ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

THESIS PRESENTED TO
ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

IN PARTIAL FULFILLEMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph. D.

BY
Edward Arne HILL

AN EVIDENCE-BASED TOOLSET TO CAPTURE, MEASURE, ANALYZE & ASSESS EMOTIONAL HEALTH

MONTREAL, APRIL 3, 2014

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

M. Pierre Dumouchel, Thesis Supervisor
Directeur général
École de technologie supérieure

M. Jérémie Voix, President of the Board of Examiners
Département de génie logiciel et des TI
École de technologie supérieure

M. Francois Coallier, Member of Jury
Département de génie logiciel et des TI
École de technologie supérieure

M. Pierrich Plusquellec, Member of Jury
Co-director Centre for Studies on Human Stress
Associate professor, École de Psychoéducation
Université de Montréal

M. Charles Moehs, External Evaluator
Occupational Medicine Associates,
Watertown, NY

THIS THESIS WAS PRESENTED AND DEFENDED
IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

JANUARY 10, 2014

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENT

This thesis was made possible by guidance from people I am privileged to know. I am most grateful to all of them.

Automatic emotion detection and speech processing guidance was provided by Dr. Pierre Dumouchel, and Dr. Najim Dehak.

Emotion Health and addiction therapy guidance was provided by:

- Dr. Charles Moehs, M.D. MPH Occupational Medicine Associates, Watertown, NY
- Rabbi Benyamin Bresinger Director, Chabad LifeLine;
- Mrs. Keren Bresinger MSW, Clinical Director and Family counselor, Chabad LifeLine;
- Ruth Weinberger ICADC International Certified Alcohol and Drug Counselor and certified Family Life Educator Chabad LifeLine;
- Shaun Sullivan, Behavioral specialist, Verdun Elementary school;
- Dr. Janice Goldfarb, M.D. M.Sc. Medical Director, Methadone Maintenance Clinic, Montreal Jewish General Hospital;
- … and members of Alcoholics Anonymous.

Training and assistance in statistical analysis over 2011 and 2012 was provided by

- Dr. Sonia Lupien, Scientific Director, Institut universitaire en santé mentale de Montréal;
- Heather Krause, Principal, Datassist;
- Dr. Georges Monette, Department of Mathematics and Statistics, York University;
- Dr. John Fox, Department of Sociology, McMaster University.

Moral support was provided by Greta & Rhys Roberts, Robert Hill, and Zak Hill.

AN EVIDENCE-BASED TOOLSET TO CAPTURE, MEASURE, ANALYZE & ASSESS EMOTIONAL HEALTH

Edward Arne HILL

ABSTRACT

This thesis describes the development and validation of an evidence-based toolkit that captures a patient's emotional state, expressiveness/affect, self-awareness, and empathy during a fifteen second telephone call, and then accurately measures and analyzes these indicators of Emotional Health based on emotion detection in speech and multilevel regression analysis.

An emotion corpus of eight thousand three hundred and seventy-six (8,376) momentary emotional states was collected from one hundred and thirteen (113) participants including three groups: Opioid Addicts undergoing Suboxone® treatment, the General Population, and members of Alcohol Anonymous. Each collected emotional state includes an emotional recording in response to "How are you feeling?" a self-assessment of emotional state, and an assessment of an emotionally-charged recording. Each recording is labeled with the emotional truth. A method for unsupervised emotional truth corpus labeling through automatic audio chunking and unsupervised automatic emotional truth labeling is proposed and experimented.

In order to monitor and analyze the emotional health of a patient, algorithms are developed to accurately measure the emotional state of a patient in their natural environment. Real-time emotion detection in speech provides instantaneous classification of the emotional truth of a speech recording. A pseudo real-time method improves emotional truth accuracy as more data becomes available. A new measure of emotional truth accuracy, the certainty score, is introduced. Measures of self-awareness, empathy, and expressiveness are derived from the collected emotional state.

Are there differences in emotional truth, self-assessment, self-awareness, and empathy across groups? Does gender have an effect? Does language have an effect? Does length of the response, as an indication of emotional expressiveness, vary with emotion or group? Does confidence of the emotional label, as an indication of affect, vary with emotion or group? Are there differences in call completion rates? Which group would be more likely to continue in data collections? Significant results to these questions will provide evidence that capturing and measuring Emotional Health in speech can:

- Assist therapists and patients in Cognitive Behavioural Therapy to become aware of symptoms and make it easier to change thoughts and behaviours;

- Provide evidence of psychotropic medication and psychotherapy effectiveness in mental health and substance abuse treatment programs;

- Accelerate the interview process during monthly assessments by physicians, psychiatrists, and therapists by providing empirical insight into emotional health of patients in their natural environment.

- Trigger crisis intervention on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions.

BOÎTE À OUTILS BASÉE SUR DES ÉVIDENCES POUR CAPTURER, MESURER, ANALYSER ET ÉVALUER LA SANTÉ MENTALE

Edward Arne HILL

RÉSUMÉ

Cette thèse décrit le développement et la validation d'une boîte à outils fondée sur des preuves qui capture l'état émotionnel, l'expressivité / affect, la conscience de soi et l'empathie d'un patient au cours d'un appel téléphonique de quinze secondes, puis mesure et analyse avec précision ces indicateurs de la santé émotionnelle basée sur la détection des émotions à partir de la voix et son analyse par régression multi-niveaux.

Un corpus d'échantillons de parole de téléphonique de 8376 (8,376) états émotionnels momentanées ont été recueillis. Cent treize (113) individus issus trois groupes ont participé à cette collecte: les toxicomanes traités avec le médicament Suboxone ®, la population en général, et les membres des alcooliques anonymes. Chaque état émotionnel recueilli comprend un enregistrement sonore de la réponse à la question "Comment allez-vous aujourd'hui?" Une auto-évaluation de son propre état émotionnel et de sa réaction à des échantillons émotionnels provenant de tierces personnes sont enregistrées. De plus, une approche non supervisée d'étiquetage automatique du véritable état émotionnel est proposée et expérimentée.

Afin de surveiller et d'analyser la santé émotionnelle d'un patient, les algorithmes sont développés pour mesurer avec précision l'état émotionnel d'un patient dans leur environnement naturel. La détection des émotions en temps réel d'un signal de parole permet la classification instantanée de la vérité émotionnelle d'un enregistrement de la parole. Une méthode en pseudo temps réel améliore la précision de la vérité émotionnelle au fur et à mesure que de nouvelles données audio deviennent disponibles. De plus, une nouvelle mesure de la précision de la vérité émotionnelle, le score de certitude, est proposée. Les mesures de la conscience de soi, d'empathie et d'expressivité sont tirées de l'état émotionnel recueilli.

Y at-il des différences dans la vérité émotionnelle, l'auto-évaluation, la conscience de soi et l'empathie entre les groupes? Est-ce que le sexe du participant influence l'étiquetage? Est-ce que la langue a un effet? Est-ce que la longueur de la réponse, comme une indication de l'expressivité émotionnelle, varie avec l'émotion ou avec le groupe auquel appartient le participant? Est-ce que la confiance de l'étiquetage émotionnel, comme une indication de l'affect, varient en fonction de l'émotion ou du groupe? Y at-il des différences dans les taux de réussite des appels? Quel groupe serait le plus susceptible à persévérer dans ce type d'analyse? Des résultats significatifs à ces questions fourniront la preuve que la capture et la mesure de la santé émotionnelle dans le discours permettent:

- D'aider les thérapeutes et les patients en thérapie cognitivo-comportementale à prendre conscience des symptômes et de faciliter les changements de pensée et de comportement;

- De fournir des preuves de l'efficacité du traitement avec des médicaments psychotropes de même que de l'efficacité des sessions de psychothérapie dans les programmes de traitement de la toxicomanie et de la santé mentale;

- D'accélérer le processus d'entrevue lors des évaluations mensuelles des médecins, des psychiatres et des thérapeutes en donnant un aperçu empirique sur la santé émotionnelle des patients dans leur environnement naturel.

- De détecter des situations de crise suite à des séquences prolongées sur plusieurs jours de non-enregistrement d'échantillons (situation de crise d'isolement) ou d'une situation de déprime exprimée par une séquence consécutives d'états négatifs.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF SOFTWARE CODE SNIPPETS

LIST OF ABREVIATIONS

AA	Alcoholics Anonymous
A-ESM	IVR Acoustic ESM
AIC	Akaike Information Criterion
AJAX	Asynchronous JavaScript and XML
AMI	Any mental illness
BECARS	Open-source software for speaker recognition
BIC	Bayesian information criterion
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CARD	Comprehensive Analysis of Reported Drugs
CBT	Cognitive Behavioural Therapy
CCXML	Call Control eXtended Markup Language
CI	Confidence Interval
CMF	Content Management Framework
CMS	Content Management System
CPU	Central Processing Unit
CRIM	Centre de recherche informatique de Montréal
CRON	Command Run On (UNIX scheduler)
CSC	Canadian Correctional Service
df	Degrees of freedom
DNA	Deoxyribonucleic acid
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders 4th edition
DTMF	Dual Tone Multi-Frequency
EBA	Evidence-Based Assessment
EC	Effortful Control
EM	Expectation-Maximization
EMA	Ecological Momentary Assessment
ESM	Experience Sampling Method

F1, F2… FN	Frequency Formants
GLLAMM	Stata's Generalized Linear Latent And Mixed Models
GMM	Gaussian Model Mixtures
GNU	GNU's Not Unix
GP	General Population
GPS	Global Positioning System
GRXML	Grammar eXtended Markup Language
HLM	Hierarchical Linear Model
HMM	Hidden Markov Models
HTK	Hidden Markov Models Toolkit
HTML	HyperText Markup Language
ICC	Intraclass Correlation Coefficient
IGT	Iowa Gambling Task
IIA	Independence from Irrelevant Alternatives
IMPACT	International Mission on Prognosis and Clinical Trial design in TBI
INTERSPEECH	International Speech Communication Association
iOS	iPhone Operating System
IQR	Interquartile Range
IRB	Institutional Review Board
IVR	Interactive Voice Response System
JFA	Joint Factor Analysis
LAMP	Linux-Apache-MySQL-PHP
LogLik	Log Likelihood
LPGL	GNU Lesser General Public License
LPGL	Lesser General Public License
MAP	Maximum-a-Posteriori
MCMCglmm	Markov Chain Monte Carlo Generalized Linear Mixed Models
MCS	Multiple Classifier Systems
MD	Medical Doctor
MFCC	Mel-Scale Cepstral Coefficients

MIT	Massachusetts Institute of Technology
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate
MLLR	maximum likelihood linear regression
MP3	Moving Picture Experts Group (MPEG) Audio Layer 3
MPH	Master of Public Health
MV	Majority Vote
MySQL	My Structured Query Language; database management system
NA	Negative Affect
NIMH	National Institute of Mental Health
NIST	National Institute of Standards and Technology
OCR	Optical Character Recognition
OLS	Ordinary Least Squares
PA	Positive Affect
Perl	Practical Extraction and Reporting Language
PHP	PHP: Hypertext Preprocessor
PIN	Personal Identification Number
PRO	Patient-Reported-Outcome
PSTN	Public Switched Telephone Network
R	The R programming language
ReCAPTCHA	CAPTCHA to improve digitization process
REML	Restricted Maximum Likelihood
SAMHSA	Substance Abuse and Mental Health Services Administration
SD	Standard Deviation
SE	Standard Error
SFTP	Secure File Transfer Protocol
SFTP	Secure File Transfer Protocol
SMS	Short Message Service
SPIDA	York University's Summer Program in Data Analysis
SQL	Software Query Language

SUBX	Opioid Addicts undergoing Suboxone® treatment
SVM	Support Vector Machines
TSD	Time sampling diary
UA	Unweighted Average
UBM	Universal Background Models
VAD	Voice Activity Detector
VoiceXML	Voice eXtended Markup Language
W3C	World Wide Web Consortium

LIST OF SYMBOLS

$participant_j$	The j^{th} patient in an experience sample collection trial
c_{ij}	The i^{th} telephone call for $participant_j$
$c_{ij}^{calltype}$	The call type of c_{ij} (inbound or outbound)
c_{ij}^{time}	The timestamp (date + time) of c_{ij}
$c_{ij}^{duration}$	The call duration in seconds of c_{ij}
c_{ij}^{state}	The call state of c_{ij}
	$c^{state} \in \{line\ busy, not\ answered\ , call\ complete\}$
X_{ij}	The i^{th} speech recording for $participant_j$ captured during call c_{ij}
$group_g$	The trial group
$e_{ij}^{self}(X_{ij})$	The i^{th} emotional self-assessment for $participant_j$ of X_{ij}
	$e^{self} \in \{okay, happy, sad, angry, and\ anxious\}$
e^{self}	Short form of $e_{ij}^{self}(X_{ij})$
$\mathrm{e}_{\mathrm{ijka}}^{\mathrm{relate}}(X_{ka})$	The i^{th} emotional assessment for $participant_j$ of a randomly selected recording X_{ka} (k^{th} speech recording from another $participant_a$). $e^{relate} \in \{okay, happy, sad, angry, and\ anxious\}$
e^{relate}	Short form of $\mathrm{e}_{\mathrm{ijka}}^{\mathrm{relate}}(X_{ka})$
$\mathrm{E}_{\mathrm{ij}}^{relate}$	Multiple $\mathrm{e}_{\mathrm{ijka}}^{\mathrm{relate}}$ assessments can be captured on the same call c_{ij}. $\mathrm{E}_{\mathrm{ij}}^{\mathrm{relate}}$ is the set of all emotional assessments.
E^{relate}	Short form of $\mathrm{E}_{\mathrm{ij}}^{\mathrm{relate}}$
ESM_{ij}	The i^{th} experience sample for $participant_j$
$e_{ij}^{truth}(X_{ij})$	The i^{th} ground truth emotion for $participant_j$ of X_{ij}
e^{truth}	Short form of $e_{ij}^{truth}(X_{ij})$
$e_{ij}^{detect}(X_{ij})$	The i^{th} approximation of $e_{ij}^{truth}(X_{ij})$ for $participant_j$ calculated by automatic emotion detection
e^{detect}	Short form of $e_{ij}^{detect}(X_{ij})$

$e_{ij}^{crowd}(X_{ij})$	The i^{th} approximation of $e_{ij}^{truth}(X_{ij})$ for $participant_j$ calculated from majority vote crowd-sourcing
e^{crowd}	Short form of $e_{ij}^{crowd}(X_{ij})$
$e_{ij}^{crowdself}(X_{ij})$	The i^{th} approximation of $e_{ij}^{truth}(X_{ij})$ for $participant_j$ calculated from majority vote crowd-sourcing and e^{self} and e^{detect}
$e^{crowdself}$	Short form of $e_{ij}^{crowdself}(X_{ij})$
$e_{ij}^{crowdselfedetect}(X_{ij})$	The i^{th} approximation of $e_{ij}^{truth}(X_{ij})$ for $participant_j$ calculated from majority vote crowd-sourcing, e^{self} and
$e^{crowdselfedetect}$	Short form of $e_{ij}^{crowdselfedetect}(X_{ij})$
$length_{ij}(X_{ij})$	The i^{th} length in seconds of X_{ij} for $participant_j$
$e_{ij}^{confidence}(X_{ij})$	The i^{th} confidence score in the approximation of $e_{ij}^{truth}(X_{ij})$ for $participant_j$
$e^{confidence}$	Short form of $e_{ij}^{confidence}(X_{ij})$
$e_{ij}^{certainty}(X_{ij})$	The i^{th} certainty score of $e_{ij}^{confidence}$ for $participant_j$ (Scaled by reCAPTCHA response factor)
$e^{certainty}$	Short form of $e_{ij}^{certainty}(X_{ij})$
$\sigma_{\bar{y}}^2$	Variance of a sample mean
$\bar{y}$	Sample mean

INTRODUCTION

Mental health and substance abuse professionals need new evidence-based methods to cost-effectively and time-effectively diagnose, monitor, assist in decision making, and treat the tens of millions of people affected by mental health disorders and substance abuse every year.

This thesis is a cross-disciplinary study combining software engineering, speech science, and elements of psychological research towards the development and validation of an evidence-based toolkit that captures a patient's emotional state, expressiveness/affect, self-awareness, and empathy during a fifteen second telephone call, and then accurately measures and analyzes these four indicators of Emotional Health based on emotion detection in speech and multilevel regression analysis. This research presents the only known statistically validated[1] system that measures a patient's emotional state, expressiveness/affect, self-awareness, and empathy in a patient's natural environment[2].

Emotional Speech

Emotional speech can be elicited by asking the quintessential question ***"how do you feel?"*** It is human nature to colour our response to this question with emotion [2]. Most people can infer something of the person's psychological state from vocal changes [3]. In the post-trial survey summarized in section 1.5.3, 85% of trial participants indicated they listened to how the speaker spoke, rather than what was said, to determine emotion.

Emotion produces changes in respiration, phonation, articulation, and energy [4]. As emotional intensity increases, frequency and duration of pauses decrease [5]. Acoustic variables strongly involved in vocal emotions include level, range, contour of the

[1] Statistical methods are described in Chapter 1. Hypothesis and analysis of trial data are presented in Chapter 5 and the General Conclusion.

[2] There are systems that have been emerging that measure distress and depression over the telephone since the start of this thesis in 2009, such as Cogito's Social Signal Platform [1]

fundamental frequency F0; vocal energy; distribution of energy in the frequency spectrum; location of the frequency formants (F1, F2,...,FN); tempo (speaking rate), rate and length of pauses [4, 6].

Unemotional speech (**Neutral**) has a much narrower pitch range than that of emotional speech [5]. Fear and **Anxiety** are characterized by an increase in mean F0, F0 range, and high-frequency energy; an accelerated rate of articulation, and pauses typically comprising 31% of total speaking time. (An increase in mean F0 is evident for milder forms of fear such as worry or anxiety) [7]. **Sadness** corresponds in a decrease in mean F0, F0 range, and mean energy as well as downward-directed F0 contours; slower tempo; irregular pauses [7]. **Happiness** produces an increase in mean F0, F0 range, F0 variability, and mean energy; and there may be an increase in high-frequency energy and rate of articulation [7].

Emotional states with high and low level of arousal are hardly ever confused, but it is difficult to determine the emotion of a person with flat affect [5]. Emotions that are close in the activation-evaluation emotional space (flat affect) often tend to be confused [7].

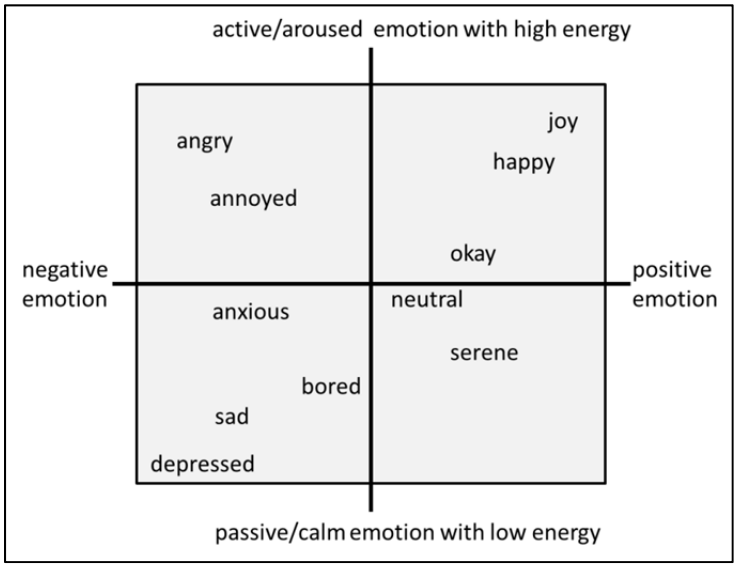


Figure 1 Activation-Evaluation Emotional Space.

Thesis Goals

1. Build an Interactive Voice Response (IVR) cloud platform to monitor and analyze the emotional health of a patient in their natural environment.

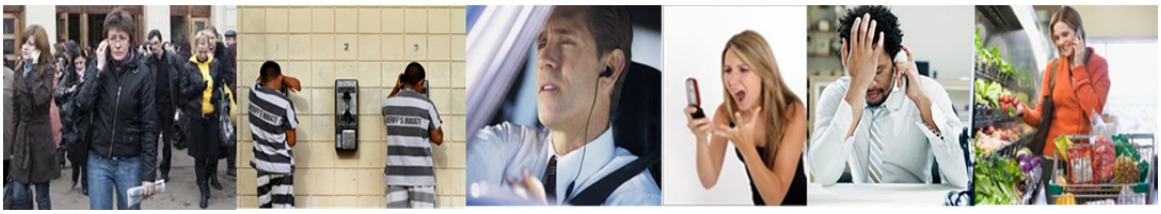


Figure 2 Monitoring patients in their natural environment

2. Sample, capture, and collect an emotional speech corpus of sufficient size to enable measurement and statistical analysis.

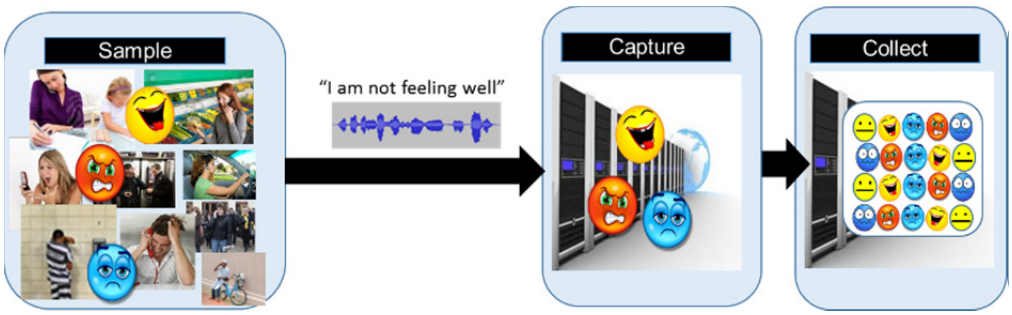


Figure 3 Sample, Capture, and Collect

3. Devise an unsupervised crowd-sourced emotional speech corpus labeling technique.

Figure 4 Unsupervised labeling

4. Accurately measure the emotional health of a patient over time.

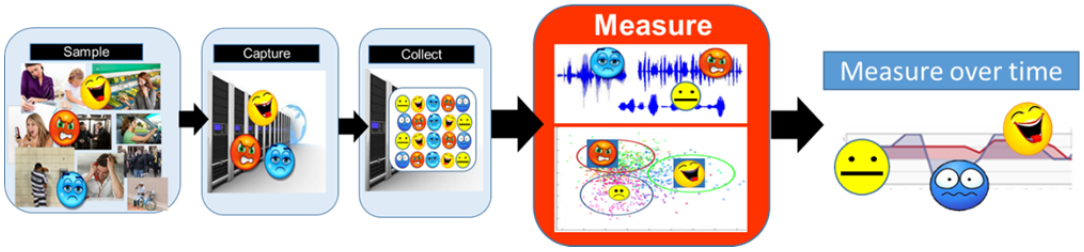


Figure 5 Measure emotional health over time

5. Devise a real-time auditable approach to emotional health measurement for monitoring patients. This method will improve the accuracy of measurements as reinforcement data becomes available; and provide an indication of the confidence and certainty of the measurement.

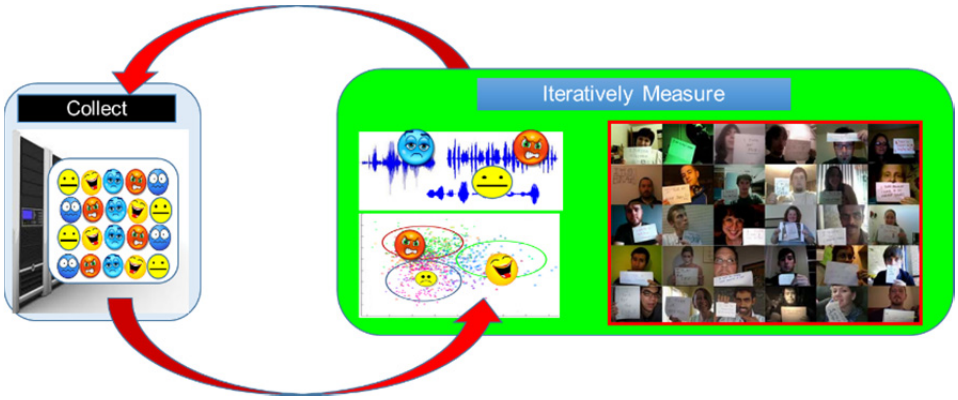


Figure 6 Iteratively improve emotional health measurements

6. Evidence-based practices are interventions for which there is consistent scientific evidence showing that they improve client outcomes [8]. In general the highest standard is several randomized clinical trials comparing the practice to alternative practices or to no intervention [8]. A key outcome of this thesis is to provide statistical evidence that capturing and measuring Emotional Health in speech can provide a mechanism:

a. To assist Cognitive Behavioural Therapy (CBT) for psychiatrists and therapists and patients to become aware of symptoms and make it easier to change thoughts and behaviors;

Figure 7 Psychotherapy

b. For evidence of psychotherapy effectiveness in mental health and substance abuse treatment programs;

c. For Medical Doctors and Psychiatrists to measure the effectiveness of psychotropic medication.

