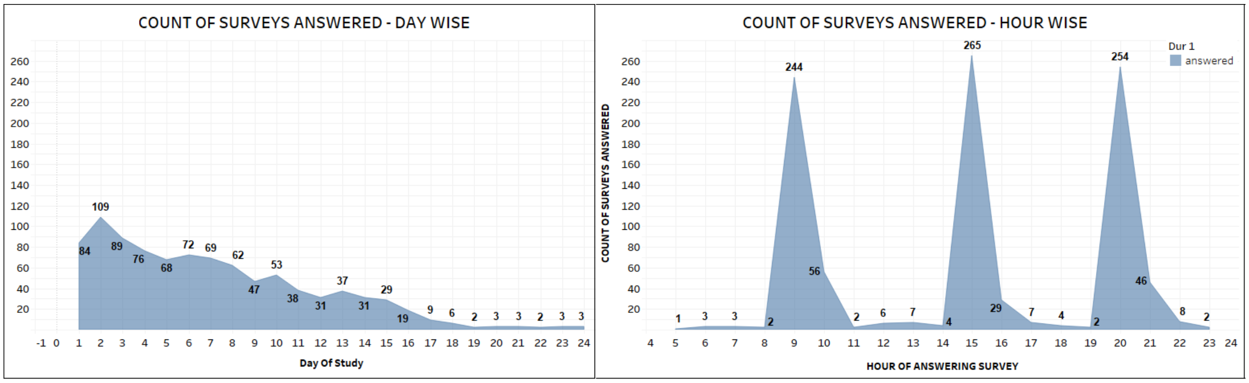


Figure A.8: Suicidal ideation Chapter 3: hours of survey reporting