Figure 8 Medication

7. Devise patient monitoring and trend analysis tools to provide empirical insight into a patient's emotional health and accelerate the interview process during monthly assessments by overburdened physicians and psychotherapists. Crisis intervention can be triggered on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions.

Figure 9 Triggering crisis intervention

Crisis in Mental Health

In Canada, only one-third of those who need mental health[3] services actually receive them [8]. An experiment conducted in 2011 in Boston to obtain a psychiatric appointment within 2 weeks after discharge from emergency services resulted in only 10 out of 64 facilities (15.6%) able to schedule an appointment [9].

The estimated American outpatient medical care expenditure for mental health is $592 Billion per year with the proportion of psychotherapy expenditure at 44.7% or $264 Billion [10]. The Canadian public health system cost for addiction treatment is $5 Billion annually [11]. In 2004, 56 % of inmates in State prisons and 45% of inmates in Federal prisons had a mental health problem in the past year [12]. 48% of the 1.6 million inmates in federal prison were serving time for a drug offence [13].

The Substance Abuse and Mental Health Services Administration (SAMSHA) reported that among the American adult population aged 18 or older (hereafter "adults") in 2011, there were an estimated 19.6% (45.6 million) with any mental illness (AMI) [12].

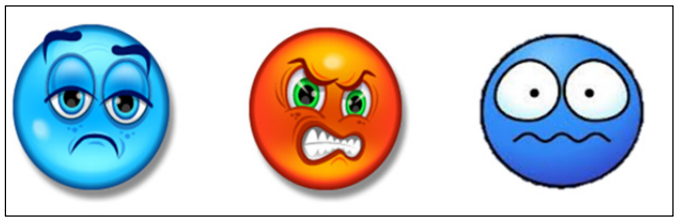


Figure 10 Mood Disorders

The National Institute of Mental Health (NIMH) estimate for diagnosed mental health disorders in the adult population is 26.2% (57.7 million)[4] per year. 9.5% (20.9 million) have

[3] The American Substance Abuse and Mental Health Services Administration definition for "any mental illness" is as follows: having (currently or at any time in the past year) a diagnosable mental, behavioral, or emotional disorder (excluding developmental and substance use disorders) of sufficient duration to meet diagnostic criteria specified within the DSM-IV.

[4] SAMSHA AMI statistics do not include homeless people, institutionalized people, or the military. Hence the

a mood disorder, 6.7% (14.8 million) have a major depressive disorder, and 18.1% (40 million) have an anxiety disorder[5] [14]. In 2008, 21% of military personnel in all services reported symptoms that suggested the need for further depression evaluation. 17% of military personnel had received mental health counselling in the past year [12].

Crisis in Substance Abuse

Mood disorder and anxiety are directly associated with substance abuse [15]. The National Epidemiologic Survey on Alcohol and Related Conditions [16] performed a survey of 43,093 respondents. Among respondents with any drug use disorder who sought treatment, 60.31% had at least one independent mood disorder, 42.63% had at least one independent anxiety disorder, and 55.16% had a comorbid alcohol use disorder. 40.7% of respondents with an alcohol use disorder had at least one current independent mood disorder, while more than 33% had one current anxiety disorder.

Conservatively, it is estimated that 80% of the 18.9 million Americans suffering from substance annually [12] do not get help for their addiction. In 2011 there were 13,720 substance abuse treatment facilities with 1.24 million patients in the United States [17]. 48% of the 1.6 million inmates in federal prison were serving time for a drug offence [13]. The American membership of Alcoholics Anonymous as of 2012 is estimated at 1.29 million [18]. There are no membership statistics on the considerably smaller drug-related fellowships including Narcotics Anonymous, Cocaine Anonymous, and Marijuana Anonymous.

Substance seeking behaviour has negative and devastating consequences for society [19]. The total costs for substance abuse (includes lost productivity, health and crime-related costs) in the United States is $416 billion annually; ($181 billion for illicit drugs [20] and $235 billion for alcohol [21]). Among the 19.6% of adults with AMI, 17.5% (8.0 million) met the criteria for substance dependence or abuse (illicit drugs or alcohol). 5.8% (10.9 million) of

---

higher estimate of 36.2% versus the NIMH estimate of 19.6%.

[5] Disorders may co-occur, thus the subsets of mood, anxiety, and depression do not add up to the total mental disorders.

the adult population who did not have mental illness in the past year also met criteria for a substance use disorder [22]. Opiate addiction is a global epidemic and is associated with many health consequences such as fatal overdose, infectious disease, and undesirable social consequences like, public disorder, crime and elevated health care costs [4].

Positive Emotions

Happy individuals are less likely to engage in harmful and unhealthy behaviours, including smoking, unhealthy eating, and abuse of drugs and alcohol [8]. Happy people are healthier, more optimistic, have higher self-esteem and self-control, and are more likely to increase their income in the future [8].

Genetics accounts for about 50% of variation in happiness, life circumstances account for 10%, and intentional activities are responsible for the remaining 40% [23, 24]. Improving Happiness through genetics is now possible. For example, Blum et al. [25] provided preliminary evidence that utilization of a customized dopamine agonist LG839 deoxyribonucleic acid (DNA) significantly increases happiness. Positive activity reinforcement is the domain of psychotherapy and treatment. Negative activity discouragement is the domain of psychotropic medication, psychotherapy and treatment.

Lyubomirsky et al. [23] examined 293 samples comprising over 275,000 participants from 225 papers studying happiness. Frequent positive affect as a hallmark of happiness has strong empirical support. The relative proportion of time that people felt positive relative to negative emotions was a good indicator of self-reports of happiness, whereas intensity of emotions was a weak indicator. People who reported high levels of happiness had predominantly positive affect (stronger positive emotions than negative) 80% or more of the time. Positive emotions might help people exert willpower and self-control over unhealthy urges and addictions.

Tugade et al. [26] determined that substantial empirical evidence supports the anecdotal wisdom that positive emotions are good for health. Those who used greater proportion of positive emotion words (versus negative emotion) showed greater positive morale and less depressed mood.

Fredrickson's broaden-and-build theory [27] suggests that multiple, discrete positive emotions are essential elements of optimal functioning. "Objective happiness" can best be measured by tracking (and later aggregating) people's momentary experiences of good and bad feelings. The overall balance of people's positive and negative emotions has been shown to predict their judgments of subjective well-being.

Dodge [28] concluded that higher depressive symptom scores significantly predicted and decreased likelihood of abstinence, after treatment center discharge, regardless of type of substance abuse, frequency of substance use, or length of stay in treatment. Dodge further stated that treatment approaches addressing the depressive symptoms are likely to enhance substance-abuse treatment outcomes.

Processing and Expressing Emotions

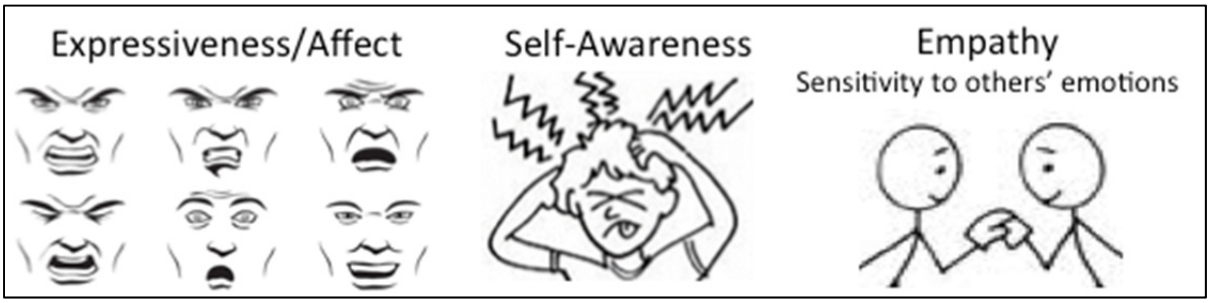


Figure 11 Affect, Self-Awareness, and Empathy

Scott [2] refers to "emotional muscle" as a necessary skill to cope with life's problems. Scott further elaborates: "For personal growth to occur, one must learn to process unpleasant feelings rather than running away from them by using drugs and alcohol. Addicts are very

inexperienced in processing feelings. As they come to understand their emotions, they develop the ability to tolerate them more and change their responses. Each time a client experiences a negative emotion without mood altering through drugs or alcohol he/she learned to take control a little more. The more clients do this, the stronger they become and the more emotional muscle they develop to cope with life's problems. Most chemically-dependent individuals cannot identify their feelings and do not know how to express[6] some effectively. Entire sessions are spent on each of the following emotions: anger, happiness, fear, depression, anxiety, and shame. Clients are asked to monitor their feelings by using a handout of a clock. At each hour on the clock they ask themselves "how am I feeling?" The goal is to become consciously aware of their internal state and how they are feeling change throughout the day and how feelings are related to other aspects of their lives."

Wurmser [29] coined the term "concretization" as the inability to identify and express emotions — a condition that often goes hand-in-hand with compulsive drug use. Wurmser further states: "it is as if these individuals have no language for their emotions of inner life; they are unable to find pleasure in every-day life because they lack the inner resources to create pleasure."

Opioid addicts on methadone maintenance (a synthetic drug used as a substitute, administered over a prolonged period of time as treatment for someone who is addicted to opioids such as heroin) appear to be less reactive to mood induction at times of peak plasma methadone concentration than non-addict controls; this suggests that methadone blunts both elative and depressive emotional reactivity [12]. There is evidence for a relationship between Substance Use Disorder and three biologically-based dimensions of affective temperament and behavior: negative affect (NA), positive affect (PA), and effortful control (EC). High NA, low EC, and both high and low PA were each found to play a role in conferring risk and maintaining substance use behaviours [10].

[6] "Affect" as defined by DSM-IV is a pattern of observable behaviors that is the expression of a subjectively experienced feeling state (emotion). Flat affect refers to a lack of outward expression of emotion that can be manifested by diminished facial, gestural, and vocal expression [15].

Introduction to Emotional Health

There has been a shift in mental health services from an emphasis on treatment focused on reducing symptoms based on health and disease, to a more holistic approach which takes into consideration quality of life [24]. Historically, the primary outcome goals for substance abuse treatment are harm reduction and cost effectiveness; with secondary outcomes including quality of life, and reduction of psychological symptoms [30]. It may be time to reconsider treatment priorities. There is evidence that happy individuals are less likely to engage in harmful and unhealthy behaviours, including abuse of drugs and alcohol [25]. In addition, treatment approaches addressing the depressive symptoms are likely to enhance substance-abuse treatment outcomes [28].

Quality of life is characterized by feelings of wellbeing, control and autonomy, a positive self-perception, a sense of belonging, participation in enjoyable and meaningful activity, and a positive view of the future. Emotional Health encompasses key aspects of quality of life including feelings, self-perception, and emotional connection with other people [2, 15, 16, 23, 26, 31, 32]. Measuring the capacity to live life with predominantly positive emotions (an average person is positive 80% of the time [23]); self-awareness of one's own emotions, emotional expressiveness (affect), and empathy for other people's emotions can provide insight into emotional health.

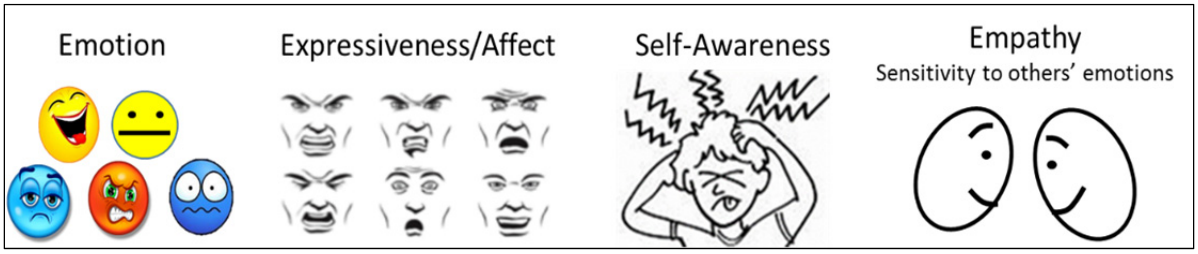


Figure 12 Components of Emotional Health

Emotion Set to Measure Quality of Life and Mood Disorders

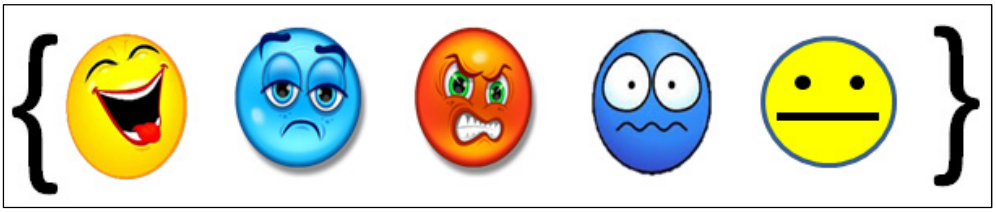


Figure 13 Emotion Set

Three sets of factors were considered in determining the set of five emotions (Neutral, Happy, Sad, Angry, and Anxious). (1) Depression, anger, and anxiety are associated with mental health disorders and substance abuse. (2) Happiness is an indicator of Quality of life [24]; (3) Human short term memory limits choices that a person can remembered to five (Miller proved that human short-term memory has a forward memory span of 7 ± 2 [33]); (4) The state of the art in automatic emotion classification is five emotions [34].

Most researchers agree [35] that emotions are short-term reactions to events or stimuli. Moods are not necessarily linked to an obvious cause or event [35]. They may influence actions and behaviour, but they do not interrupt ongoing behaviour and do not prepare immediate actions like emotions can [35]. The usual intensity of a mood is low to medium, and may last hours or even days or weeks, e.g. depression [35].

Many researchers have attempted to define the primary human emotions. In 1995, Goleman [36] grouped emotions into 8 primary emotions: (anger, sadness, fear, enjoyment, love, surprise, disgust, shame); but faith, encouragement, forgiveness, complacency, and boredom do not map neatly into these primary categories. In 1999, Ekman [37] proposed 15 primary emotions.

In 2011, four lead emotion researchers' theoretical models of basic emotions were compared by Tracy et al. [38] and were found to share Happiness, Sadness, Fear and Anger in common.

Table 1 Comparison of four lead authors' theoretical emotion models

IZARD	PANKSEPP & WATT	LEVENSON	EKMAN & CORDARO
Happiness	Play	Enjoyment	Happiness
Sadness	Panic/Grief	Sadness	Sadness
Fear	Fear	Fear	Fear
Anger	Rage	Anger	Anger
Disgust		Disgust	Disgust
Interest	Seeking		Contempt
	Lust		Surprise
	Care		

Happiness is an indicator of positive emotional health. Anger, depression and anxiety are key emotions in mood disorder and substance abuse. Fear and anxiety are overlapping, aversive, activated states centered on threat; clinical anxiety has been described as an ineffable and unpleasant feeling of foreboding [39].

The five emotions categories map well to Cognitive Behaviour Therapy (CBT) [40]as shown in Figure 14.

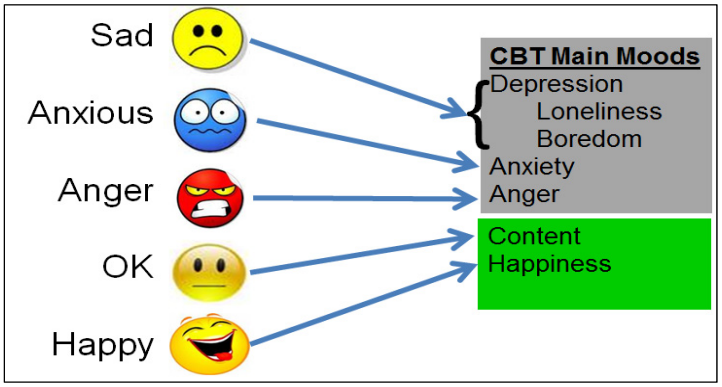


Figure 14 Five Emotions Mapped to CBT

Most emotions can be clustered into these five categories as shown in Figure 15.

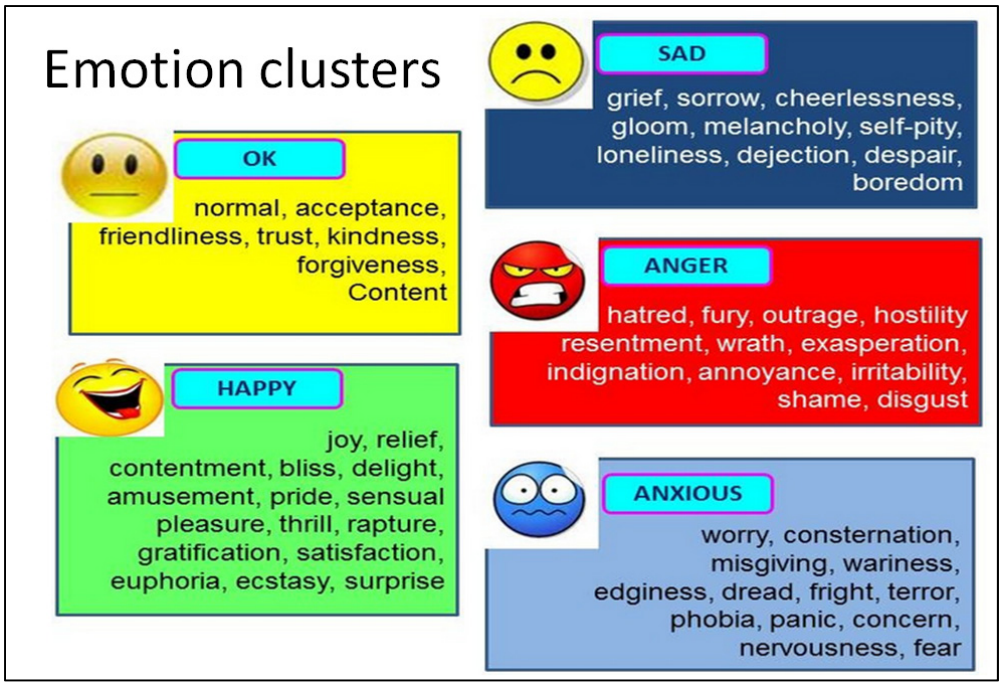


Figure 15 Emotion Cluster Mapping

State-of-the-art in Emotion Detection

What level of emotional truth accuracy is required for a viable commercial emotion classification system? In deterministic automatic classification problems like image recognition and speech recognition, the 80% accuracy benchmark [41] is a good threshold for viable commercialization of automatic classification systems. However, we discover in this thesis that emotional truth accuracy is not a black and white measurement. Determining the emotional truth for people with flat affect can be confusable. This confusability could provide insight on a person's expressiveness/affect.

Nwe et al. [42] conducted experiments to measure the performance of human classification of utterances into six classes: (Anger, Dislike, Fear, Joy, Sadness, Surprise). The average performance was 65.7%. The language of the utterances presented to the human subject was neither his mother tongue nor any other language that he has any knowledge to perceive linguistically; thus assuring only acoustic features were considered.

Steidl et al. [43] measured the performance of transcribers who listened to speech utterances and labeled the emotional content. In most cases, three out of five people could agree on the emotional content.

Five emotion classes is the current state-of-art in automatic emotion detection. The INTERSPEECH 2009 Emotion Challenge [34], was held in conjunction with INTERSPEECH 2009 in Brighton, UK, September 6-10. This challenge was the first open public evaluation of speech-based emotion recognition systems with strict comparability where all participants were using the same corpus. The German FAU Aibo Emotion Corpus of spontaneous, emotionally coloured speech of 51 children served as a basis. The results was 41.65% unweighted (UA) recall for the five-class problem (Angry, Emphatic, Neutral, Positive and Rest) by Kockmann et al. [34]. Dumouchel et al. [34] achieved 39.40% recall. The Dumouchel et al. algorithm [44], described in detail in chapter 5, is the basis for emotion detection in this thesis.

It is hypothesized that providing confidence and certainty scores of the classified emotion in speech will enable statistical analysis and allow professionals to monitor patients even when emotion classification is confusable and nondeterministic.

Sampling a Person's Experience over the Telephone

The Experience Sampling Method (ESM) is the best method to collect momentary emotional states in a person's natural environment [45]. The benefits of the Ecological Momentary Assessment method (EMA which includes ESM) are avoidance of recall and bias by collecting data on momentary states, realization of ecological validity by collecting data in the real-world, and achievement of temporal resolution enabling an analysis of dynamic processes over time [45].

Stone et al. [46] examined Patient-Reported-Outcome (PRO) ESM data collection and concluded PRO ESM places considerable demands on participants. Stone states that the

success of an ESM data collection depends upon participant compliance with the sampling protocol. Participants must record an ESM at least 20% of the time when requested to do so; otherwise the validity of the protocol is questionable. The problem of "hoarding" – where reports are collected and completed at a later date – must be avoided. Stone found that only 11% of pen-and-pencil diary studies are compliant; 89% of participants missed entries, or hoarded entries and bulk entered them later.

Hufford [47] also concluded that subject burden is a factor effecting compliance rates. Hufford states that at least six different aspects affect participant burden: Density of sampling (times per day); length of PRO assessments; the user interface of the reporting platform; the complexity of PRO assessments (i.e. the cognitive load, or effort, required to complete the assessments); duration of monitoring; and stability of the reporting platform. Researchers [46] have been known to improve compliance through extensive training of participants.

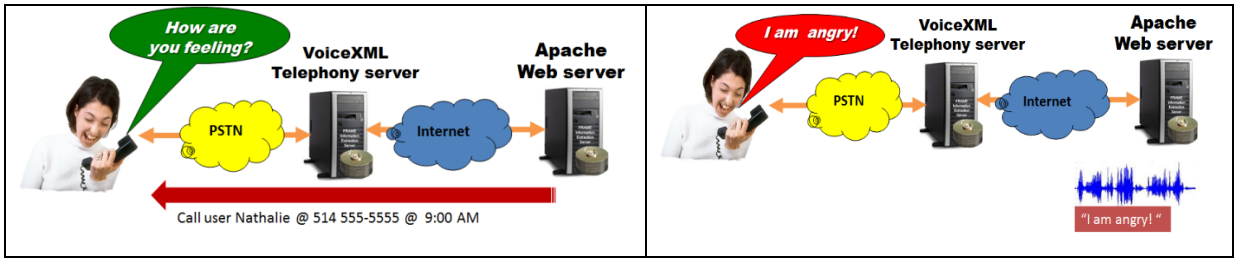


Figure 16 Emotional Health Sampling over the Telephone

Interactive Voice Response (IVR), as depicted in Figure 16, overcomes hoarding by time-sampling and improve compliance by allowing researchers to actively place outgoing telephone calls to participants in order to more dynamically sample their experience [47]. Rates of compliance in IVR sampling literature vary from 40% to 96% [47].

IVR ESM, avoids deployment costs associated with self-report systems on smartphones. There are 5 Billion mobile and phone users worldwide; only 1.5 Billion have access to a smartphone [48]. To deploy on all smartphones, you must build Apple iPhone Operating System (iOS), Android, Blackberry, and Symbian applications. Providing patients with a

smartphone is expensive; typical cost is $500 with reoccurring monthly telephony carrier charges of $30 or more. A severely afflicted addict may sell their smartphone for drugs.

Subject burden is addressed by limiting call duration to as little as 15 seconds, and providing an intuitive user interface design with no need for training. Calling subjects at times of their convenience further maximizes compliance rates. A patient is called and prompted with "How are you feeling?" The audio response (e.g. "I am angry!") is recorded in the cloud.

Emotional Health Statistical Analysis and Monitoring

Longitudinal regression **analysis** can provide evidence of the effectiveness of psychotropic medication, psychotherapy, and substance abuse rehabilitation. **Monitoring** and trend analysis can provide empirical insight and accelerate the interview process during monthly assessments by overburdened physicians and psychotherapists. Crisis **intervention** can be triggered on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions.

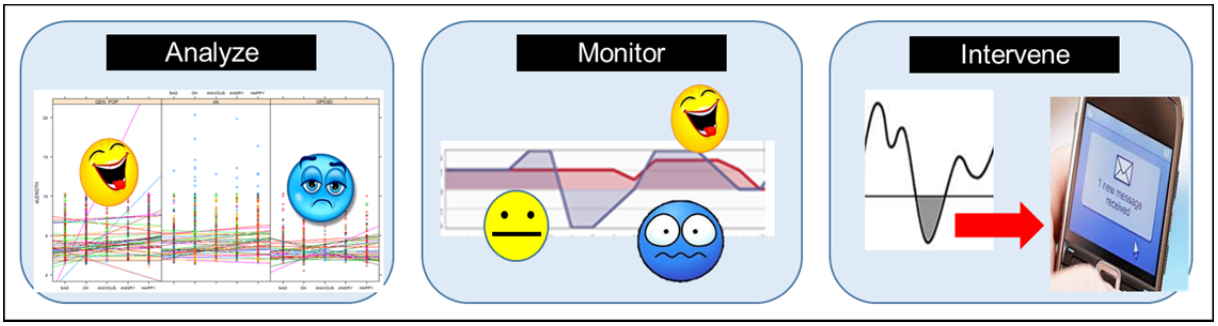


Figure 17 Analyze, Monitor, and Intervene

Patient Monitoring Benefits for Physicians and Psychotherapists

The average community mental health services' psychiatric follow-up is once a month for unstable patients and once every three months for stable patients [49]. A stable psychiatric outpatient session is estimated at 20 minutes (15 minute interview + 5 minutes for documentation). A session for an unstable outpatient is estimated at 40 minutes (30-minute interview + 10 minutes for documentation) [49]. Patient recall of events, feelings, and behaviours during the month(s) between sessions may not be reliable or objective, as detailed in APPENDIX A.

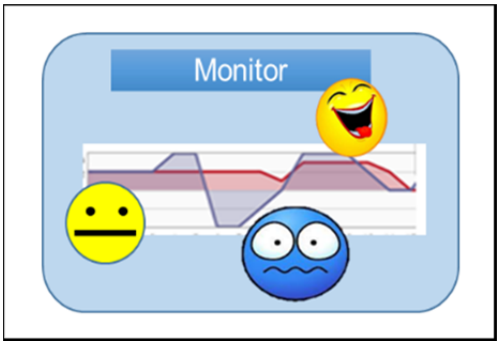


Figure 18 Monitor Patients' Emotional Health over Time

Empirical emotional health data and trend analysis using a toolkit that records emotion data should improve understanding of a patient between sessions. Emotional recordings can be played back to trigger recall of events and behaviours associated with peaks and valleys of longitudinal emotional state charts. Historical data can be reviewed for evidence of progress.

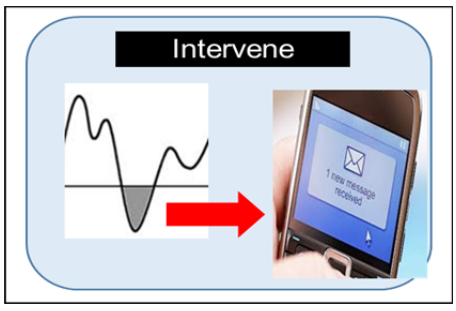


Figure 19 Detect Anomalies and Notify Professionals for Intervention

Crisis intervention can be triggered on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions (possible indication of relapse or an episode of mood disorder such as depression).

Figure 20 Monitor Emotional Health Effects of Medication

Automated daily emotional health measurements and trend analysis can also provide insight into the effectiveness of medication. For example, opioid addiction maintenance treatment with buprenorphine or Suboxone® [50] consists of three phases: (1) induction, (2) stabilization, and (3) maintenance. During the stabilization phase, patients are seen on a weekly basis. Once a stable dose is reached and toxicologic samples are free of illicit opioids, the physician may determine that less frequent visits are acceptable [50]. A patient may be in the maintenance phase indefinitely [50]. During the maintenance phase, attention must be focused on the psychosocial and family issues that have been identified during the course of treatment as contributing to a patient's addiction [50].

CBT [51], developed by Dr. Aaron T. Beck, is a form of psychotherapy in which the therapist and the client work together as a team to identify and solve problems. CBT is one of the few forms of psychotherapy that has been scientifically tested and found to be effective in hundreds of clinical trials for many different disorders [51]. In contrast to other forms of psychotherapy, CBT is usually more focused on the present, more time-limited, and more problem-solving oriented [51]. In addition, patients learn specific skills that they can use for the rest of their lives. These skills involve identifying distorted thinking, modifying beliefs, relating to others in different ways, and changing behaviours [51].

Figure 21 CBT Application Domains

Statistics on the use of CBT in psychotherapy are difficult to find. According to the Harley clinic in London [52], 45% of psychotherapy is CBT. As shown in Figure 21, common CBT interventions [40] include:

1) setting realistic goals and learning how to solve problems
2) learning how to manage stress and anxiety
3) identifying situations that are often avoided and gradually approaching feared situations
4) identifying and engaging in enjoyable activities such as social activities and exercise
5) identifying and challenging negative thoughts

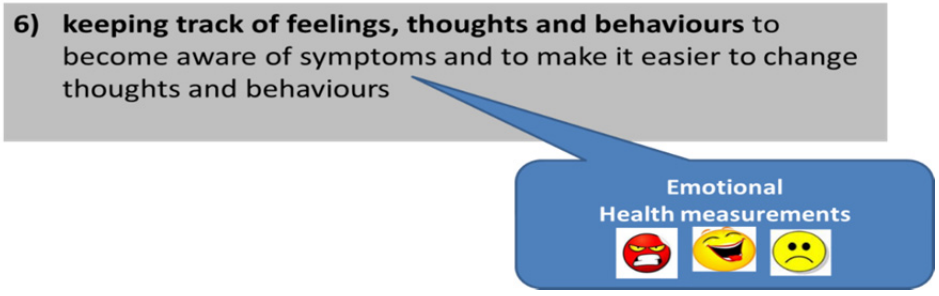


Figure 22 Emotional Health Measurement in CBT

The emotional health toolkit is well suited to automate step 6 of the CBT process.

Statistical Analysis Benefits for Evidence-Based Practice

"Clinical Efficacy is key to acceptance of technology in evidence-based practice" (Sonia Lupien, 2011).

Evidence-based assessment of drug effectiveness in clinical practice starts at the patient level by collecting relevant and validated data at the right time [53]. These data are collected and analysed at individual and group levels by health care professionals as a part of their daily work [53]. National agencies then make their assessments according to their mandate [53].

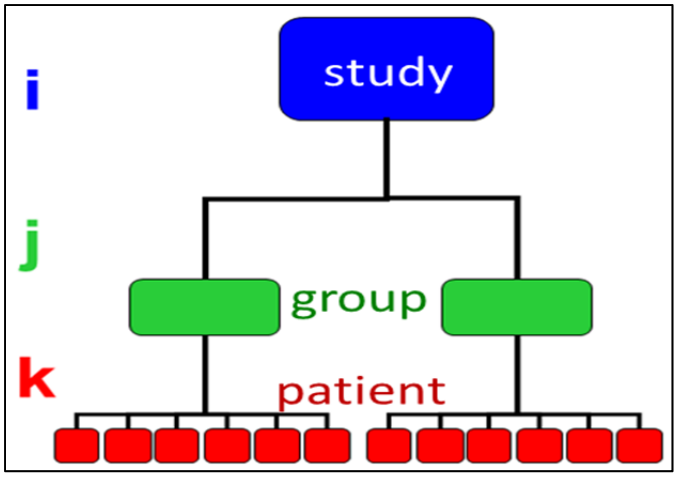


Figure 23 Hierarchical Study

Evidence-based treatment of drug and alcohol addiction through pharmaceuticals such as Methadone, Buprenorphine, Naltrexone for Heroin addiction; Naltrexone, Acamprosate, Disulfiram, Topiramate for Alcohol addiction; have proven effectiveness from a predominant standpoint of harm reduction [31]. Measuring emotional health and quality of life could enhance drug effectiveness assessment.

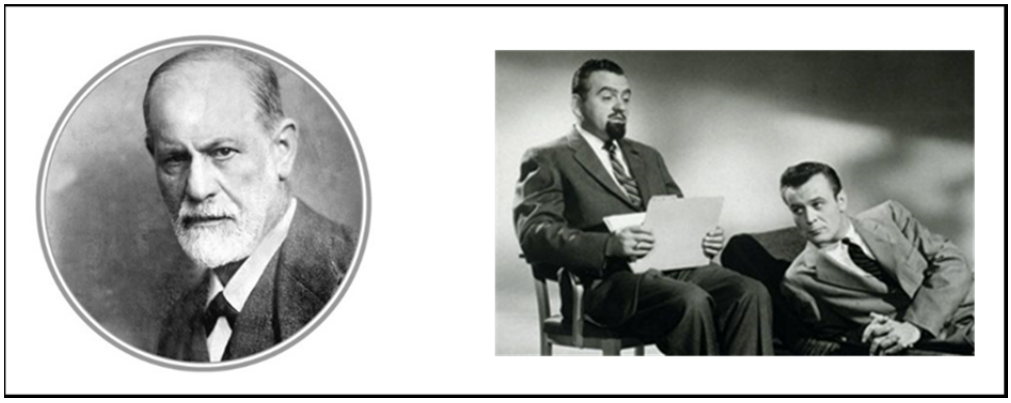


Figure 24 Psychotherapy

Decades of careful scientific research have documented the effectiveness of psychotherapy. Both qualitative and quantitative reviews of thousands of scientific studies have shown that about 75–80% of patients who enter psychotherapy show benefit [53]. Despite the importance of Evidence-Based Assessment (EBA), much available evidence suggests that clinicians are not engaged in assessment practices consistent with EBA, including what is arguably the core component of EBA: use of standardized assessment tools with research support for their reliability and validity [53]. Surveys of practicing psychologists suggest that the unstructured clinical interview is the most common, and often the only, assessment method used [54].

Carroll calls for the integration of empirically supported therapies into behavioural therapy as standard practice [55, 56]. Client outcomes are the bottom line for mental health services, like profit in business [55]. In mental health, productivity measures, such as the number of counselling sessions or the number of client served, tell us very little, if anything, about the effects of services on clients and their welfare [55]. For information to be useful data must be reliable and valid and collected at irregular and short intervals; it is important to measure progress towards substance abuse recovery [56]. Outcomes such as satisfaction, quality of life, and recovery are multifaceted and difficult to measure objectively [57].

Quality of life measurement as well as patient monitoring to detect isolation and relapse could help standardize substance abuse treatment outcome statistics. SAMHSA [57] and the

National Institute on Drug Abuse [31] emphasize that evidence-based treatment is needed to support treatment client outcomes. However measuring treatment and rehabilitation center (rehab) recovery effectiveness is a controversial subject. The effectiveness of treatment is not always measured, measurement methods are non-standardized, and statistics are scarce. Treatment facilities focus on abstinence as a key metric, but length of abstinence varies, and numbers are rarely publicized. A review of several outcome studies [58] indicate 13% - 36% of patients maintain continuous abstinence from drugs and alcohol for 6 months to 2 years after treatment. A study of crack cocaine addicts [59] with high attendance rates at a Behavioral Day Treatment measured 20% abstinence after 12 months. Survival rate statistical analysis [60] seems to be the predominant method in measuring recovery: $S(t) = \Pr(T > t)$, where T is length of abstinence, and t is a specified time.

CHAPTER 1

METHODOLOGY

1.1 Overview

The objective of this thesis is to develop and validate an evidence-based toolkit that captures a patient's emotional state, expressiveness/affect, self-awareness, and empathy on a telephone call, and then accurately measures and analyzes these indicators of Emotional Health. We need to establish accurate methods to measure emotional health from a telephone call; sound scientific statistical design of experiment and statistical methods to analyze trial data towards clinical efficacy; and methods to monitor a patient's emotional health and clinical protocol compliance over time.

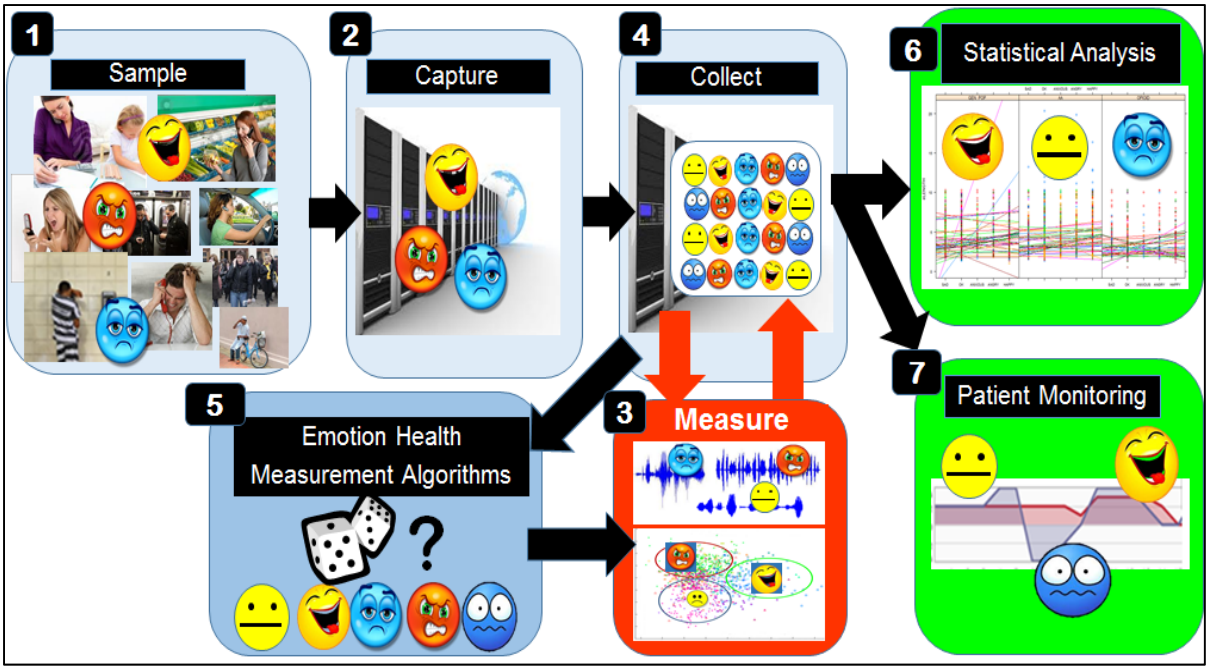


Figure 25 Emotional Health Toolkit Process Flow

Figure 25 and Table 2 describe the seven steps in the emotional health toolkit. This overview briefly describes the steps. Subsequent sections in this chapter provide more detail. Subsequent chapters provide still greater detail and results.

Table 2 Emotional Health Toolkit Process Steps

Step	Purpose	Description
1	Data Collection	Sample a patient's emotional state in their natural environment.
2	Data Collection	Capture the emotional state to a secure cloud database.
3	Measurement	Measure the patient's emotional health.
4	Data Collection	Collect emotional health measurements over time.
5	Emotional Health Algorithms	Develop emotion classification and emotional health algorithms. Improve accuracy as data is collected.
6	Statistical Analysis	Analyze patients to establish evidence-based practices.
7	Patient Monitoring	Monitor patients' emotional health over time.

Steps 1 (sample), 2 (capture), 4 (collect) are required for step 5 (emotional health measurement algorithm development). Once sufficient data has been collected to develop, train and test emotional health measurement algorithms, we can label the collected data and all subsequently captured data with the emotional health measurements by step 3 (measure). Scientifically designed experiments can be conducted on the collected data in step 6: (statistical analysis). Daily emotional health experience samples are sampled, captured, measured, and collected to enable step 7 (patient monitoring).

1.1.1 Step 5: Emotional Health Algorithm Development

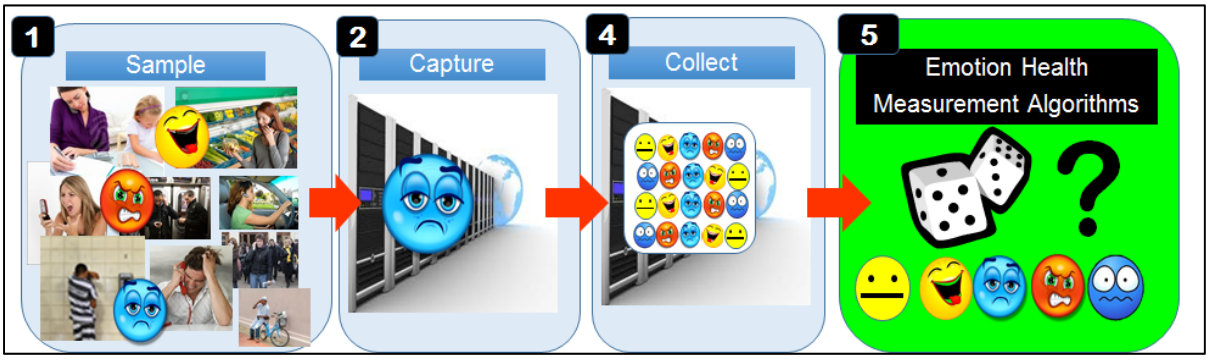


Figure 26 Emotion Health Algorithm Development Process Flow

Table 3 and Figure 26 describe the process to develop emotional health algorithms.

Table 3 Emotional Health Algorithm Development Process Steps

Step	Purpose	Description
1	Data Collection	Sample a patient's emotional state in their natural environment.
2	Data Collection	Capture the emotional state to a secure cloud database.
4	Data Collection	Collect emotional states to establish a corpus for algorithm development.
5	Emotional Health algorithm development	Develop, train and test emotional health measurement algorithms.

Once emotional health measurement algorithms have been developed, we can add the momentary emotional health measurements through step 3 (measure) to the data collection towards patient monitoring and emotional health analysis.

1.1.2 Step 6: Statistical Analysis

To demonstrate clinical efficacy, data collection must go one step further; sample patients' emotional health as a scientific experiment. Montgomery [61] describes scientific statistical design of experiments must follow three basic principles: randomization, replication, and blocking. Statistical methods require that observations (or errors) be independently distributed random variables (normal distribution). Replication means that the experiment can be independently repeated. Replication allows the estimation of experimental error as a basic unit for determining whether observed differences are really statistically different. If $\bar{y}$ is the sample mean and σ^2 is the variance of an individual observation, and there are n replicates, then the variance of a sample mean is $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$. Replication must also reflect the variability between trials and within trials. Blocking is a set of homogeneous experimental conditions (e.g. group size, trial duration). Selection of response variables should provide

useful information about the process under study. Factors should be included that may influence the performance of the process or system.

The analysis process steps are outlined in and the process flow is depicted in Figure 27 and described in Table 4.

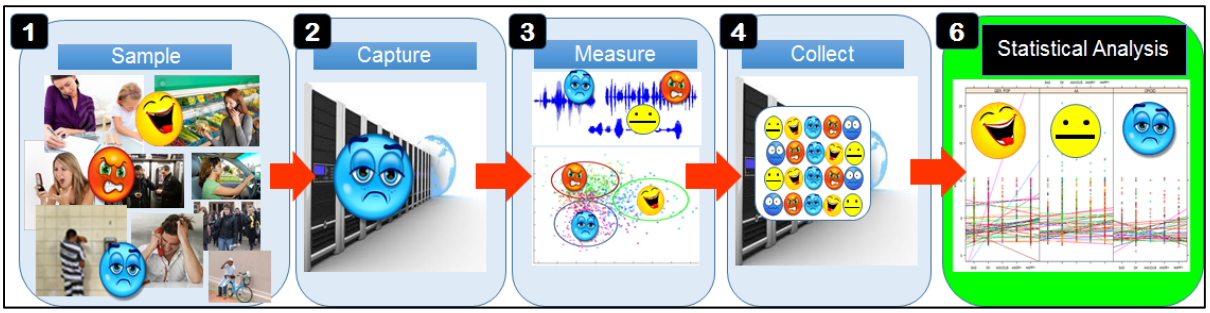


Figure 27 Emotional Health Statistical Analysis Process Flow

Table 4 Emotional Health Analysis Process Steps

Step	Purpose	Description
1	Data Collection	Sample a patient's emotional state in their natural environment.
2	Data Collection	Capture the emotional state to a secure cloud database.
3	Measure	Measure the patient's emotional state.
4	Data Collection	Collect emotional health emotional state and emotional health measurements over time.
7	Analyze	Analyze groups of patients to establish clinical efficacy.

1.1.3 Step 7: Patient Monitoring

Monitoring and trend analysis can provide empirical insight and accelerate the interview process during monthly assessments by overburdened physicians and psychotherapists. Crisis intervention can be triggered on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions.

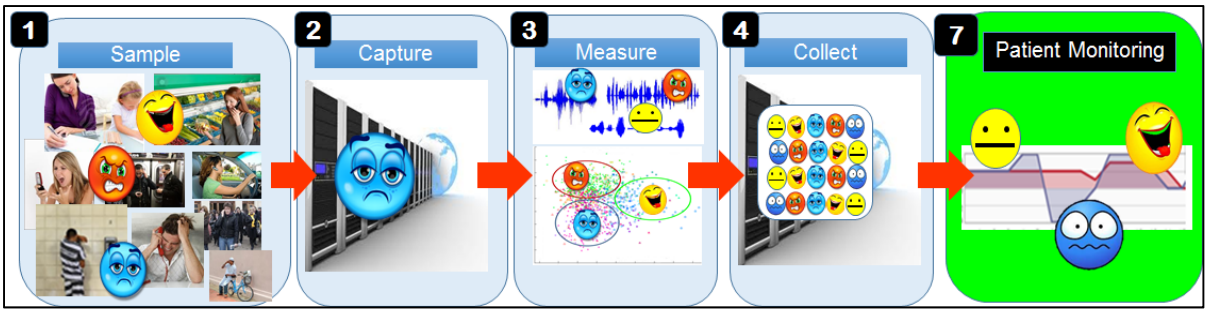


Figure 28 Emotional Health Patient Monitoring Process Flow

1.2 Subjects

This project originated from the department of Software Engineering and Information Technologies at École de Technologie Supérieure. A consent form, approved by the University of Quebec Ethics Committee (Canadian equivalent to the American Institutional Review Board (IRB) informed consent), was signed by each participant (see appendix K). We did not ask participants any information other than gender and language due to ethics committee restrictions.

Thirty-six Opioid Addicts undergoing Suboxone®[7] treatment (hereafter "SUBX") were randomly urine screened for the presence of SUBX. The urine screening revealed the presence of SUBX in 100% of these patients. Testing was performed by off-site by Quest Diagnostics (727 Washington St., Watertown, NY 13601, USA), and on-site at Occupational Medicine Associates of Northern New York, using the Proscreen drug test kit provided by US Diagnostics (2007, Bob Wallace Avenue, Huntsville, AL 35805, USA). Table 5 provides a breakdown of gender and language for the subjects.

[7] Suboxone® is a medication based on buprenorphine and naloxone. Buprenorphine is a pharmacological treatment for opioid addiction and is used in both maintenance and withdrawal programs. Naltrexone is an opioid antagonist that blocks the effects of heroin and most other opioids. There are two main modalities for the treatment of opioid addiction: pharmacotherapy and psychosocial therapy. Pharmacotherapies now available for opioid addiction include (1) agonist maintenance with methadone; (2) partial-agonist maintenance with buprenorphine or buprenorphine plus naloxone; (3) antagonist maintenance using naltrexone; and (4) the use of antiwithdrawal ("detoxification") agents (e.g., methadone, buprenorphine, and/or clonidine) for brief periods, and in tapering doses, to facilitate entry into drug-free or antagonist treatment. Psychosocial approaches (e.g., residential therapeutic communities), mutual-help programs (e.g., Narcotics Anonymous), and 12-Step- or abstinence-based treatment programs are important modalities in the treatment of addiction to heroin and other opioids, either as stand-alone interventions or in combination with pharmacotherapy [7].

Table 5 Gender and Language of the Research Participants

Group	Females	Males	English	French
General Population (GP)	29	15	25	19
AA	4	29	33	0
SUBX	23	13	36	0
Totals	56	57	94	19
	113		113	

1.3 Step 1 and 2: Emotional State Sample and Capture

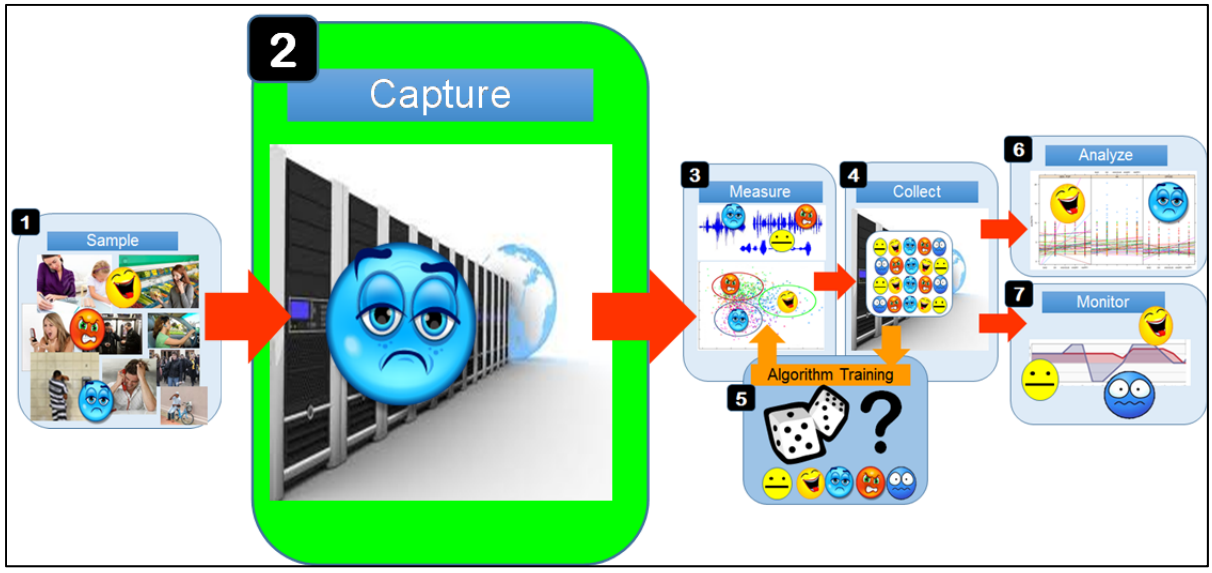


Figure 29 Step 2: Emotional Health Sample and Capture

Once a patient is registered, the system can start making and receiving daily telephone calls. Multiple momentary emotional states (experience samples) are collected over time for each patient. The i^{th} momentary emotional state for $patient_j$ (denoted ESM_{ij}) captured during a

call is recorded to a secure cloud database for subsequent emotion heath measurement and analysis[8].

1.3.1 Call Processing

Scheduled outbound dialing performs both pre-arranged and random time sampling over the Public Switched Telephone Network (PSTN) through the power of Call Control eXtended Markup Language[9] (CCXML) and Command Run On UNIX scheduler (CRON)[10]. The CRON daemon invokes a PHP: Hypertext Preprocessor[11] (PHP) script that checks the database for "ripened" call times. Once the call is successfully answered, the Voice eXtended Markup Language[12] (VoiceXML) application, with speech recognition and Dual Tone Multi-Frequency[13] (DTMF) recognition grammars coded in Grammar eXtended Markup Language[14] (GRXML) is invoked. Call status and user responses are captured to a database indexed by user and timestamp.

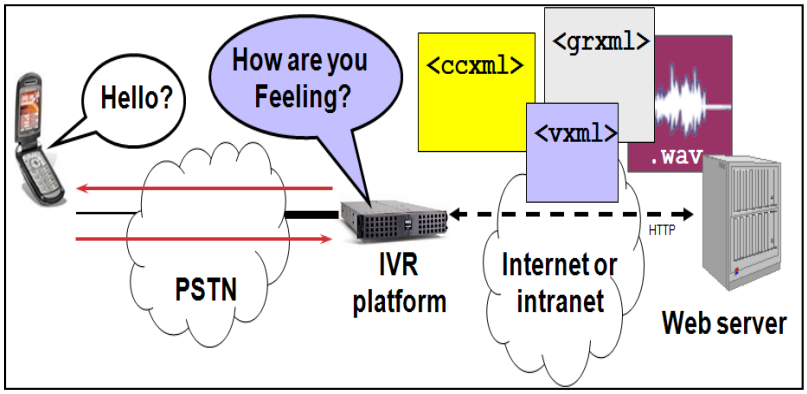


Figure 30 IVR Network Architecture

[8] See APPENDIX B for further details
[9] The World Wide Web Consortium (W3C) standard CCXML is an event driven markup language that is designed to provide telephony call control support for VoiceXML and has features such as answer machine detection, busy detect, and connection time out (no answer).
[10] CRON is the time-based job scheduler in Linux computer operating systems.
[11] PHP is an open-source server-side scripting language designed to produce dynamic Web pages
[12] VoiceXML is the W3C's standard XML format for specifying an interactive voice dialogues between a human and a computer.
[13] DTMF is used for telecommunication signaling over analog telephone lines and corresponds to the numbers on a telephone keypad.
[14] W3C's Speech Recognition Grammar Specification

CCXML automatic answer detection was experimented with to detect whether a human, voice mail, or fax, answered the call. The algorithm is based on a human's trait to answer with a short interrogative like "Hello?" versus a long voice/DTMF sequence from voice mail or a fax. The algorithm gets confused with excessive background noise that occurs in public places such as restaurants. The answer detection feature was therefore removed from the CCXML state machine. Instead, the call state is tracked over VoiceXML dialogue legs – logging dialogue progress. If the participant does not respond to the first question in the dialogue, it can be assumed the call was unsuccessful, and logged and processed as such. Call completion statistics are then mined from the call state records captured in the database.

1.3.2 VoiceXML Dialogue

The telephone dialogue to capture the emotional state is as follows:

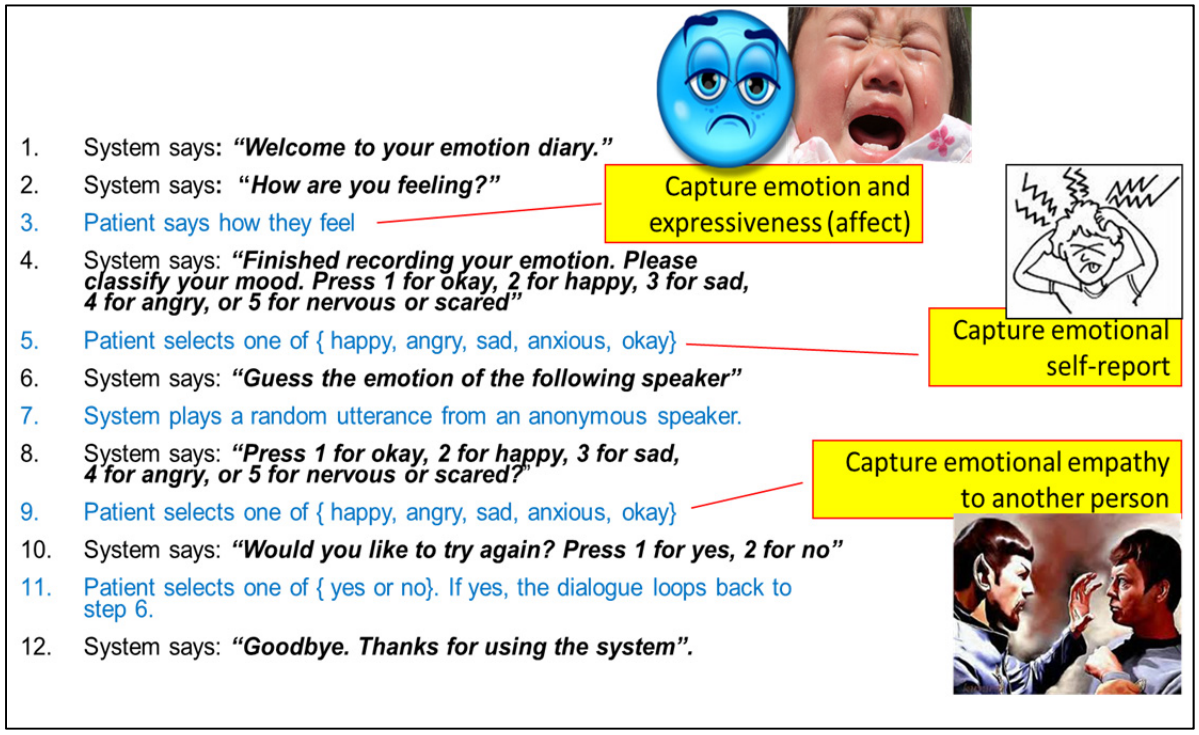


Figure 31 Emotional State (ESM_{ij}) Capture Telephone Dialogue

Prompts in the dialogue can be recorded by the patients' physician or psychologist to provide a sense of familiarity which may improve compliance; though this hypothesis is untested. This was done for Dr. Moehs' trial of his SUBX patients. The dialogue can also be modified to collect information such as a detailed diary or cravings.

Speech recognition was experimented to capture choice responses in steps 5, 9 and 11 in Figure 31. However, in noisy public places such as restaurants, the IVR system would not recognize the response, and have to reprompt "*I'm sorry, I did not understand, please repeat your choice*". Participants were asking to be removed from the trial as the annoyance factor was too high. As such, the dialogue was simplified with only keypad choices, which is robust in all noisy environments;

1.3.3 Emotional Experience Capture

In Figure 31, the recording containing the patient's emotion and expressiveness is captured in step 3 and designated X_{ij} (the i^{th} utterance from $patient_j$). The patient's self-reported emotion e_{ij}^{self} is captured in step 5. Zero or more empathic responses, e^{relate} are captured to the set E^{relate} in step 9. Equation 1.1 is the captured parameter set during the 15-second telephone call. Each parameter of ESM_{ij} described in Table 6.

$$ESM_{ij} = \{patient_j, c_{ij}^{calltype}, c_{ij}^{time}, c_{ij}^{state}, X_{ij}, e_{ij}^{self}, E^{relate}, c_{ij}^{duration}\} \quad (1.1)$$

Table 6 Experience Sample ESM_{ij} parameters

Parameter	Description
$patient_j$	The j^{th} patient in an experience sample collection trial
c_{ij}	The i^{th} telephone call for $patient_j$
$c_{ij}^{calltype}$	The call type of c_{ij} (inbound or outbound)
c_{ij}^{time}	The timestamp (date + time) of c_{ij}
$c_{ij}^{duration}$	The call duration in seconds of c_{ij}
c_{ij}^{state}	The call state of c_{ij} $c^{state} \in \{line\ busy, not\ answered\ , call\ complete\}$
If the call was complete:	
X_{ij}	The i^{th} speech recording for $patient_j$ captured during call c_{ij} in response to "How are you feeling?"
$e_{ij}^{self}(X_{ij})$ or e^{self}	The i^{th} emotional self-assessment for $patient_j$ of X_{ij} $e^{self} \in \{okay, happy, sad, angry, and\ anxious\}$ in response to "please classify your mood. Press 1 for okay, 2 for happy, 3 for sad, 4 for angry or 5 for nervous"
E^{relate}	Zero or more empathic responses e^{relate} to randomly selected anonymous emotionally charged recordings following the prompt "Guess the emotion of the following speaker. Press 1 for okay, 2 for happy, 3 for sad, 4 for angry or 5 for nervous", and looped with "would you like to try again?" The empathy $e^{empathy}$ is computed by comparing the empathic responses e^{relate} to the actual emotion of the anonymous recording.

1.3.3.1 Capture of Empathic Responses

The capture of E^{relate} on the i^{th} call c_{ij} from $patient_j$ (E_{ij}^{relate}) represents the set containing zero or more e^{relate} responses. The response e^{relate} is actually $e_{ijka}^{relate}(X_{ka})$; X_{ka} is the k^{th} utterance from an anonymous $patient_a$. $e_{ijka}^{relate}(X_{ka})$ is the empathic response on the i^{th} call c_{ij} from $patient_j$ after listening to the utterance X_{ka}. Both the anonymous patient and the utterance are randomly chosen; and never played twice to the same $patient_j$. To illustrate, suppose we have three patients in the system. There are 10 ESMs captured so far for $patient_2$ and 22 ESMs captured for $patient_2$ from previous calls in Table 7.

Table 7 Utterances for $Patient_2$ and $Patient_3$

i	$Patient_2$ $\mathbf{X_{i,2}}$
1	"I feel good"
2	"Not bad"
3	"I am mad!"
….	
10	"I am nervous"

i	$Patient_3$ $\mathbf{X_{i,3}}$
1	"Having a bad day"
2	"My dog ate my homework"
3	"I am happy!"
….	
22	"What a day!"

$Patient_1$ is currently on their 5th emotional experience capture call: $ESM_{5,1}$. The system prompts $patient_1$ with "Guess the emotion of the following speaker, and then randomly chooses an utterance from all other patients, which currently includes $patient_2$ and $patient_3$.

The system randomly chooses the utterance $X_{3,2}$, the 3rd utterance from $patient_2$. The IVR dialog then proceeds:

- Prompts $patient_1$ with "*Guess the emotion of the following speaker*"
- Plays utterance $X_{3,2}$ "*I am mad!*"
- Prompts *"Was the speaker happy, angry, sad, nervous or OK?"*
- $Patient_1$ selects "angry" (DTMF 3) on their telephone dial pad.
- The empathic response $e_{5,1,3,2}^{relate}(X_{3,2})$ = "angry" is captured in $ESM_{5,1}$ for $Patient_1$

1.4 Step 3: Emotional Health Measurement

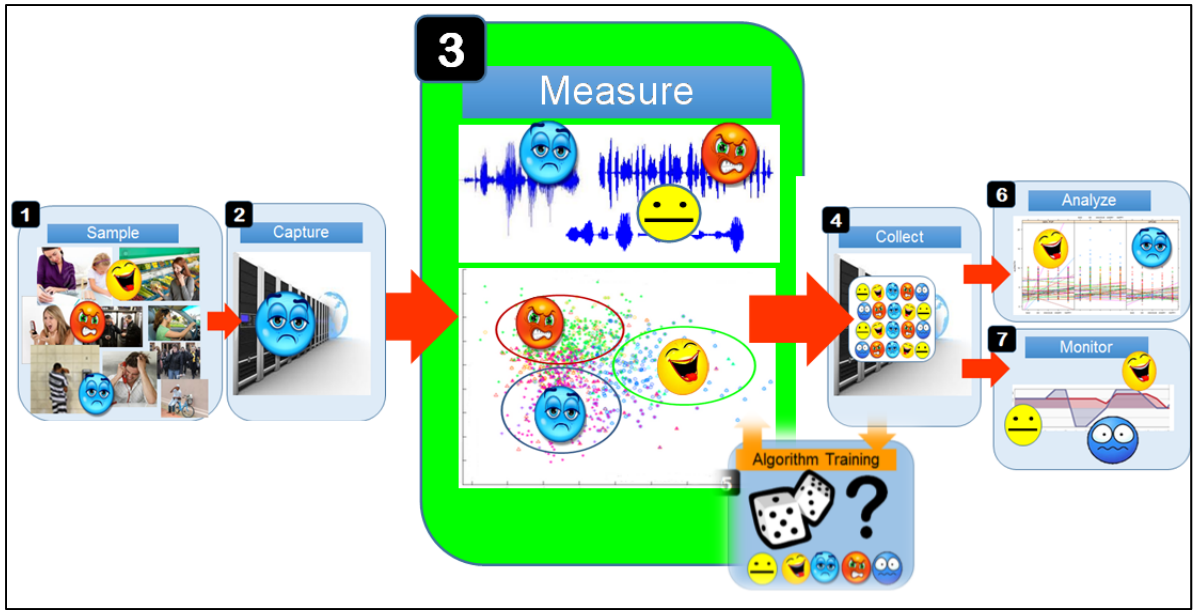


Figure 32 Step 3: Emotional Health Measurement

The next step is for speech processing algorithms to classify the emotion in the audio (hereafter the "emotional truth") and to calculate expressiveness and affect. Once the emotional truth is established, self-awareness and empathy can be measured by comparison with the patient's DTMF keypad choices.

1.4.1 Emotional Truth

Based on emotion classification algorithms which will be developed in Step 5 on page 35, we classify the emotional truth $e_{ij}^{truth}(X_{ij})$, of the emotional speech recording X_{ij}, from the emotion set (Neutral, Happy, Sad, Angry and Anxious).

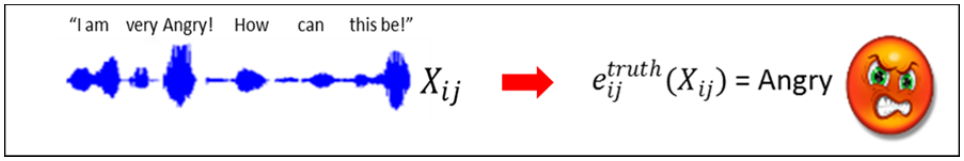


Figure 33 Classified Emotional Truth

1.4.2 Self-Awareness Emotional Concordance

Figure 34 Self-Awareness

Self-awareness $e_{ij}^{self-aware}(X_{ij})$ is computed as in equation 1.2 by comparing the emotional truth $e_{ij}^{truth}(X_{ij})$ of the recording to the patient's self-assessment, which is captured in response to the prompt "Are you happy, angry, sad, anxious or okay?"

$$\boldsymbol{if}\ \left(e_{ij}^{truth} == e_{ij}^{self}\right)\ \boldsymbol{then}\ e_{ij}^{self-aware} = TRUE$$
$$\boldsymbol{else}\ \ e_{ij}^{self-aware} = FALSE \qquad (1.2)$$

In Figure 35 a visual analysis of a random chosen participant early in data collection clearly indicates a discordance between self-assessment and assessment by others.

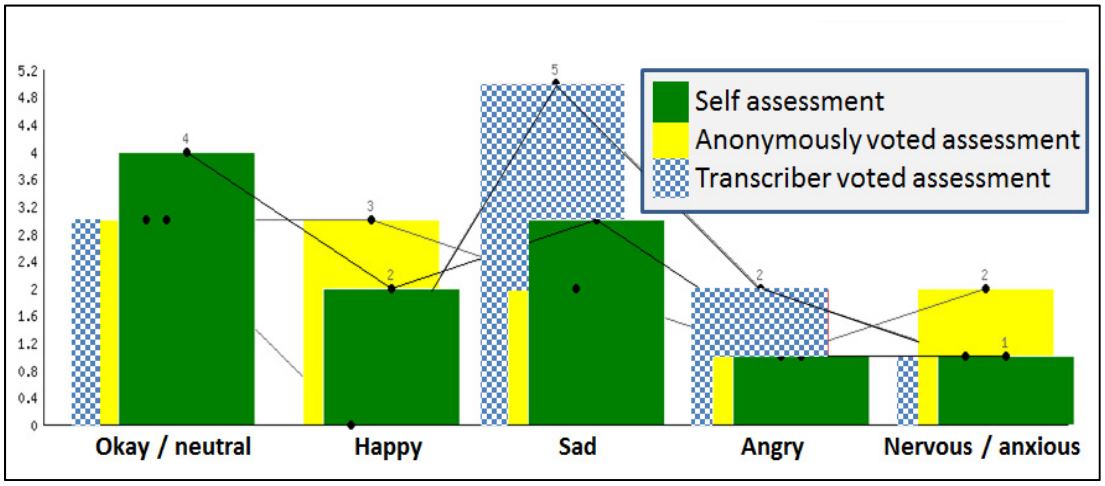


Figure 35 Concordance of Self-Assessment and Empathy Assessments

1.4.3 Empathy Concordance

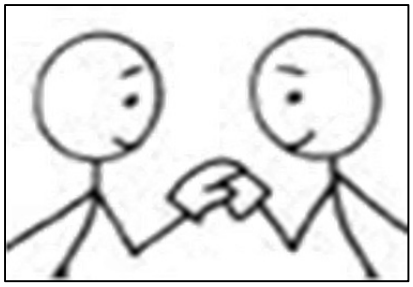


Figure 36 Empathy for another Human Being

Empathy $e_{ijka}^{empathy}(X_{ka})$ is calculated as in equation 1.3 by comparing the relate response $e_{ijka}^{relate}(X_{ka})$ of $patient_j$ of a randomly selected anonymous recording to the emotional truth $e_{ka}^{truth}(X_{ka})$ of that same anonymous recording. The anonymous recording is played following the prompt "Guess the emotion of the following speaker".

$$\begin{aligned} \boldsymbol{if}\ \left(e_{ijka}^{relate} == e_{ka}^{truth}\right)\ \boldsymbol{then}\ e_{ijka}^{empathy} &= TRUE \\ \boldsymbol{else}\ \ e_{ijka}^{empathy} &= FALSE \end{aligned} \qquad (\ 1.3\ )$$

If $e_{ijka}^{empathy}(X_{ka}) = TRUE$ then $patient_j$ correctly determined the k^{th} emotion of $speaker_a$ captured in ESM_{ka}. Otherwise, the patient could not determine the emotion correctly. Since we ask the patient "would you like to try again?" there can be many responses $e_{ijka}^{relate}(X_{ka})$ within the set E^{relate}. For each response, we calculate $e_{ijka}^{empathy}(X_{ka})$ that denotes the empathy of $patient_j$ during the i^{th} momentary emotional state collection towards $patient_l$ recording X_{ka} captured during the k^{th} momentary emotional state collection.

1.4.4 Emotional Expressiveness

Figure 37 Expressiveness/Affect

One measure of expressiveness/affect is the length of speech $length_{ij}(X_{ij})$. Length-of-speech is short for people who respond with phrases like "fine", "ok", "not bad"; and longer for people who are more expressive about how they feel (e.g. "having a great day! The sun is shining!"). "Affect" as defined by the Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM-IV) [26] is "a pattern of observable behaviours that is the expression of a subjectively experienced feeling state (emotion)." Flat affect refers to a lack of outward expression of emotion that can be manifested by diminished facial, gestural, and vocal expression. An additional measure of affect is the emotional-truth calculation's confidence score $e_{ij}^{confidence}(X_{ij})$. A high confidence score indicates high concordance amongst classifiers that can be interpreted that the emotion was easily recognized and the person has high affect. A low score indicates confusability among classifiers that may indicate flat affect or lack of emotion in the audio.

1.4.5 Emotional Experience Sample

We add calculated measurements to the captured parameters of ESM_{ij} from equation 1.1.

$$ESM_{ij} = \{captured\ parameters\} + \{measured\ parameters\} \quad (1.4)$$

Expanding equation 1.4 gives:

$$ESM_{ij} = \{patient_j, c_{ij}^{calltype}, c_{ij}^{time}, c_{ij}^{state}, X_{ij}, e_{ij}^{self}, E^{relate}, c_{ij}^{duration}\} + \{e_{ij}^{truth}, e_{ij}^{self-aware}, E^{empathy}, length_{ij}, e_{ij}^{confidence}\} \quad (1.5)$$

Table 8 summarizes the captured and measured emotional health parameters required to calculate emotional truth, self-awareness, empathy, and expressiveness/affect.

Table 8 Emotional Health Parameters

Emotional Health Requirement	**Description**	**Captured**	**Measured**
	Speech recording captured during call c_{ij}	X_{ij}	
Emotional truth	Classification algorithms determine the emotional truth in X_{ij}		$e_{ij}^{truth}(X_{ij})$
	Self-assessment during call c_{ij} of X_{ij}	$e_{ij}^{self}(X_{ij})$	
Self-Awareness	How aware is $patient_j$ of own emotion?		$e_{ij}^{self-aware}(X_{ij})$
	Emotional relate of a randomly selected recording X_{kl} from the k^{th} telephone call from anonymous $patient_l$	$\mathrm{e}_{\mathrm{ijkl}}^{\mathrm{relate}}(X_{kl})$	
Empathy	How well can $patient_j$ determine the emotion of another?		$e_{ijka}^{empathy}(X_{ka})$
Expressiveness /Affect	Longer speech is more expressive		$length_{ij}(X_{ij})$
	Less confusable speech has more affect		$e_{ij}^{confidence}(X_{ij})$

1.5 Step 4: Emotional Health Data Collection

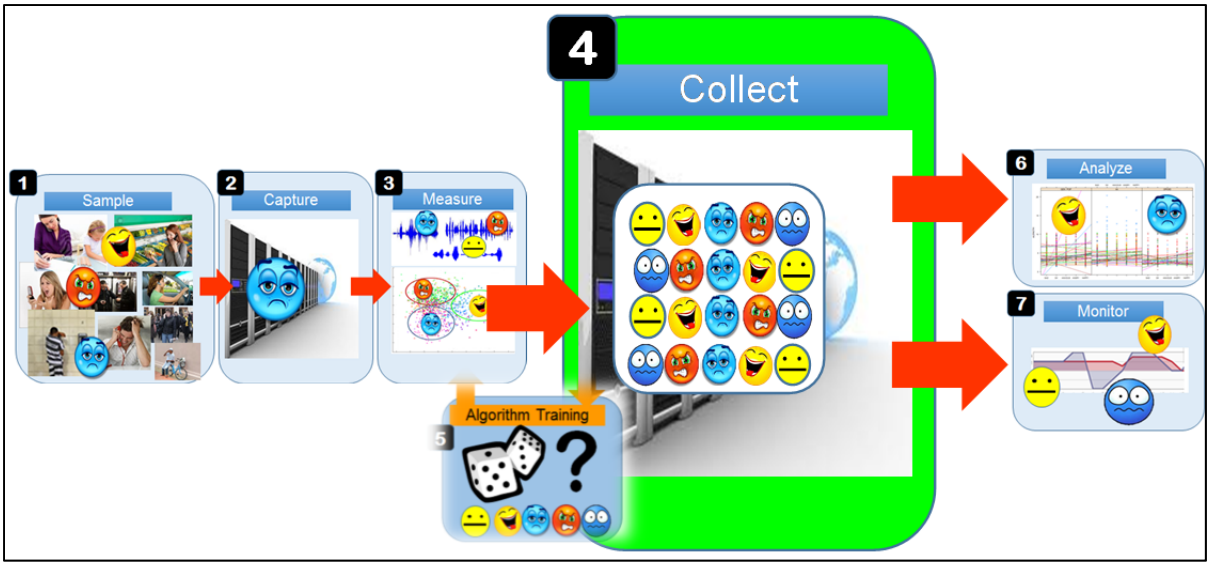


Figure 38 Step 4: Emotional Health Data Collection

Nineteen thousand five hundred and thirty-nine (19,539) telephone calls were made to the 129 trial participants. 8,376 of the 19,539 ESMs were successful ($c_{ij}^{state} = call\ completed$) resulting in 8,249 momentary emotional states, once bad recordings were pruned. The 129 trial participants included 113 subjects (detailed in section 2.1 on page 19), and 16 test and demonstration pseudo-participants. The multilevel data collected is summarized in equation 1.6. A post-trial survey in section 1.5.3 indicates low subject burden. Analysis of call rates in section 5.14 indicates high compliance rates.

$$\sum_{i,j} ESM_{ij} = \sum_{j\,(particpants)=129}^{1} \sum_{i\,(\frac{samples}{participant})}^{1} = 8{,}249 \qquad (1.6)$$

1.5.1 Data Warehousing

Three Linux-Apache-MySQL-PHP (LAMP) web servers were commissioned to handle the six trials: www.emotiondetect.com, www.emosub.com, and www.emotoolkit.com. MySQL auto-increment indexing was interleaved to ensure unique primary keys across the servers. A base Linux computer was used as the Data Warehouse for merging data collections from the three web servers. A daily CRON daemon invoked Practical Extraction and Reporting Language (Perl) and PHP scripts to remotely execute a My Structured Query Language; database management system (MySQL) dump on each server, and transfer the database dumps and speech recordings to the base computer. A Pentaho Spoon[15] shell script extracts ESMs from each server's database dump and merges them into a single MySQL Data Warehouse.

The Data Warehouse is used for emotion detection model training and testing, as well as statistical analysis using the R programming language [62].

[15] Pentaho's Spoon is an open source ETL or Extract, Transform and Load tool to load data from various formats into a MySQL server. http://wiki.pentaho.com/display/EAI/.01+Introduction+to+Spoon

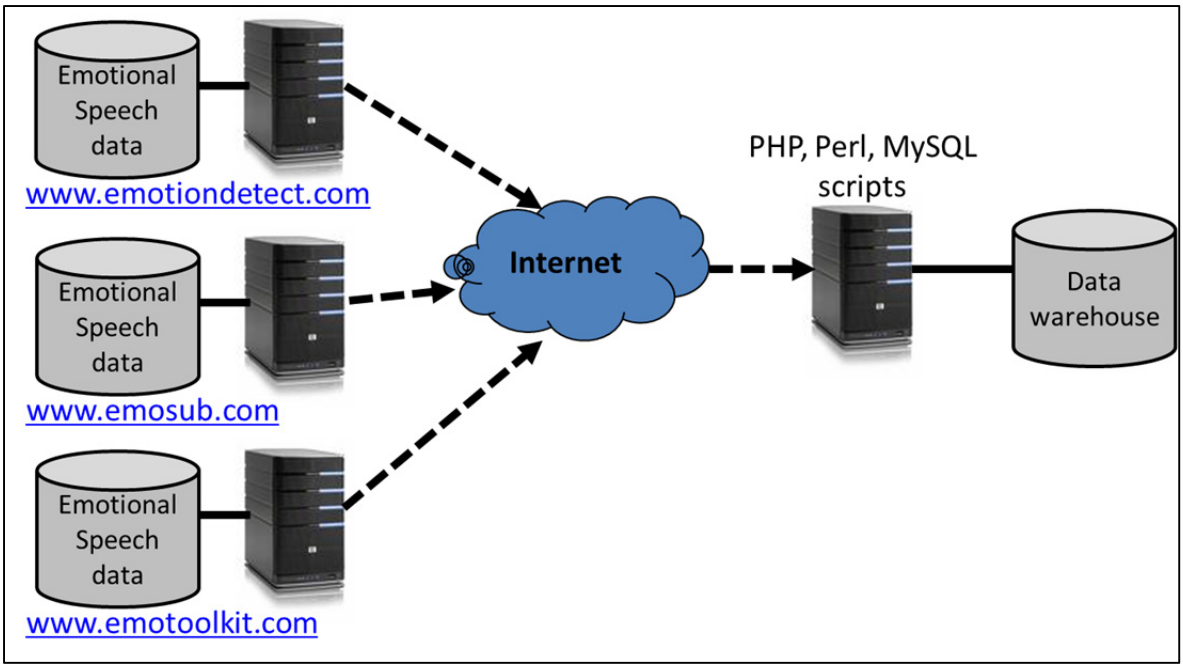


Figure 39 Extract, Transform and Load the Data Warehouse

1.5.2 Unsupervised Crowd-Sourced Corpus Labeling

Labeling speech is a time-consuming and labour-intensive process. Typically, as for the FAU Aibo Emotion corpus [63], raw audio recordings are first segmented manually into small, syntactically meaningful 'chunks' using syntactic-prosodic criteria that are subsequently labeled by paid professional transcribers. An unsupervised automatic crowd-source labeling method passed on a fused classifier of Majority Votes (MV) was devised. This method is described in detail in chapter 2.

1.5.3 Post-Trial Survey

A survey was conducted in September 2010 after the combined AA and English GP trial that started August 16th. There were 26 AA and 9 general population respondents. Summary of the responses are as follows:

- To determine the emotion of the anonymous recording:
  - 85% listened to how the speaker spoke (acoustics)

- 10% listened to the acoustics and the words spoken
- 5% listened to the words spoken.

- 56% of AA members and 74% of the General Population indicated Anger and Sadness as emotions most difficult to express.
- "Were 5 emotions were sufficient to report your emotion?" Average score was 3.3 on a scale of 1 to 5 (1=no; 5=sufficient).
- "Were you able to express your emotion in a telephone recording?" Average score was 4.1 on a scale of 1 to 5 (1=no; 5=full willingness).
- "Was 10 seconds sufficient time to express an emotion?" Average score was 4.0 on a scale of 1 to 5 (1=no; 5=plenty of time).

Positive Comments

"Good for self-evaluation." "It's an accountability tool for myself!" "It's a useful personal growth tool." "Good for introverted people; I can express myself without violating trust." "I got to track how I was feeling over time." "I liked it!" "Helpful!" "It's an Electronic sponsor!" "Interesting!" "Why is it over?" "Breaks out my day!" "I looked forward to call."

Negative Comments

"Need to prepare people!" "Therapy is relational / art /intuitive - rely more on instincts than technology." "Most people seem okay! 5 choices made it easy to pick okay." "People said ok allot!" "You need softer emotions like confusion, frustration, tired, ambivalent!" "Definitive emotions easy, but when I'm just cruising through day, it's tough to express vague feeling." "I am not always sure how I am feeling." "Add a preparation Question: 'are you in touch with your feelings?'" "There is not enough variety - too mechanical (I got used to it)" "Not a true reflection, I got used to it, A little pavlovian." "Instinct is to say 'ya I am OK' sometimes on the spot! Too much pressure to produce emotion! It came too fast! Like a machine gun. Not enough time to think!" "Happened too fast sometimes - misrated myself sometimes!" "I prefer to talk to human than computer." "Need option to listen to my emotion history!" "Trial should be at least 3 months! Too short time to develop a routine."

1.6 Step 5: Emotional Health Measurement Algorithms

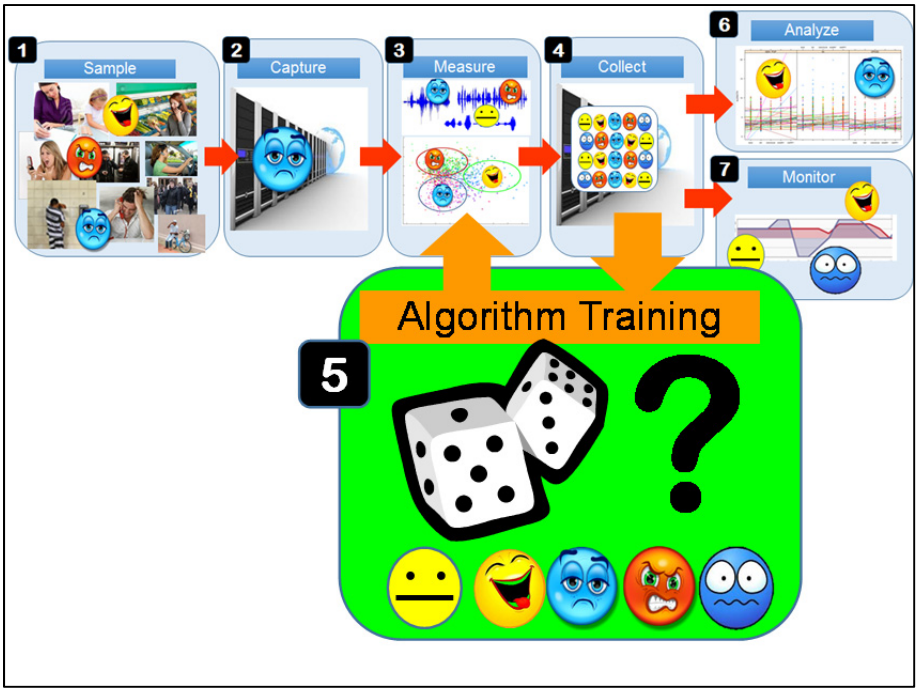


Figure 40 Step 5: Emotional Truth Algorithm Development and Training

Clinical efficacy for evidence-based practices requires reliable emotional health measurements. Successful commercialization of software depends on minimizing support costs. For the Emotional Health toolkit, minimizing support costs depends on automating emotional classification.

What is the actual emotion ("ground truth") in a speech recording? How confident are we of the emotional truth classification? Can the emotion classification be automatically calculated? What level of confidence is needed to trust the classification? How can we measure expressiveness, self-awareness, and empathy? These are the fundamental core research questions for the emotional health measurement aspect of this thesis.

1.6.1 Emotional Truth Calculation

The preferred approach to emotional truth determination is automatic real-time emotion detection in speech as this will provide instantaneous results. The core algorithm has been developed through a collaborative of scientists from the Massachusetts Institute of Technology (MIT) and the University of Québec [44]. This method is described in detail in chapter 3. The overall accuracy of the emotion detector is 42% (Neutral=64%, Happy=47%, Sad=29%, Angry=23%, Anxious=16%).

To improve accuracy, the crowd-sourced MV classifier described in chapter 2 was fused to the automatic emotion detector described in chapter 3. This pseudo real-time automated method is described in detail in chapter 4. The emotion is calculated in real-time with the emotion detector for monitoring and intervention purposes. The emotion is subsequently recalculated as votes become available, which may enforce the confidence in the emotion classification, or result in a new emotion classification.

1.7 Step 6: Emotional Health Statistical Analysis

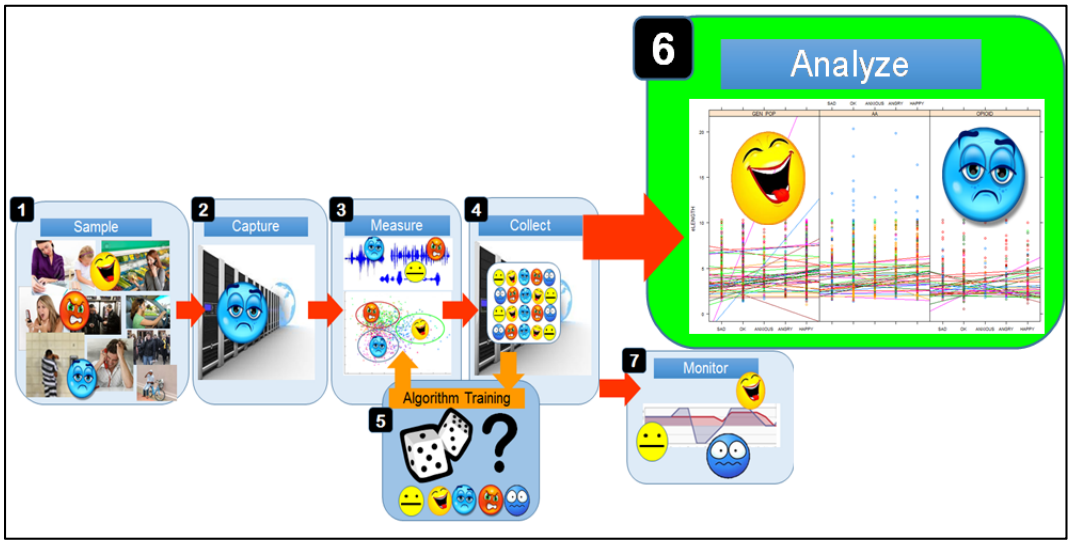


Figure 41 Step 6: Emotional Health Statistical Analysis

To demonstrate clinical efficacy, we analyze the data collected from the three groups (GP, AA members, and SUBX) using statistical analysis techniques.

1) Are there differences in emotional truth, self-assessment, self-awareness, and empathy across groups (GP, AA members, and SUBX)? Does gender (Male, Female) have an effect? Does language (English, French) have an effect? Do emotional health indicators vary with the time of day?

2) Does length of the response vary with emotion or group? Does the confidence score (confusability) of the emotional label vary with emotion or group?

3) Are there differences in call completion rates? Which group would be more likely to continue in data collections?

Significant results to these questions will provide evidence that capturing and measuring Emotional Health in speech can provide a mechanism (a) to assist CBT for therapists and patients to become aware of symptoms and make it easier to change thoughts and behaviors; (b) for Medical Doctors and Psychiatrists to measure the effectiveness of psychotropic medication; (c) for evidence of psychotherapy effectiveness in mental health and substance abuse treatment programs.

1.7.1 Statistical Regression Analysis

Statistical regression models specify how a set of dependent variables, functionally depend on another set of independent variables (predictor or explanatory variables). The functional relationship does not necessarily reflects a causal relationship - i.e. the independent variables do not necessarily describe the cause. Statistical models explain the value of the dependent variable by values of the independent variables [64].

1.7.2 P-values – Measuring Statistical Significance

The following is a simple description of p-values from StatsDirect [65]. The p-value or calculated probability is the estimated probability of rejecting the null hypothesis (H0) of a study question when that hypothesis is true. The null hypothesis is usually a hypothesis of "no difference" e.g. no difference between blood pressures in group A and group B. The alternative hypothesis (H1) is the opposite of the null hypothesis. For example, question is "is there a significant (not due to chance) difference in blood pressures between groups A and B if we give group A the test drug and group B a sugar pill?" and alternative hypothesis is " there is a difference in blood pressures between groups A and B if we give group A the test drug and group B a sugar pill". If the p-value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample gives reasonable evidence to support the alternative hypothesis. Conventionally, 5% (less than 1 in 20 chance of being wrong), 1% and 0.1% ($P < 0.05$, 0.01 and 0.001) significance levels are used. Most authors refer to statistically significant as $P < 0.05$ and a trend as $P < 0.1$.

1.7.3 Confidence Intervals

If the p-value of the difference in sample means indicates statistical significance, we still have to answer the question: how well does the observed pattern of sample means represent the underlying pattern of population means [66]? Confidence Intervals (CI) are designed to directly measure the population means [66]. The 95% confidence interval is invaluable in estimation because it provides a range of values within which the true population mean is likely to lie [67]. The 95% confidence interval is typically calculated from a single sample using the mean and Standard Error (SE) (derived from the SD, as described above). It is defined as follows: (sample mean – 1.96 SE) to (sample mean + 1.96 SE) [67].

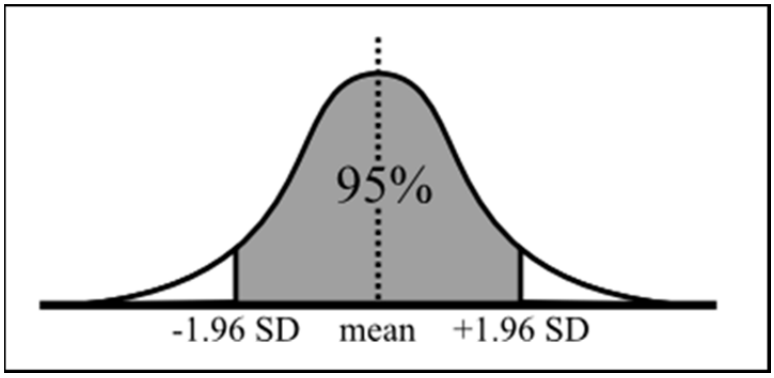


Figure 42 95% Confidence Interval

1.7.4 Pooled Ordinary Least Squares Regression Analysis

The emotional data collected is multilevel data grouped within participants. A common approach in social research with two-level data is to aggregate the micro-level data to the macro-level and perform Ordinary Least Squares (OLS) regression analysis. The standard OLS linear model:

$$Y_{ij} = \beta_{0j}+\beta_{1j}x_{ij} + \cdots + \beta_m x_{ij} + \varepsilon_{ij;}\ \varepsilon_{ij} \sim NID(0, \sigma^2); \qquad (\ 1.1\ )$$

has one random effect, the error term ε_{ij} [68]. OLS for multilevel analysis is known as "pooled analysis"; i.e. OLS analysis estimates β coefficients as a combination of β_w (within-group) and β_B (between-group). Pooled analysis suffers from: (a) Bias due to "causal" heterogeneity (i.e. the effect of X varies. there is no consideration for this interaction). (b) Incorrect standard errors due to clustering of observations (the number of observations differ per individual, and aggregated Y do not include corresponding weights).

In Figure 43 the black lines are the within regressions; the red line is the between regression; the green line is the pooled regression, which is a "compromise" between the within and between regression [69].

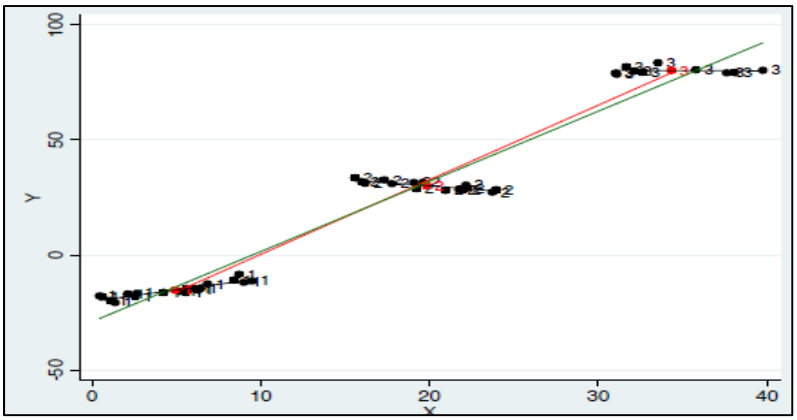


Figure 43 Illustration of Causal Heterogeneity

1.7.5 Multilevel Analysis

The best approach to analyze multilevel data is one that represents within-groups and between-groups relations within a single analysis [68]. The main statistical model of multilevel analysis is the Hierarchical Linear Model (HLM) which extends multiple linear regression analysis to a model that includes nested random coefficients [68]. All statistical analysis performed in this thesis is with the R programming language [62].

HLMs incorporate fixed-effects parameters, which apply to an entire population; and random effects, which apply to observational units. The random effects represent levels of variation in addition to the per-observation noise term that is incorporated in common statistical models such as linear regression models, generalized linear models and nonlinear regression models.[70, 71] [72].

The random-effect terms in HLMs are more appropriate for representing hierarchical clustered dependent data – arising, as in the emotion data collection, when ESMs are gathered over time from the same $participant_j$. Each $\bar{y}_j$ for $participant_j$ gives some information towards calculating the overall population average γ. Some $\bar{y}_j$ provide better information than others; i.e. $\bar{y}_j$ from a participant with more ESMs (number of observations n_j) will give better information than $\bar{y}_j$ from a participant with less observations. How do

you weigh $\bar{y}_j's$ from all participants in an optimal manner? ***Answer: weigh $\bar{y}_j$ by the inverse of its variance.*** ALL OBSERVATIONS then contribute to the analysis; including participants who have as few as one observation, since the observations are inversely weighted by within-group variance [72].

1.7.6 Comparison of HLM, OLS, and Average to Calculate the Population Mean

To illustrate the power of HLMs, we will calculate the log mean of the continuous outcome variable, length of emotional response length-of-speech $length_{ij}(X_{ij})$ (represented in R as eLENGTH), across all participants. One method would be to take the average of $\bar{y}'s$ (mean of means):

$$\bar{Y}_{MM} = \frac{\sum \bar{y}_{0j}}{M} = 1.274454 \qquad (1.7)$$

Alternatively, each $\bar{y}_j$ is weighted by the number of observations n_j for that participant and a weighted sum grand means is calculated:

$$\bar{Y}_{GM} = \frac{1}{\sum n_j} \sum (n_j \, \bar{y}_{0j}) = 1.37039 \qquad (1.8)$$

However, neither $\bar{Y}_{MM}$ or $\bar{Y}_{GM}$ take into consideration variable number of observations n_j per participant, as illustrated in Figure 44.

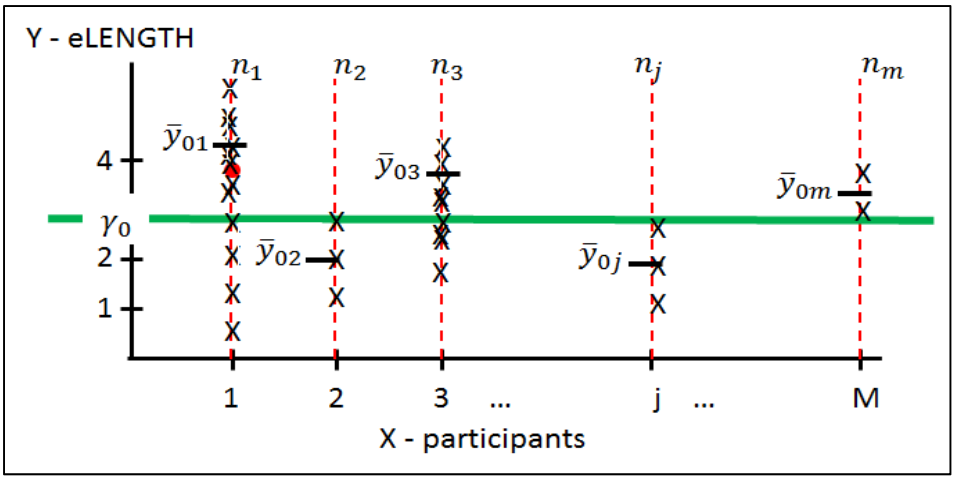


Figure 44 Mean of Unbalanced Multilevel Data

In the R programming language, the average calculation is performed as follows:

```
#mean of eLENGTH
ag1 <- aggregate(eLENGTH ~ p, IDr, mean)
ag2 <- aggregate(eLENGTH ~ p, IDr, length)
ag1$samples <- ag2$eLENGTH
> head(ag1)
  idUsers   eLENGTH   samples
1       1   2.058967     112
2       3   2.570243     176
3       5   0.935225      50
4       6   1.580125       2
5       7   1.435186      66
6       8   2.026909      55
> mean(ag1$eLENGTH)
[1] 1.274454
> sum(ag1$samples*ag1$eLENGTH)/sum(ag1$samples)
[1] 1.37039
```

Code Snippet 1 Aggregated Means Calculation in R

If we apply OLS to the mean of eLENGTH, we get a simple regression:

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij;} \qquad (1.9)$$

```
> # compute Linear model (OLS. Computes intercepts where B = Bw + Bb)
> secsLM <- lm(ag1$speechSECS ~ 1)
> summary(secsLM)

Call:
lm(formula = ag1$eLENGTH ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1945 -0.6412 -0.2001  0.3704  3.5256

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.27445    0.07616   16.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8915 on 136 degrees of freedom
```

Code Snippet 2 OLS Model Calculation in R

The Intercept (β_0), in this case, is equal to the expected value $E(Y) = 1.27445$. The standard error $= \frac{standard\ deviation}{\sqrt{sample\ size}} = 0.07616$. This model does not consider β_{within}, the within-group variance as depicted in Figure 45.

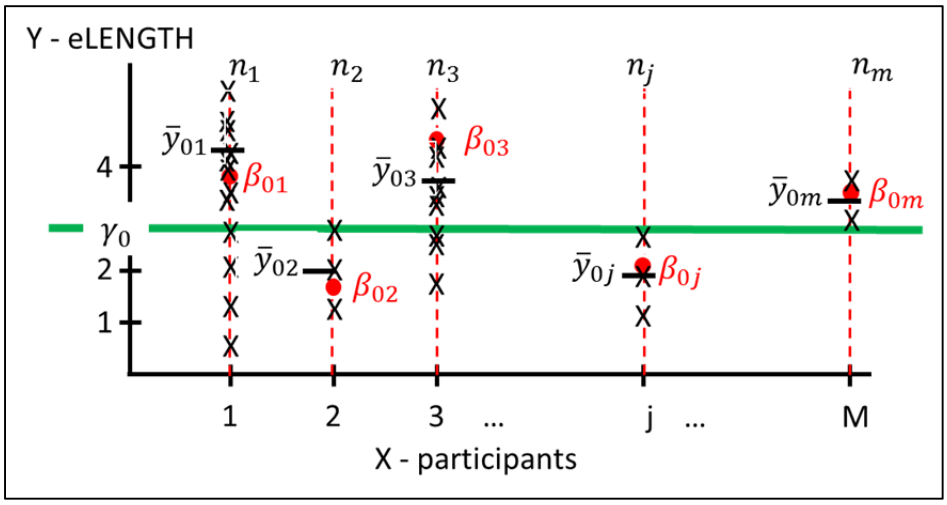


Figure 45 Level 1 Unbalanced Observation Clusters

To compute estimate γ_{00} with the HLM algorithm, each participant's β_{0j} is leveraged to calculate the overall mean (intercept) γ_{00} and the Standard Deviation $\sqrt{g_{00}}$. g_{00} is a 1x1 matrix, denoted G. Adding coefficients $\beta_{pj}, p > 0$ increases the size of G correspondingly.

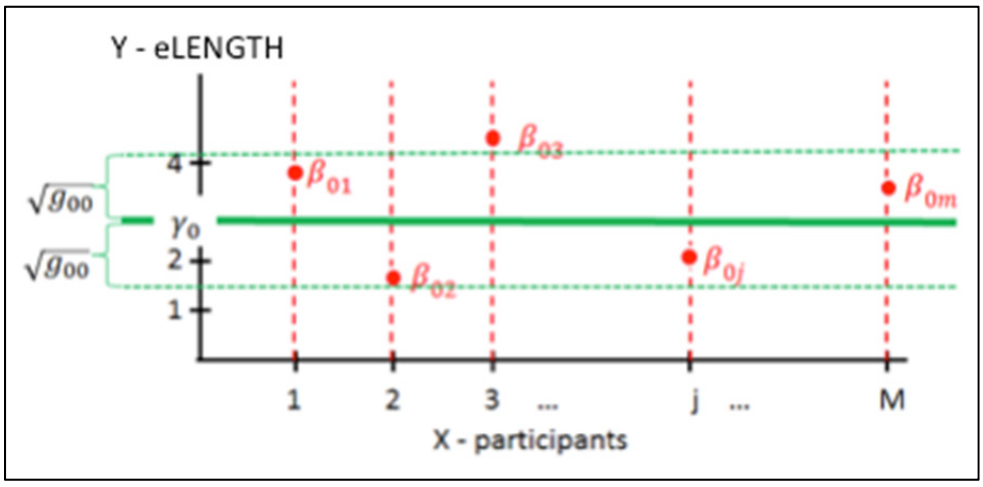


Figure 46 Overall Level 2 Mean Calculated from Cluster B's

Code Snippet 3 shows the R code to calculate the HLM model with R code. $E(Y) = \gamma_{00} + \mu_{0j} + \varepsilon_{ij;}$ $E(Y) = 1.265842$, with Standard Error of 0.068933448 which is a better estimate of $\beta_{between}$ than OLS. In addition, β_{within} is estimated as a random effect with SE 0.7400109.

```
> secsLME_ANOVA <- lme(speechSECS ~ 1, random = ~1|idUsers, data=IDr, na.action=na.omit)
> summary(secsLME_ANOVA)
Linear mixed-effects model fit by REML
 Data: IDr
       AIC      BIC    logLik
  29074.65 29095.64 -14534.32

Random effects:
 Formula: ~1 | idUsers
        (Intercept) Residual
StdDev:   0.7400109 1.428901

Fixed effects: speechSECS ~ 1
               Value  Std.Error   DF  t-value p-value
(Intercept) 1.265842 0.06893448 7956 18.36298       0

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.6712390 -0.5108697 -0.1679570  0.2132545  8.0058592

Number of Observations: 8093
Number of Groups: 137
> intervals(secsLME_ANOVA)
Approximate 95% confidence intervals

 Fixed effects:
               lower     est.    upper
(Intercept) 1.130713 1.265842 1.400972
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: idUsers
                    lower      est.     upper
sd((Intercept)) 0.6432936 0.7400109 0.8512693

 Within-group standard error:
   lower     est.    upper
1.406871 1.428901 1.451275
```

Code Snippet 3 HLM Calculation using lme() in R

The 95% confidence interval is 1.130713 to 1.400972. The length of speech as an indicator of expressiveness is analyzed further in section 5.13.1.

1.7.7 More on Hierarchical Linear Models

The following is based on a tutorial provided by Georges Monette at York University's Summer Program in Data Analysis (SPIDA) in 2012 [72]. The simplest example to move from OLS to a HLM is the one regression coefficient problem $Y_{ij} = \beta_{0j} + \varepsilon_{ij}$ where β_{0j} is the intercept (population average), and ε_{ij} is the residual effect of micro-unit i within macro-unit j. Applying HLM proceeds as follows:

$$\text{Level 1 model: } Y_{ij} = \boldsymbol{\beta_{0j}} + \varepsilon_{ij} \quad (1.10)$$

$$\text{Level 2 model: } \boldsymbol{\beta_{0j}} = \gamma_{00} + U_{0j} \quad (1.11)$$

Combining equations 1.10 and 1.11 produces the mixed-model HLM:

$$Y_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij} \quad (1.12)$$

where γ_{00} is the fixed effect, and $U_{0j} + \varepsilon_{ij}$ are the random effects. The overall variance for $participant_j$ is:

$$Var(\beta_{0j} - \gamma_0) = g_{00} \quad (1.13)$$

But this does not tell us how to apply the participant's variance $\frac{\sigma^2}{n}$ as an estimator of $\gamma_0 = Var(\bar{y}_{0j} - \gamma_0)$. We need to calculate:

$$Var(\bar{y}_{0j} - \gamma_0) = Var(\bar{y}_{0j} - \beta_{0j} + \beta_{0j} - \gamma_0) \quad (1.14)$$

$$Var(\bar{y}_{0j} - \gamma_0) = Var(\bar{y}_{0j} - \beta_{0j}) + Var(\beta_{0j} - \gamma_0) \quad (1.15)$$

Substituting the variance estimator $\frac{\sigma^2}{n}$ and equation 1.13 into equation 1.15:

$$Var(\bar{y}_{0j} - \gamma_0) = \frac{\sigma^2}{n_j} + g_{00} \quad (1.16)$$

The overall population average is:

$$\bar{Y}_{Mixed\ Model} = \left[\sum \frac{1}{\frac{\sigma^2}{n_j} + g_{00}}\right]^{-1} \sum \frac{1}{\frac{\sigma^2}{n_j} + g_{00}} \bar{y}_{0j} \quad (1.17)$$

$\bar{Y}_{Mixed\ Model}$ is an optimized estimator of overall mean that takes into account, in an optimal way, information contained in each participant's mean. Weight contribution from each participant depends on n_j and g_{00}. Thus a participant with 100 samples will contribute more than a participant with 1 sample, but the 1 sample cluster can still be leveraged to improve the overall estimate.

Complexity increases as coefficients are added. A one-level, two-regression-coefficient OLS model is formulated as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \qquad (1.18)$$

The intercepts β_{0j} as well as the regression coefficients β_{1j} are group-dependent. To move to a mixed-effect model, the group-dependent coefficients can be divided into an average coefficient in equation 1.19 and the group-dependent deviation in 1.20.

$$\beta_{0j} = \gamma_{00} + U_{0j} \qquad (1.19)$$

$$\beta_{1j} = \gamma_{10} + U_{1j} \qquad (1.20)$$

Substituting equations 1.19 and 1.20 into equation 1.18 gives:

$$Y_{ij} = \gamma_{00} + U_{0j} + \gamma_{10} x_{ij} + U_{1j}\, x_{ij} \beta_{1j} x_{ij} + \varepsilon_{ij;} \qquad (1.21)$$

Grouping fixed and random effects of equation 1.21:

$$Y_{ij} = \overbrace{\gamma_{00} + \gamma_{10} x_{ij}}^{\text{fixed effect}} + \overbrace{U_{0j} + U_{1j}\, x_{ij} \beta_{1j} x_{ij} + \varepsilon_{ij;}}^{\text{random effect}} \qquad (1.22)$$

1.7.8 Assumptions of the Hierarchical Linear Model

Like all statistical models, the HLM is based on assumptions [68]. If these assumptions are not satisfied, the procedures for estimating and testing coefficients may be invalid. The assumptions are [68]:

1. Are the response variables are independent? This is an experimental design issue adequately addressed at the beginning of this chapter.
2. Are the residuals normally distributed? We test for this.
3. Do the residuals have constant variance (Homoscedasticity)? We test for this.

For example, in the one regression coefficient model $Y_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij}$; the level-two residual U_{0j} and the level-one residual ε_{ij} must both be homoscedastic and normally distributed.

1.7.9 Homoscedasticity and Heteroscedasticity

According to Zuur et al. [73], heteroscedasticity (a violation of homoscedasticity) happens if the spread of data is not the same at each X value, and this can be checked by comparing the spread of residuals for different X values. The residuals are pooled and plotted against fitted values. The spread should be roughly the same across the range of fitted values.

According to the National Institute of Standards and Technology (NIST) [74], residuals are estimates of experimental error obtained by subtracting the observed responses from the predicted responses. Residuals can be thought of as elements of variation unexplained by the fitted model. Since this is a form of error, the same general assumptions apply to the group of residuals that we typically use for errors in general: one expects them to be (roughly) normal and (approximately) independently distributed with a mean of 0 and some constant variance.

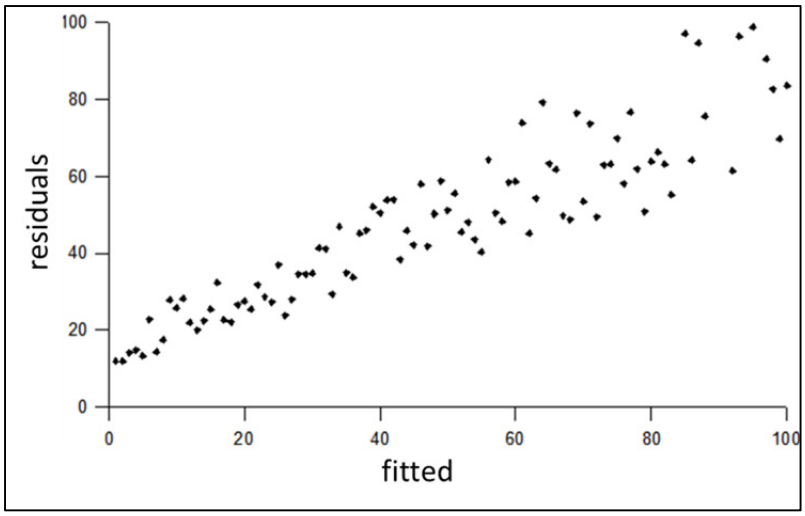


Figure 47 Heteroscedastic Residuals

Figure 47 depicts heteroscedastic residuals. According to Bradley [75] heteroscedasticity is apparent if the residuals seem to increase or decrease in average magnitude with the fitted values, it is an indication that the variance of the residuals is not constant. The points in the plot lie on a curve around zero, rather than fluctuating randomly. A few points in the plot lie a long way from the rest of the points.

As examined in detail in appendix J, the length-of-speech $length_{ij}(X_{ij})$ is a non-normal Gamma distribution. Plotting the residuals of length-of-speech produces the left chart in Figure 48 which is heteroscedastic. Log-Normalizing length-of-speech produces the right chart in Figure 48, which is much closer to homoscedasticity.

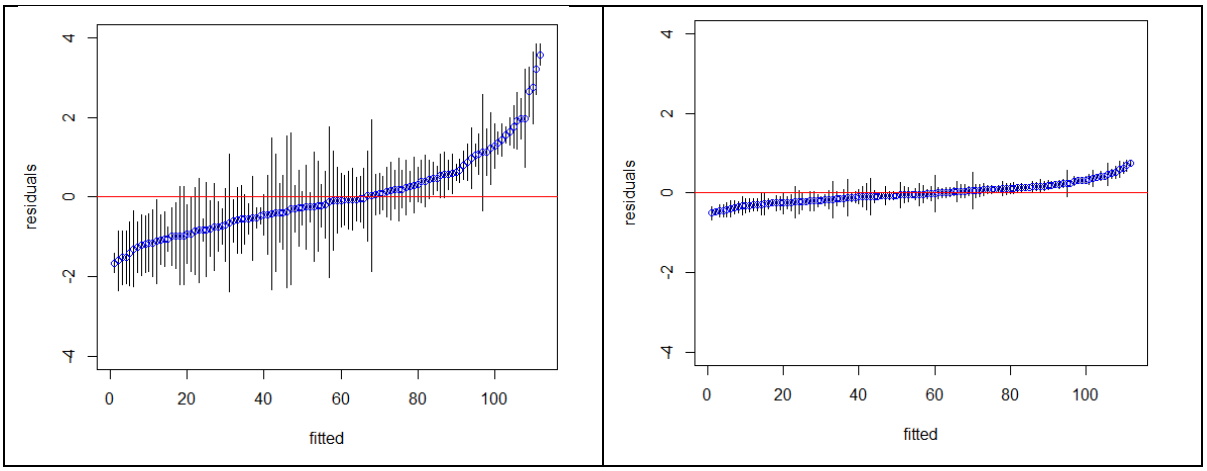


Figure 48 Residuals of Length-of-Speech and Log-Normalized Length-of-Speech

1.7.10 Normality

The Shapiro-Wilk test is one of the most powerful normality tests, especially for small samples [74]. Normality is tested by matching two alternative variance estimates: a non-parametric estimator got by a linear combination of ordered sample values and the usual parametric estimator [74]. The R command to perform Shapiro-Wilk test is shapiro.test() [62] and it supplies W statistic and the p-value. Small values of W are evidence of departure from normality and percentage points for the W statistic. If the p-value is higher than significance levels (0.05), then we accept null hypothesis that is sample data belong from a normal distribution.

In Code Snippet 4 Shapiro Wilk Normality Test in R Code Snippet 4, we generate a normally distributed random vector of 100 with mean of 5, and standard deviation of 2. We run the shapiro.test(), which gives a high W statistic and a p-value > 0.05, indicating a normal distribution.

```
shapiro.test(rnorm(100, mean = 5, sd = 3))
        Shapiro-Wilk normality test
data:  rnorm(100, mean = 5, sd = 3)
W = 0.9935, p-value = 0.9154
```

Code Snippet 4 Shapiro Wilk Normality Test in R

1.7.11 Distribution Description – Skewness and Kurtosis

If Shapiro-Wilk normality test should fail, we need to explore the distribution. A powerful tool to visually explore distributions in R is descdist() [76] that computes descriptive parameters of a distribution and provides a Cullen and Frey Skewness-Kurtosis plot to visualize the type of distribution. On this plot, values for common distributions are also displayed as a tools to help the choice of distributions to fit to data. For some distributions (normal, uniform, logistic, exponential for example), there is only one possible value for the skewness and the kurtosis (for a normal distribution for example, skewness = 0 and kurtosis = 3), and the distribution is thus represented by a point on the plot. For other distributions, areas of possible values are represented, consisting in lines (gamma and lognormal distributions for example), or larger areas (beta distribution for example).

Skewness is a measure of symmetry, or more precisely, the lack of symmetry [74]. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point [74]. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution [74]. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails; Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak [74].

As an example, in Code Snippet 6 we first generate a Poisson distribution of 100 points using the rpois() [62] . Next we generate a histogram of the distribution in Figure 49 using hist() [62].

```
> x.poiss <- rpois(100,lambda=2)
> hist(x.poiss,main="Histogram of observed data")
```

Code Snippet 5 Poisson distribution in R

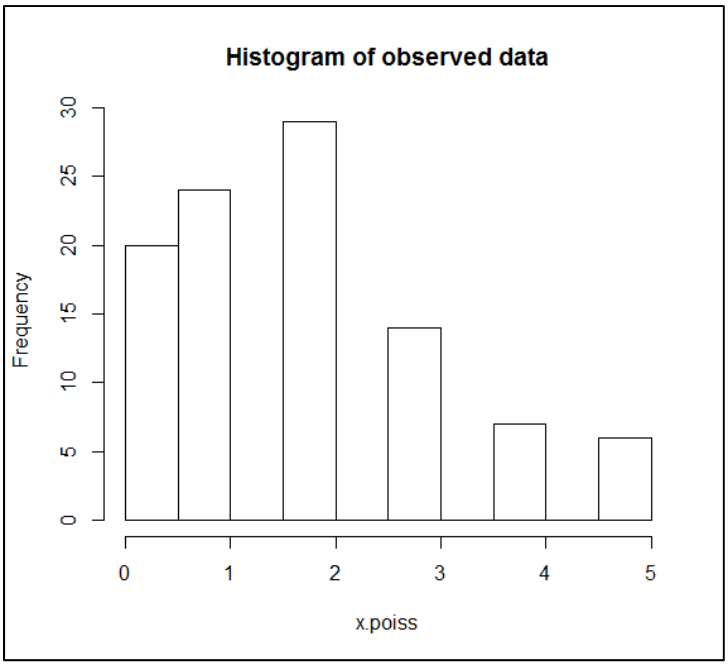


Figure 49 Histogram of Poisson distribution

In Code Snippet 6 the function descdist() [76] is executed to generate Figure 50 . In order to take into account the uncertainty of the estimated values of kurtosis and skewness from data, the data set is bootstrapped by fixing the argument boot to an integer above 10. Boot values of skewness and kurtosis corresponding to the boot bootstrap samples are then computed and reported in blue color on the skewness-kurtosis plot [76].

```
> descdist(x.poiss,discrete=TRUE,boot=500)
summary statistics
------
min:  0   max:  5
median:  2
mean:  1.82
estimated sd:  1.409778
estimated skewness:  0.5695758
estimated kurtosis:  2.703005
```

Code Snippet 6 Description of a Poisson distribution in R

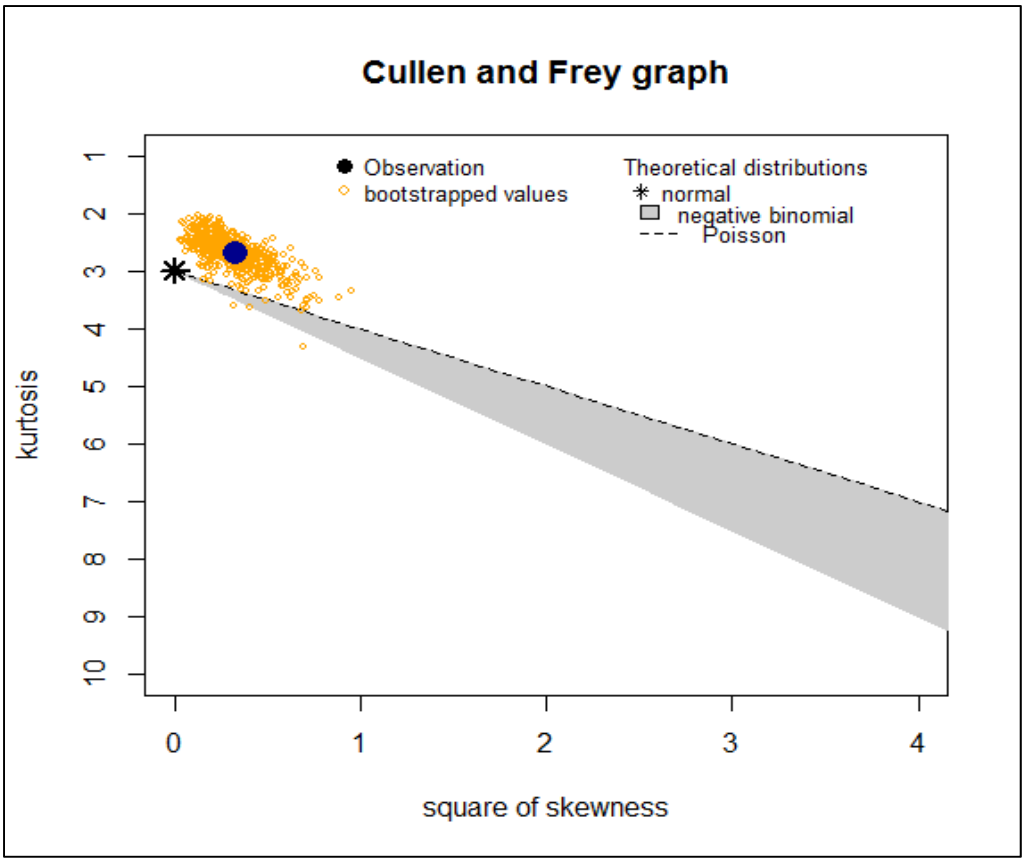


Figure 50 Cullen and Frey Plot of a Poisson distribution

1.7.12 Distribution of Emotional Health Dependent Variables

1.7.12.1 Binomial Emotional Health Variables

Emotional truth, self-awareness, and empathy categorical variables are dependent discrete-choice outcome variables (a.k.a. unordered polytomous variables). The standard statistical model for discrete-choice is logistical regression; where each binomial choice is split out from the multinomial category and independent logistical regressions are performed on each binomial (e.g. emotional truth variable ***eTRUTH*** split into binomials ***happyTRUTH***, ***sadTRUTH***, ***angryTRUTH***, ***anxiousTRUTH***, and ***neutralTRUTH*** variables).

For each $emotion \in \{okay, happy, sad, angry, nervous\}$, an R function is executed to explore for significant effects ($p<0.05$) and trends ($p<0.1$) against categorical factors group3, language, and gender. Goodness of fit tests is run as well. This process is repeated for each $category \in \{eTRUTH, eSELFAWARE, eEMPATHY\}$.

Splitting a categorical variable into multiple discrete-choice variables depends on the mutual exclusivity of choices and as such violates the independence assumption known as "Independence from Irrelevant Alternatives" (IIA) [77] where the odds of preferring one class over another do not depend on the presence or absence of other "irrelevant" alternatives. Multilevel multinomial logit modelling enables the analysis of discrete-choice dependent variables accommodating dependence at unit and cluster levels [78]. Markov Chain Monte Carlo Generalized Linear Mixed Models (MCMCglmm) [79] are experimented in chapter 6 and compared to the discrete-choice (binomial) method in chapter 5.

1.7.12.2 Continuous Emotional Health Variables

As examined in detail in Appendices J and K respectively, both two-level length-of-speech $length_{ij}(X_{ij})$ and confidence score $e_{ij}^{confidence}(X_{ij})$ have Gamma distributions.

1.7.12.3 Count Variables

In chapter 2 we analyze two-level votes collected from anonymous and transcriber sources. Both these variables have Poisson distributions.

1.7.13 Generalized Linear Mixed Models in R

In section 1.7.6 we introduced HLMs with the lme() R function from the nlme package [80]. Lme() can only be applied to normally distributed data and as such is not suitable for the analysis of Binomial and Poisson distributions. Glmer() in the R package lme4 [81] fits generalized linear mixed models on these distribution types. Running the R command help(glmer) gives details on glmer.

```
> help(glmer)
glmer(formula, data, family = dist, start = NULL, verbose = FALSE, nAGQ = 1, doFit
= TRUE, subset, weights, na.action, offset, contrasts = NULL, model = TRUE,
      control = list(), ...)
glmer(formula, data, family = dist)

 formula   a two-sided linear formula object describing the fixed-effects part of
           the model, with the response on the left of a ~ operator and the terms,
           separated   by + operators,   on   the   right.   The   vertical   bar
           character "|" separates an expression for a model matrix and a grouping
           factor.
 data      an optional data frame containing the variables named in formula. By
           default the variables are taken from the environment from which lmer is
           called.
 family    a GLM family, binomial(link = "logit"), gaussian(link = "identity"),
           Gamma(link = "inverse"), inverse.gaussian(link = "1/mu^2"), poisson(link
           =   "log"),   quasi(link   =   "identity",   variance   =   "constant"),
           quasibinomial(link = "logit"), quasipoisson(link = "log")If family is
           missing then a linear mixed model is fit; otherwise a generalized linear
           mixed model is fit.
```

Code Snippet 7 help(glmer) in R

The family parameter in glmer() is set to the distribution of the data. All families work well except the Gamma distribution. To compensate for this bug, we log-normalize the data, retest for normality using the Shapiro-Wilk test, and use the Gaussian family. The function call lmer() is the same as glmer() with the family set to Gaussian.

1.7.14 Glmer example

Before applying glmer() to emotional data (example of happiness in section 1.7.20 and all emotional data in chapter 5), it is important to provide a reference to the reader from a reputable source. The example from Snijders chapter 17 example 17.5 [68] is leveraged to describe glmer() output and subsequent measurements. This example analyzes data from a

study was conducted by Ruiter et al. [82] with data from 60 nations obtained from the European/World Values Surveys. Multilevel logistic regression analyses show that religious regulation in a country diminishes religious attendance and that there are only small negative effects of people's own education and average educational level of the country. Religious attendance is strongly affected by personal and societal insecurities and by parental and national religious socialization and level of urbanization.

Applying glmer() to Snijders' example 17.5 [68] to analyze the question: "is religious attendance affected by income or gender?". The dependent variable is religious attendance (ra), and the explanatory variables are income (a continuous variable) and gender (FEMALE is true=1, or false=0). The multilevel grouping variable is COUNTRY. The previously loaded data set is level12. The mixed model equation with two factors is:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \beta_{2j} x_{ij} + \varepsilon_{ij} \qquad (1.23)$$

Substituting income and gender

$$Y_{ij} = \beta_{0j} + \boldsymbol{income} x_{ij} + \boldsymbol{FEMALE} x_{ij} + \varepsilon_{ij} \qquad (1.24)$$

Converting the equation to glmer() format:

$$glmer(ra \sim income + FEMALE + (1 \,|\, COUNTRY), \quad family = binomial, \quad data = level12) \qquad (1.25)$$

Running equation 1.25 on the data set level12 produces the following output in R Code Snippet 8. Line numbers have been added, so we can analyze each line in Table 9.

There are significant effects for income (p = 0.007532) and gender (p = 0.553214). The higher the income, the less attendance. Females are more likely to attend than males. Replacing coefficients from Code Snippet 8 into equation 1.25:

$$Y_{ij} = -1.758533 - 0.127767x_{ij} + +0.553214x_{ij} + \varepsilon_{ij} \qquad (1.26)$$

```
> summary(mlm1 <- glmer(ra ~  income + FEMALE + (1 |COUNTRY), data=level12))
   1) Generalized linear mixed model fit by the Laplace approximation
   2) Formula: ra ~ income + FEMALE + (1 | COUNTRY)
   3) Data: level12
   4) AIC    BIC logLik deviance
   5) 117509 117548 -58751   117501
   6) Random effects:
   7) Groups  Name        Variance Std.Dev.
   8) COUNTRY (Intercept) 1.9661   1.4022
   9) Number of obs: 135508, groups: COUNTRY, 59
   10) Fixed effects:
   11) Estimate Std. Error z value Pr(>|z|)
   12) (Intercept) -1.758533   0.183183   -9.60   <2e-16 ***
   13) income      -0.127767   0.007532  -16.96   <2e-16 ***
   14) FEMALE       0.553214   0.014912   37.10   <2e-16 ***
   15) ---
   16) Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
   17) Correlation of Fixed Effects:
   18) (Intr) income
   19) income  0.001
   20) FEMALE -0.047  0.054
```

Code Snippet 8 A glmer() Example with Numbered Output in R

Table 9 Glmer Results Description

Line	Description
4-5	**Goodness of fit measurements** AIC Akaike information criterion = 117509 BIC Bayesian information criterion = 117548 logLik log likelihood *L* = -58751 Deviance -2L = 117501 (measurement used and explained below)
6-8	**random effects** Groups the multilevel grouping variable (COUNTRY) Variance σ^2 (within group Variance) = 1.9661 Std.Dev. σ (Standard Deviation) = 1.4022
10 – 13 12 12 12 12 13	**fixed effects** Estimate $\beta_0 = -1.758533$ Std. Error $S.E.(\beta_0) = (\sqrt{\frac{1}{\sum_j s_j{}^{-2}}}\ where\ s_j = \frac{\sigma}{\sqrt{\sum n_j}})$ = 0.183183 Z value standard score = $\frac{\beta_0 - \mu}{\sigma}$ = -9.60 Pr(>|z|)Type 1 significance test result <2e-16 (significant p< 0.5) Estimate $\beta_1 = -0.127767$
16-18 17	**Correlation of fixed effects** Calculated from covariance matrix. Lower the better. 0.001 indicates good correlation for income.

1.7.15 Confidence Intervals of Generalized Linear Mixed Models

The wald() function from the R package spidadev [83] calculates the 95% CI for linear models and linear mixed models; including glmer(). Running wald on the glmer example in section 1.7.14:

```
> wald(mlm1)
 numDF denDF  F.value p.value
     3   Inf 599.1602 <.00001

Coefficients   Estimate Std.Error  DF   t-value p-value Lower 0.95 Upper 0.95
  (Intercept) -1.758533  0.183183 Inf  -9.59986 <.00001  -2.117565  -1.399500
  income      -0.127767  0.007532 Inf -16.96306 <.00001  -0.142530  -0.113005
  FEMALE       0.553214  0.014912 Inf  37.09834 <.00001   0.523987   0.582441
```

Code Snippet 9 Confidence Intervals Calculated with the wald() Function in R

Applying these results to equation 1.26 produces CI: -0.142530 to -0.113005 for the income effect, and CI: 0.523987 to 0.582441 for the gender effect.

Another example is presented in Code Snippet 10 . We calculate the confidence interval for the null model (overall mean) for length-of-speech $length_{ij}(X_{ij})$ (represented as eLENGTH in R). In section 5.13.1, we determine the distribution of eLENGTH is Gamma, so we first log-normalize eLENGTH into eLENLOG.

```
> EMO$eLENLOG = log(EMO$eLENGTH)
> nullmodel <- lmer(eLENLOG ~(1|p), data=EMO, REML=FALSE)
> wald(nullmodel)
 numDF denDF  F.value p.value
     1   Inf 1389.598 <.00001

Coefficients  Estimate Std.Error  DF  t-value p-value Lower 0.95 Upper 0.95
  (Intercept) 1.120054  0.030047 Inf 37.27731 <.00001   1.061164   1.178944

> exp( 1.120054)
[1] 3.06502
> exp( 1.061164)
[1] 2.889733
> exp( 1.178944)
[1] 3.250939
```

Code Snippet 10 Wald Confidence Intervals of a glmer Model in R

We then compute the lmer() two-level null model of the eLENGTH within participants. Running wald(nullmodel) produces the 95% CI. We then convert back from log scale to

linear using the R function exp() to get the results. In reporting terms we can say the population mean of length-of-speech for an emotional response is 3.06502 seconds ($p<0.0001$; CI: 2.889733-3.250939).

1.7.16 Boxplot of Predicted Probabilities

This section is a summary from Kirkman [84]. The box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.

In the box plot depicted in Figure 51, the central rectangle spans the first quartile to the third quartile (the interquartile range or IQR). A segment inside the rectangle shows the median and "whiskers" above and below the box show the locations of the minimum and maximum. This box plot displays the full range of variation (from min to max), the likely range of variation (the IQR), and a typical value (the median).

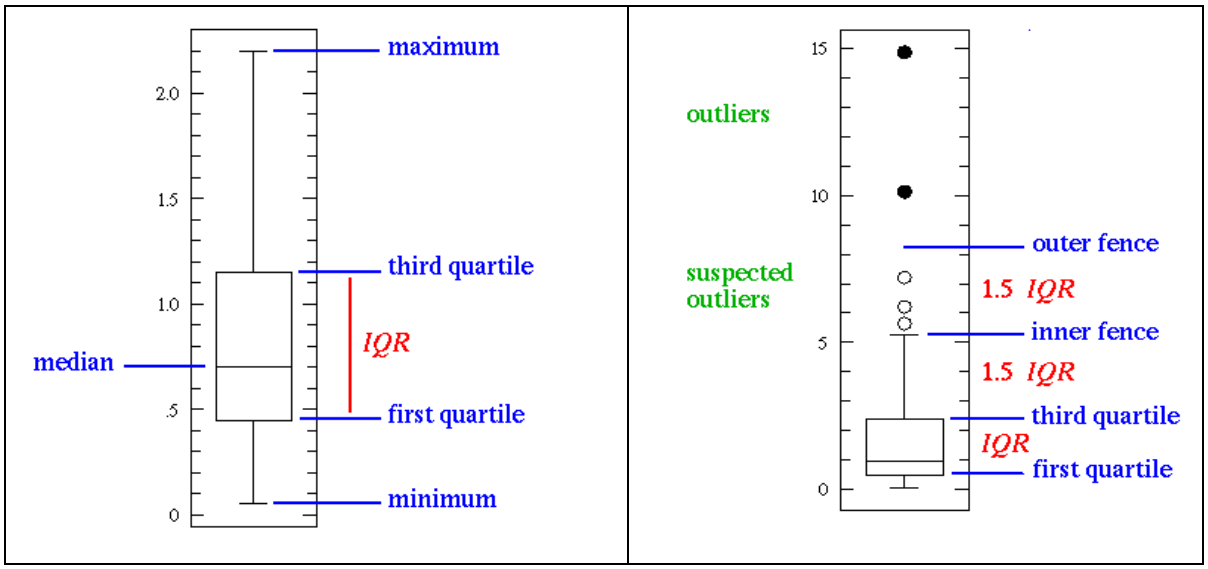


Figure 51 Box Plot and Box Plot with Outliers

Real datasets will display surprisingly high maximums or surprisingly low minimums called outliers. Outliers are either 3×IQR or more above the third quartile or 3×IQR or more below the first quartile. Suspected outliers are slightly more central versions of outliers: either

1.5×IQR or more above the third quartile or 1.5×IQR or more below the first quartile. Outliers are displayed in a box plot as in the right side of Figure 51.

In order to analyze regression models, we need to examine fitted values. The R command fitted() [62] extracts the fitted values from an R regression model object. This command is wrapped in a function THboxPLOT in Code Snippet 11; which takes a glmer model and factor as parameters, and produces a box plot.

```
THboxPLOT <- function(glmm,factor) {
  predprob <-fitted(glmm) # only works if VGAM not loaded!
  plot(predprob ~ factor, data = EMO)
}
```

Code Snippet 11 Function for a Box Plot of a Regression Model's Fitted Values in R

As an example, in Code Snippet 12, we generate the generalized linear mixed model on anxCROWD, and then execute THboxPLOT on Anxiety levels across the three groups (GP, AA, and SUBX[16]).

```
glmm <- glmer(anxCROWD~group3+(1|p) ,family = binomial("logit"), data=EMO)
THboxPLOT(glmm,EMO$group3)
```

Code Snippet 12 Box Plot of Anxiety Levels across Groups in R

The resulting box plot is displayed in Figure 52. Notice the large IQR and higher population mean of anxiety for AA members.

[16] SUBX is actually labeled OPIOID in this example

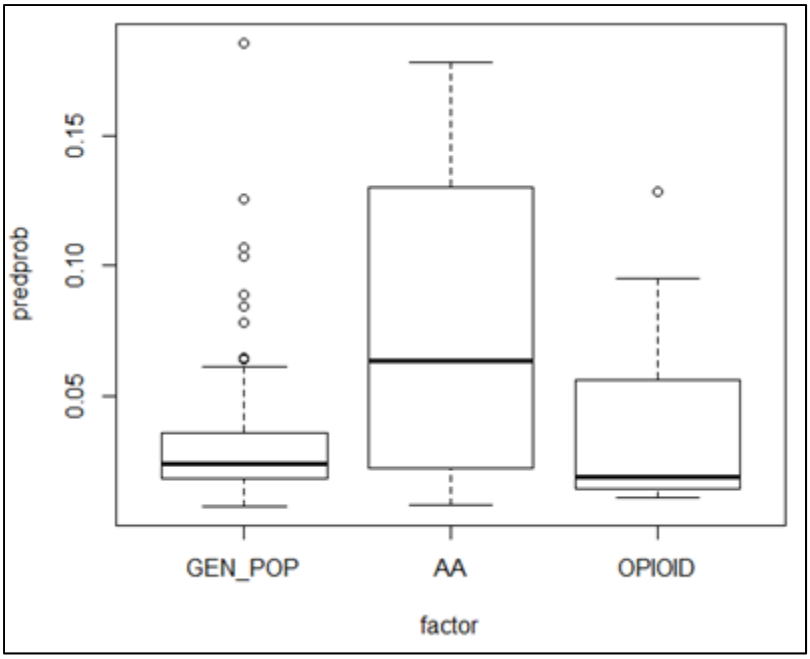


Figure 52 Box Plot of Fitted value for Anxiety versus Group in R

1.7.17 Explained Variance

This section is a summary from Snijders [68]. Calculation of explained variance is demonstrated in the continuous glmer() example in section 1.7.14 and the binomial example in section 1.7.20.

The simplest definition of effect size is:

$$\frac{\mu_{experimental\ group} - \mu_{control\ group}}{\sigma} \qquad (1.27)$$

In OLS regression analysis, the coefficient of determination , R^2, is the proportion of variance explained by the statistical model. The most general definition is:

$$R^2 = \frac{residual\ sum\ of\ squares}{total\ sum\ of\ squares} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \qquad (1.28)$$

For the two-level model given in equation 1.12 the level two (micro-level) residual is U_{0j} with variance τ_0^2 . The level two (macro-level) residual ε_{ij} has variance σ^2. The mean square predicted error is:

$$var\left(Y_{ij} - \sum_h \gamma_h X_{hij}\right) = \sigma^2 + \tau_0^2 \qquad (1.29)$$

Level-one explained proportion of variance is the proportional reduction in the mean squared prediction error. To estimate R_1^2 we compute the null model $Y_{ij} = \gamma_{00} + U_{0j} + \varepsilon_{ij}$ as well as the fitted model and compute 1 minus the ratio.

$$R_1^2 = 1 - \frac{var\left(Y_{ij} - \sum_h \gamma_h X_{hij}\right)}{var\left(Y_{ij}\right)} \qquad (1.30)$$

Calculation of R^2 and R_1^2 for multilevel regression is considerably more complicated and controversial. Snijders extends mcKelvey and Zavoina's measure of single-level logistic R^2 to multilevel [68].

$$\widetilde{Y_{ij}} = \gamma_{00} + \sum_h \gamma_h X_{hij} + U_{0j} + \varepsilon_{ij} \qquad (1.31)$$

The fixed part is $\widehat{Y_{ij}} = \gamma_{00} + \sum_h \gamma_h X_{hij}$. The variance of $\widehat{Y_{ij}}$ is σ_F^2 . The intercept variance is τ_0^2. The level one residual is $\text{var}(\varepsilon_{ij}) = \sigma_R^2 = \frac{\pi^2}{3} = 3.29$. $\text{Var}(\widetilde{Y_{ij}}) = \sigma_F^2 + \tau_0^2 + \sigma_R^2$. The explained variance is σ_F^2. The unexplained variance is $\tau_0^2 + \sigma_R^2$; with τ_0^2 at level two, and σ_R^2 at level one.

$$R_{mm}^2 = \frac{\sigma_F^2}{\sigma_F^2 + \tau_0^2 + \sigma_R^2} \qquad (1.32)$$

```
mlm1 <- lmer(religiousattendance ~  income + FEMALE +
+                      (1 |COUNTRY),family = binomial, data=level12_nT))
> X1 <- attr(mlm1,"X")
> # The parameter estimates for the fixed effects are available as
> (beta1 <- fixef(mlm1))
(Intercept)      income      FEMALE
 -1.7585328  -0.1277673   0.5532137
> # The linear predictor, i.e., linear combination of the rows of X1
> # with weights being the estimated fixed effect parameters, is
> pred1 <- X1 %*% beta1
> # and has variance
> (sigma2_F <- var(pred1))
            [,1]
[1,] 0.09734382
> # The explained variance according to formula (17.22) is
> sigma2_F/(sigma2_F + VarCorr(mlm1)$COUNTRY[1,1] + pi^2/3)
            [,1]
[1,] 0.01818397
```

Code Snippet 13 Pseudo R-Squared Binomial example in R

The explained variance σ_F^2 is 18.18397%.

1.7.18 Intraclass Correlation

The degree of micro-units belonging to the same macro-unit is expressed as the Intraclass Correlation Coefficient (ICC) [68]. Calculation of ICC is demonstrated in the continuous glmer() example in section 1.7.14 and the binomial example in section 1.7.20. For Linear data, the ICC is defined as [68]:

$$\rho_I = \frac{variance\ between\ macro-units}{total\ variance} = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} \qquad (1.33)$$

1.7.19 Goodness of Fit

Measures such as R^2 based on residual errors are not very informative as a measure of fit for multilevel models [68]. There are many proposals for measures of fit in literature, but there is no standard. In Snijders [68] Deviance measurement is preferred. Summarizing from Baroni [85]: Deviance is an important measure of fit of a model, used also to compare models. The larger the deviance; the worse the fit. As parameters are added, deviance should decrease. The difference in deviance between a simpler and a more complex model approximates a χ^2

(chi-squared) distribution with the difference in number of parameters as appropriate degrees of freedom (df). Improvement is significant ($\alpha = 0.05$) if the deviance difference is larger than the parameter difference. A model can also be compared against the "null" model.

Akaike's An Information Criterion (AIC) is a preferred measure for model "goodness of fit" by statisticians John Fox [71], Georges Monette [72], and Heather Krause[86] ; the lower the AIC, the better the fit. Log Likelihood is calculated from the AIC and is used as a measure of Goodness of fit.

The χ^2 distribution deviance differences between the null model and models with factors use binomial example in section 1.7.20, and throughout chapter 5.

1.7.20 Glmer() Discrete-Choice Outcome Variable Analysis Example

We analyze "Happy" emotional truth (happyTRUTH) across all patients to demonstrate the statistical analysis process of a discrete-choice multilevel outcome. This section follows the procedure of Szmaragd et al [87]. We begin by fitting the null two-level model with an intercept β_0 and random effect μ_{0j} to determine the mean happiness for an average participant.

$$logit(\pi_{ij}) = Ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \mu_{0j} \qquad (1.34)$$

The intercept β_0 is the between effect, and μ_{0j} is specific to each patient. μ_{0j} is assumed to follow a normal distribution. This assumption will be tested. The formula-grouping factor in the glmer() function call is set to (1|p) in order to group factors (e.g. group, gender, language) by $participant_j$, denoted by p.

```
Generalized linear mixed model fit by the Laplace approximation
Formula: happyTRUTH ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 8016 8030  -4006     8012
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.86772  0.93151
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.29258    0.09515  -13.59   <2e-16 ***
```

Code Snippet 14 Happiness Emotion Truth 2-level Null Model in R

The log-odds mean for happiness for an average participant ($\mu_{0j} = 0$) is highly significant (p<2e-16) and is estimated at -1.29258 which is a probability of $\frac{e^{-1.29258}}{(1+e^{-1.29258})} = 21.54\%$.

The log-odds for $participant_j$ is estimated at $-1.2961 + \widehat{\mu_{0j}}$, where $\widehat{\mu_{0j}}$ is the participant residual. A participant with $\widehat{\mu_{0j}} > 0$ has a log-odds higher than average, while $\widehat{\mu_{0j}} < 0$ is a below-average participant. The between-participant (level 2) variance of $\widehat{\mu_{0j}}$ is estimated at $\hat{\sigma}^2_{\mu 0} = 0.908412$.

The goodness-of-fit likelihood ratio statistic for testing the hypothesis $\sigma^2_{\mu 0} = 0$ can be calculated by comparing the two-level model with the corresponding single-level model:

```
> happy.glm <- glm(happyTRUTH ~ 1 ,family = binomial, data=EMO)
> logLik(happy.glm)-logLik(g1)
'log Lik.' -320.2453 (df=1)
```

Code Snippet 15 Happiness Two-Level and One-Level Model Comparison in R

The Log Likelihood test statistic is 620 (-2 x -320.2453) indicating strong evidence that between-patient variance is non-zero.

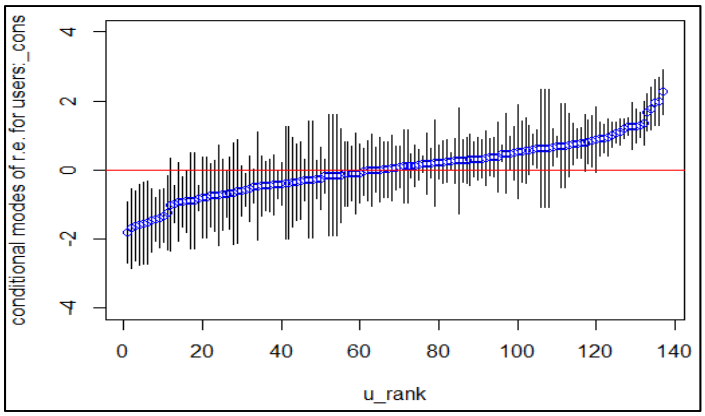


Figure 53 Estimates of the Residuals $\hat{\mu}_{0j}$ for each $patient_j$

The plot in Figure 53 shows the residuals for all patients is close to homoscedastic; all residuals are within the 95% confidence interval. The distribution is normal as the Shapiro-Wilk test null hypothesis is rejected (p >0.05):

```
> shapiro.test(ranef(g1)$p[,1])
        Shapiro-Wilk normality test

data:  ranef(g1)$p[, 1]
W = 0.9905, p-value = 0.55
```

Code Snippet 16 Shapiro-Wilk Normality Test in R

Dummy regressors from Table 10 are added to the logit equation for the group (GP, AA, and SUBX) effect producing equation 1.36.

Table 10 Dummy Regressors

Factor	**D1**	**D2**
GP	0	0
AA	1	0
SUBX	0	1

$$logit(\pi_{ij}) = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \mu_{0j} \quad (1.35)$$

$$logit(\pi_{ij}) = \text{GP} + \text{AA}D_{i1} + \text{SUBX}D_{i2} + \mu_{0j} \qquad (1.36)$$

We calculate the generalized linear mixed model using glmer():

```
Formula: happyTRUTH ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 6959 6987  -3476     6951
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.85201  0.92305
Number of obs: 6650, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.11687    0.15934  -7.009 2.39e-12 ***
         AA  -0.03508    0.23719  -0.148    0.882
       SUBX -0.60438    0.25324  -2.387    0.017 *
> inv.logit(-1.11687) # General population intercept
[1] 0.2465923
> inv.logit(-1.11687-0.60438) # OPIOID intercept
[1] 0.1517102
```

Code Snippet 17 Happiness versus Group Two-Level Regression Model in R

We use the Wald function to explore significance differences of happiness between the General Population versus AA members and SUBX patients as well as confidence intervals. A p-value less than 0.05 indicates statistical significance.

```
>wald(g2) # wald Confidence intervals (GP versus…)
numDF denDF  F.value p.value
     3   Inf 56.19705 <.00001

Coefficients    Estimate Std.Error  DF   t-value p-value Lower 0.95 Upper 0.95
  (Intercept)  -1.116871  0.159340 Inf -7.009359 <.00001  -1.429172  -0.804570
        AA     -0.035078  0.237191 Inf -0.147888 0.88243  -0.499963   0.429808
       SUBX    -0.604382  0.253240 Inf -2.386597 0.01701  -1.100724  -0.108041

> inv.logit(-1.116871)
[1] 0.2465921
> inv.logit(-1.116871-0.604382)
[1] 0.1517098
```

Code Snippet 18 Happiness versus Group Two-Level Confidence Intervals in R

There is an effect for SUBX (p = 0.01701). From the coefficient estimates in Code Snippet 18, and equation 1.36 and eliminating the effect of AA by setting the dummy to 0:

$$logit(\pi_{ij}) = -1.1168 - 0.604\text{SUBX} + \mu_{0j} \quad (1.37)$$

To calculate the intercept for the GP, we set the SUBX regressor to 0:

$$logit(\pi_{ij}) = -1.1168 - 0.604(0) + \mu_{0j} \quad (1.38)$$

$$logit(\pi_{ij}) = -1.1168 \quad (1.39)$$

$$\pi_{ij} = 24.65921\% \quad (1.40)$$

To calculate the intercept for SUBX, we set its dummy regressor to 1:

$$logit(\pi_{ij}) = -1.1168 - 0.604(1) + \mu_{0j} \quad (1.41)$$

$$\pi_{ij} = 15.17397\% \quad (1.42)$$

Indicating the probability of a SUBX patient being happy is 9.5% less than the GP (p = 0.017).

In order to determine if there is an effect between AA members and SUBX patients, we must relevel, and set AA members as the reference.

```
> EMO$group3 <- relevel(EMO$group3,ref="AA")  # relevel to compare (AA versus…)
>wald(g2)
Coefficients      Estimate Std.Error  DF   t-value p-value Lower 0.95 Upper 0.95
  (Intercept)   -1.151949  0.175699 Inf -6.556362 <.00001  -1.496313  -0.807584
  group3GEN_POP  0.035078  0.237191 Inf  0.147889 0.88243  -0.429808   0.499963
  group3OPIOID  -0.569072  0.263839 Inf -2.156894 0.03101  -1.086187  -0.051958
> inv.logit(-1.151949  )
[1] 0.2401333
> inv.logit(-1.151949 -0.569072   )
[1] 0.1517397
```

Code Snippet 19 Re-Leveling of Group Factor to Reveal AA versus SUBX Effect in R

Code Snippet 19 reveals a significant difference of 8.8% between AA and SUBX (p = < 0.031).

- General Population pr(happyTRUTH) = 24.7% (95% CI, 19.2%–31.0%)

- AA Member pr(happyTRUTH) = 24.0% (95% CI, 16.4%–33.7%)
- SUBX pr(happyTRUTH) = 15.2% (95% CI, 9.7%–22.9%)

The $R^2_{binomial}$ = 0.009683 indicating that group3 describes a small amount of variance. The ICC is 0.205707 indicating a good degree of correlation within groups.

```
> anova(g1,g2)
Data: EMO
Models:
g1: emotion ~ (1 | p)
g2: emotion ~ group3 + (1 | p)
   Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
g1  2 8016.1 8029.9 -4006.0
g2  4 6959.5 6986.7 -3475.7 1060.6      2  < 2.2e-16 ***
```

Code Snippet 20 Analysis of Variance in R

With four degrees of freedom, the 99% cut-off for X^2 distribution is 13.3; anova(g1,g2) is highly significant with $X^2 = 1060.6$.

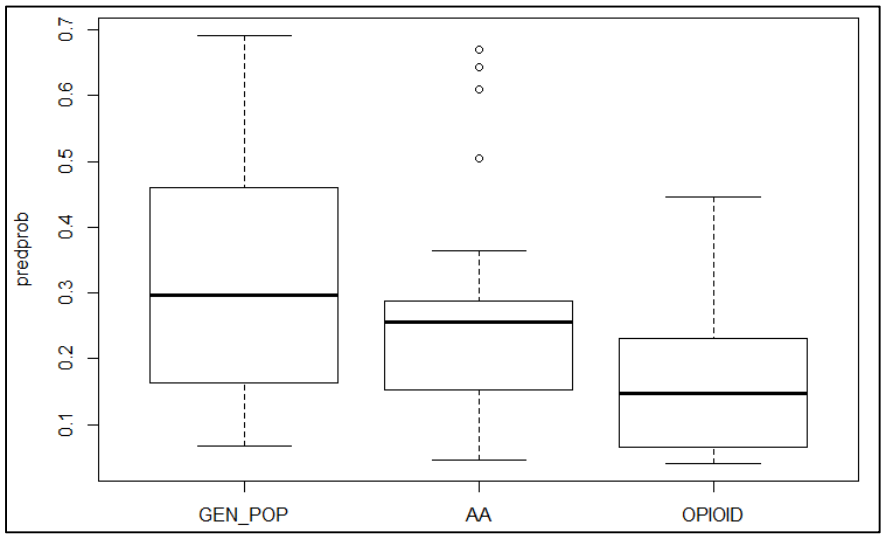


Figure 54 Predicted Probabilities of happyTRUTH versus Group3

The boxplot of predicted probabilities of happyTRUTH~group3 is presented in Figure 54. The T-shaped whiskers provide the minimum and maximum range for the population; the box spans an IQR of 25% - 75% quartiles; the dark line in the box marks the median value; circles represent outliers.

1.7.21 Kaplan-Meier Survival Estimate

As described in Table 5 on page 30, emotional data was collected in trials from GP, SUBX, and AA groups of varied gender and language. In the context of emotional trial participation, "survival" is a measurement of whether a participant completed the trial; or quit before the trial end. This metric is important to measure patient acceptance, and measure compliance in order to establish protocol validity. The Kaplan-Meier estimate is applied to measure trial survival in section 5.14.

The Kaplan-Meier estimate is a nonparametric Maximum Likelihood Estimate (MLE) of the survival function, $S(t)$ [88]. This estimate is a step function with jumps at observed event times, t_i. In the formula below, it is assumed the t_i are ordered: $t_1 < t_2 < \cdots t_D$. If the number of individuals with an observed event time t_i is d_i, and the value Y_i represents the number of individuals at risk at time t_i (where at risk means individuals who do not survive time t_i or later), then the Kaplan-Meier estimate of the survival function [89, 90] is given by equation 1.43 and its estimated variance is given by 1.44.

Kaplan-Meier survival estimate:

$$\hat{S}(t) = \prod_{t_i \le t} \begin{cases} 1 & if\ t < t_1 \\ \left[1 - \frac{d_i}{Y_i}\right] & if\ t_1 \le t \end{cases} \qquad (\ 1.43\ )$$

Kaplan-Meier variance:

$$\hat{V}\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \sigma_S^2(t) = \left[\hat{S}(t)\right]^2 \sum_{t_i < t} \frac{d_i}{Y_i(Y_i - d_i)} \qquad (\ 1.44\ )$$

The Kaplan-Meier estimate is survfit() in the R package survival [88].

The R package OIsurv [91] is used as an example. The tongue data set in OIsurv is from a study on the prognosis of patients with cancer of the tongue. Code Snippet 21 executes survival analysis on the tongue data set.

```
> library(survival)
> library(OIsurv)
> data(tongue)
> attach(tongue)
The following object(s) are masked from 'tongue (position 3)':
    delta, time, type
The following object(s) are masked from 'tongue (position 4)':
    delta, time, type
> mySurv <- Surv(time[type==1], delta[type==1])
> (myFit <- survfit(mySurv ~ 1))
Call: survfit(formula = mySurv ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
     52      52      52      31      93      67      NA
> plot(myFit, main="Kaplan-Meier estimate with 95% confidence bounds",
+ xlab="time", ylab="survival function")
```

Code Snippet 21 Survival Estimate for Patients with Cancer of the Tongue in R

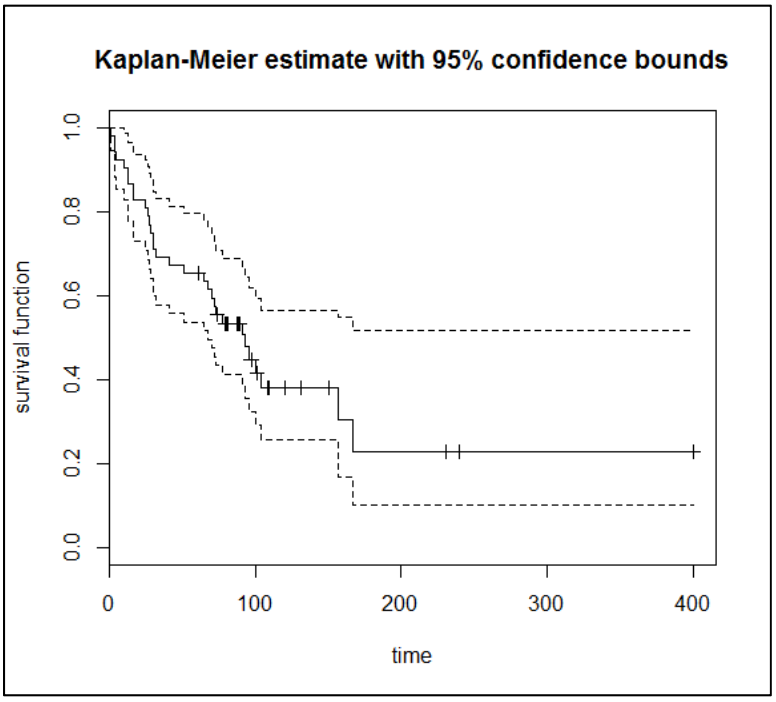


Figure 55 Survival Estimate for Patients with Cancer of the Tongue

Figure 55 indicates that 80% of patients died within the first 200 weeks.

1.8 Step 7: Monitoring Patients' Emotional Health

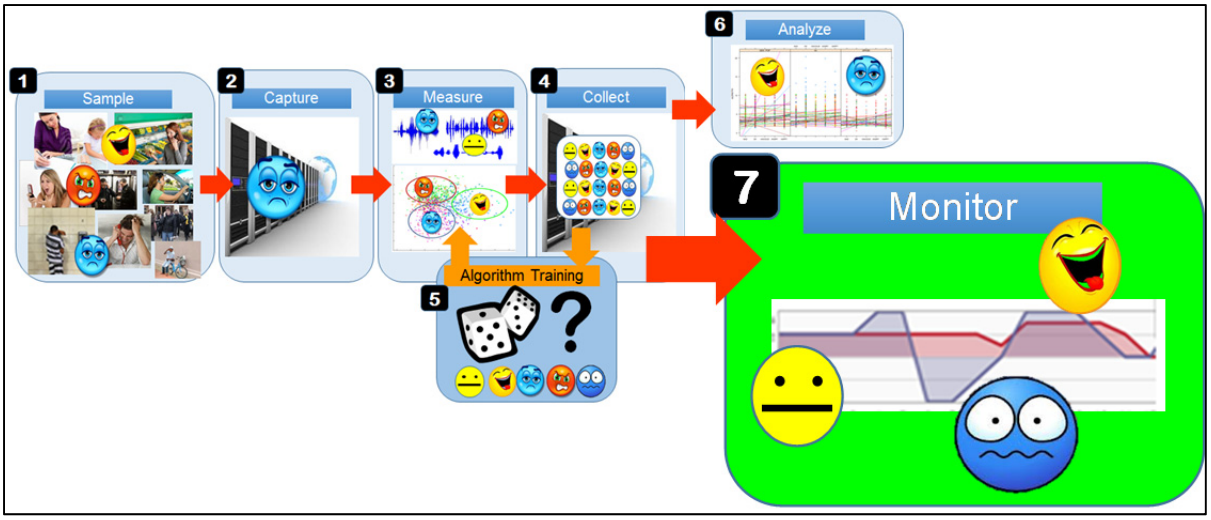


Figure 56 Step 7: Emotional Health Monitoring

Empirical emotional health data and trend analysis using this toolkit should improve understanding of a patient's emotional health between sessions. For example, monitoring can discover episodes or the chronic presence of depression, anxiety, and resentment. Emotional recordings can be played back to trigger recall of events and behaviours associated with peaks and valleys of longitudinal emotional state charts. Historical data can be reviewed for evidence of progress.

Crisis intervention can be triggered on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions (possible indication of relapse or an episode of mood disorder such as depression).

In Figure 57, the toolkit has been customized for Dr. David Nussbaum's 2013 study of emotional traits of gamblers (University of Toronto. Conducted by PhD candidate Lucas Ogura).

Figure 57 Emotional Health Toolkit Login Web Page

In Figure 58, patients are signed up by entering their name (or alias), phone number, PIN, and times during the day the system will call them. The PIN allows the participant to also call in by dialing a 1-800 number. If the professional wants the participant to view their emotional health progress on the web, an email and password is entered.

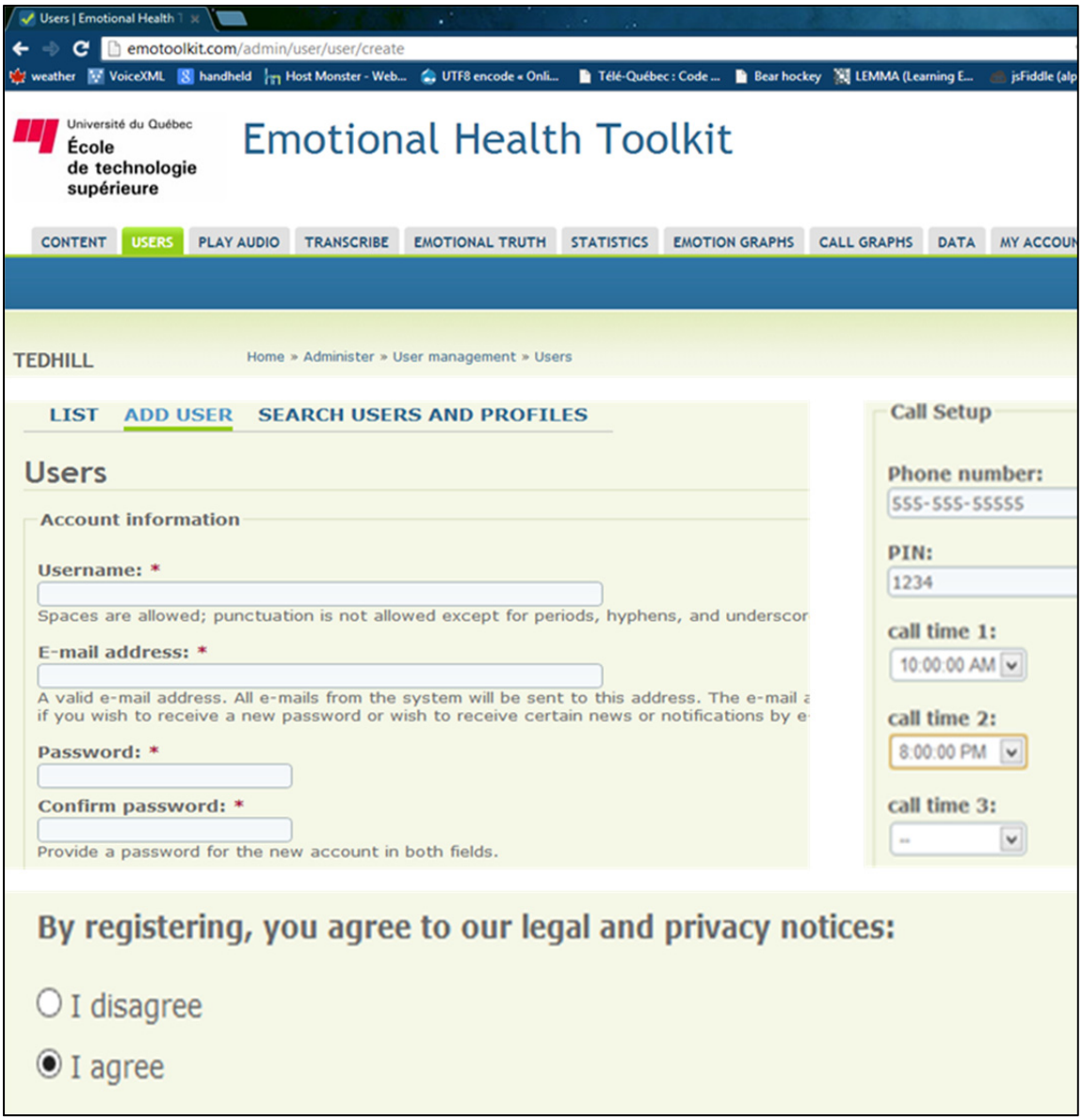


Figure 58 Patient Registration

Figure 59 provides a view of daily call completion rates for a SUBX patient. The highlighted lapse in calls could indicate isolation, depression, or drug relapse. A crisis intervention alert via Short Message Service (SMS) or email can be triggered on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions.

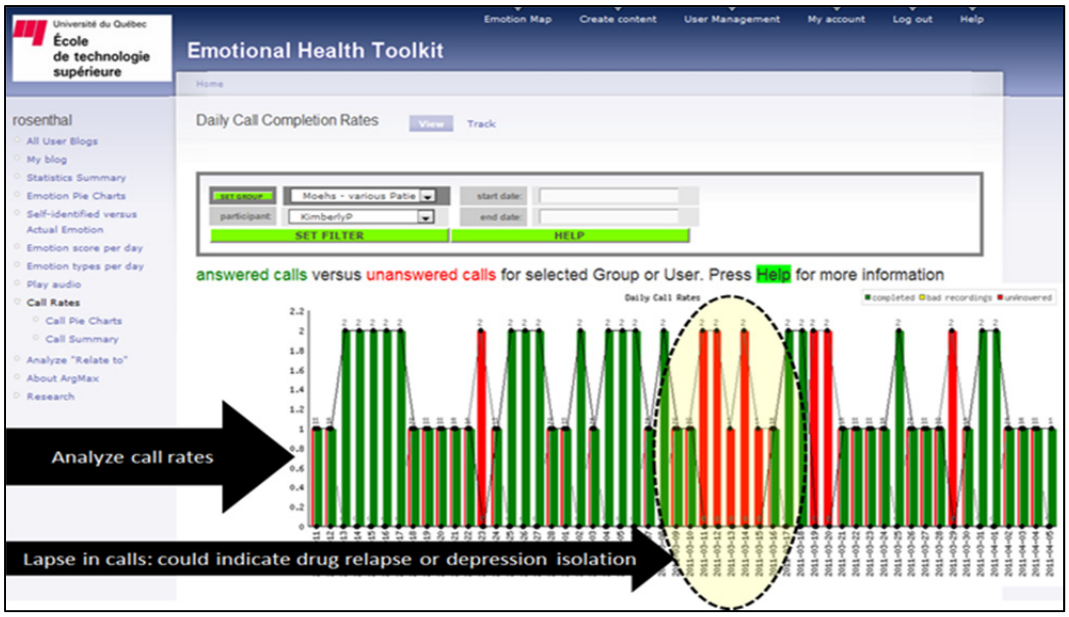


Figure 59 Daily Call Completion Rates

Figure 60 shows the view where a therapist can playback recordings associated with emotional health samples in order to explore feelings and behaviours during monthly assessments.

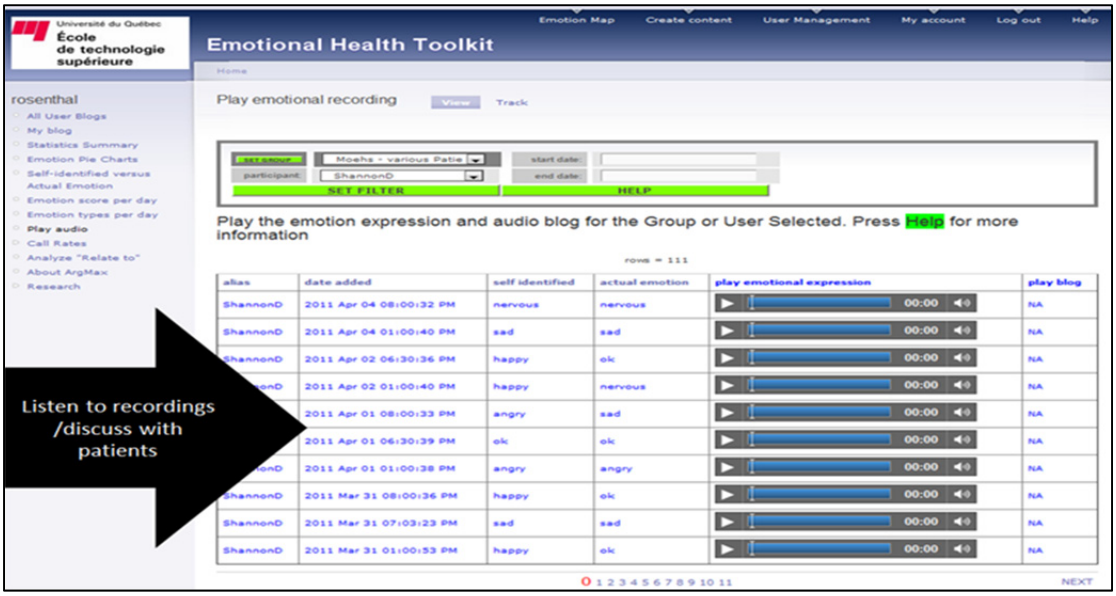


Figure 60 Emotional Recording Playback Tool

The patient's emotions can be graphed over time to visualize problems. In Figure 61, a SUBX patient is negative for a period of 8 days. Analysis of recordings indicate a traumatic experience occurred August 30th. The emotional residual lasted a period of 8 days. The therapist could have intervened during this period, or explore the episode during the next therapy session.

Figure 61 Monitoring Patient Emotions over Time

CHAPTER 2

UNSUPERVISED CROWD-SOURCED CORPUS LABELING

Eight thousand three hundred and seventy-six (8,376) audio recordings and momentary emotional states were collected from 2010 to 2011, from one hundred and thirteen (113) participants including three groups: SUBX patients at Dr. Charles Moehs MD MPH clinic (Occupational Medicine Associates of Northern New York) N = 36 [13 men; Expressions = 1054] with an average SUBX continued maintenance period of 1.66 years (Standard Deviation (SD) = 0.48); General Population (GP), N = 44 [15 men; Expressions = 2440]; and Alcohol Anonymous (AA), N = 33 [29 men; Expressions = 3848].

The emotional truth of each emotionally charged audio recording must be accurately labeled in order to develop emotion detection algorithms and to perform statistical analysis.

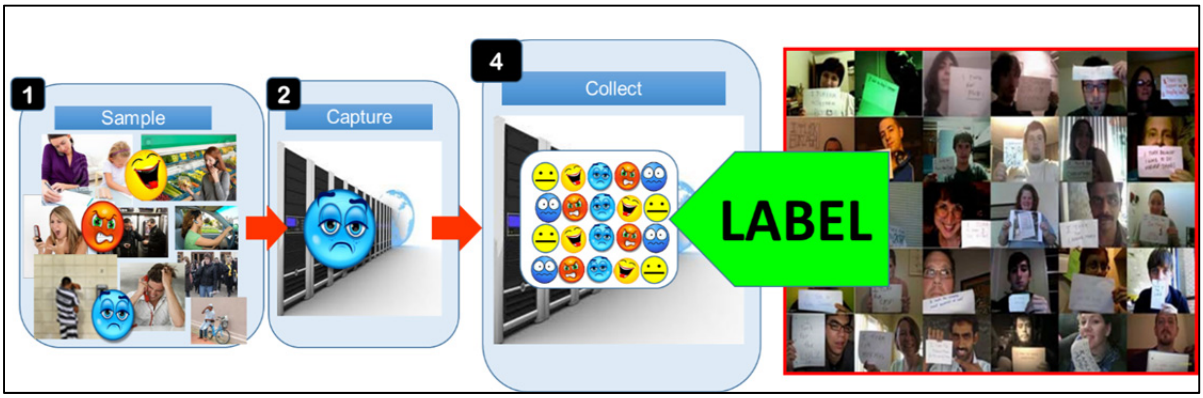


Figure 62 Crowd-Sourced Corpus Labeling

Labeling speech utterances is a time-consuming and labour-intensive process. Typically, as for the FAU Aibo Emotion corpus [63], raw audio recordings are first segmented manually into small, syntactically meaningful 'chunks' using syntactic-prosodic criteria that are subsequently labeled with an emotion tag by paid professional transcribers [63].

Unsupervised emotional truth corpus labeling requires automatic chunking of audio into an utterance with a single emotion, and unsupervised automatic emotional truth labeling. Unsupervised emotional truth labeling was experimented by leveraging the response to the IVR dialogue prompt "Guess the emotion of the following speaker". The accuracy of unsupervised emotional truth labeling is compared to an emotional truth label with maximized certainty.

A confidence measure, as a measure of emotional truth confusability and affect, is introduced. A certainty measure, based on the total number of votes and the confidence measure, provides insight into the accuracy of the emotional truth label.

To maximize the certainty measure of the emotional truth for emotion detection algorithm training and statistical analysis, a fused MV emotional truth classifier is constructed from the unsupervised classifier, transcriber classifier, and self-report.

2.1 Automatic Chunking

Emotion Data collected by asking participants to respond to the open Question ***"How are you Feeling?"*** could result in an utterance with multiple emotions. As an example, in Figure 63 there are three emotions (sad, angry, and happy) in the utterance.

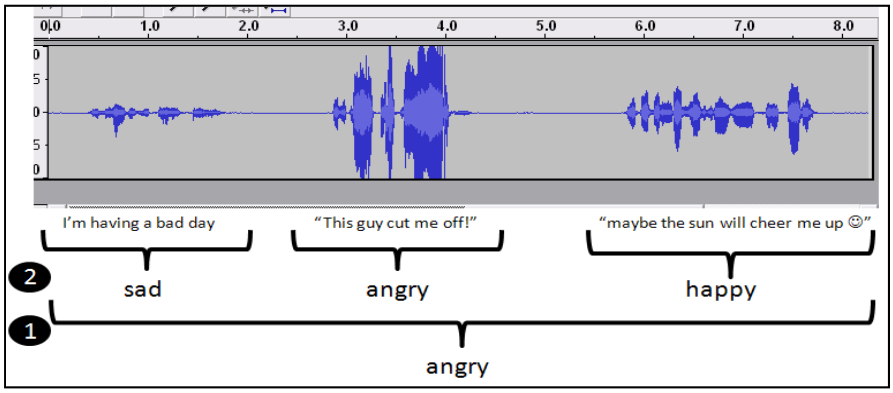


Figure 63 Example Utterance with Multiple Emotions

It would be incorrect to label the utterance with the predominant emotion angry. To avoid multiple emotions, two remedies were combined. The first remedy was to limit the maximum speech duration to 10 seconds. 4/5 respondents in a post data collection survey indicated 10 seconds was sufficient time to express an emotion (see section 2.6.3 on page 34). The second remedy is intuitive training. After recording their emotion, the participant is asked to self-report: "Please classify your mood. Press 1 for okay, 2 for happy…" After a few telephone calls, the participant intuitively knows to express a single emotion. Manual screening of the corpus confirmed that these remedies are effective in limiting the utterance to one emotion.

2.2 Accuracy of FAU Aibo Emotion corpus emotion labels

Steidl et al. states that normally only in a few cases did labelers agree on one common emotion label [43]. In most cases, only 3 out of 5 labelers agreed on emotional content [43] as depicted in Figure 64.

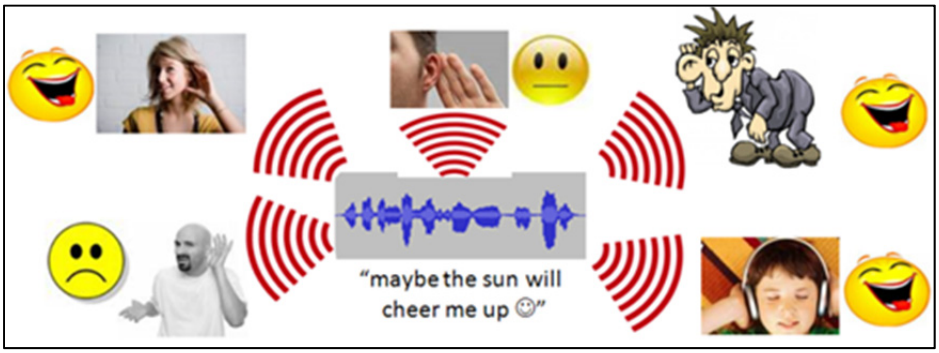


Figure 64 Human Labeling of Emotional Audio Content

This is 60% agreement on an emotional label amongst 5 voters. We will name the criteria of minimum 3 votes with 60% agreement threshold: $Steidl_{3:60\%}$. For example, vote counts of 3/5, 3/4, and 4/6 satisfy $Steidl_{3:60\%}$. Vote counts of 2/2 and 2/3 do not meet the minimum vote criteria. A vote count of 4/10 does not satisfy the agreement threshold.

2.3 ReCAPTCHA Crowd-Sourced Automatic Corpus labeling

The crowd-source approach towards unsupervised emotional truth corpus labeling was derived from an investigation of reCAPTCHA [92], used by over 30 million users per day, which improves the process of digitizing books by voting on the spelling of words that cannot be deciphered by Optical Character Recognition (OCR). CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a challenge response test used on the World Wide Web to determine whether a user is a human or a computer. Whereas standard CAPTCHAs display images of random characters rendered by a computer, reCAPTCHA displays words taken from scanned texts. The solutions entered by humans are used to improve the digitization process. To increase efficiency and security, only the words that automated OCR programs cannot recognize are sent to humans [92].

2.3.1 ReCAPTCHA Accuracy

The accuracy of reCAPTCHA crowd-sourcing is startling: *"From analysis of our data, 67.87% of the words required only two human responses to be considered correct, 17.86% required three, 7.10% required four, 3.11% required five, and only 4.06% required six or more. The reCAPTCHA system achieves an accuracy of 99.1% at the word level. An accuracy of 99.1% is within the acceptable 'over 99%' industry standard guarantee for 'key and verify' transcription techniques in which two professional human transcribers independently type the data and discrepancies are corrected. As an anecdote, manual transcription performed as ground truth achieved 99.2% accuracy"* [93].

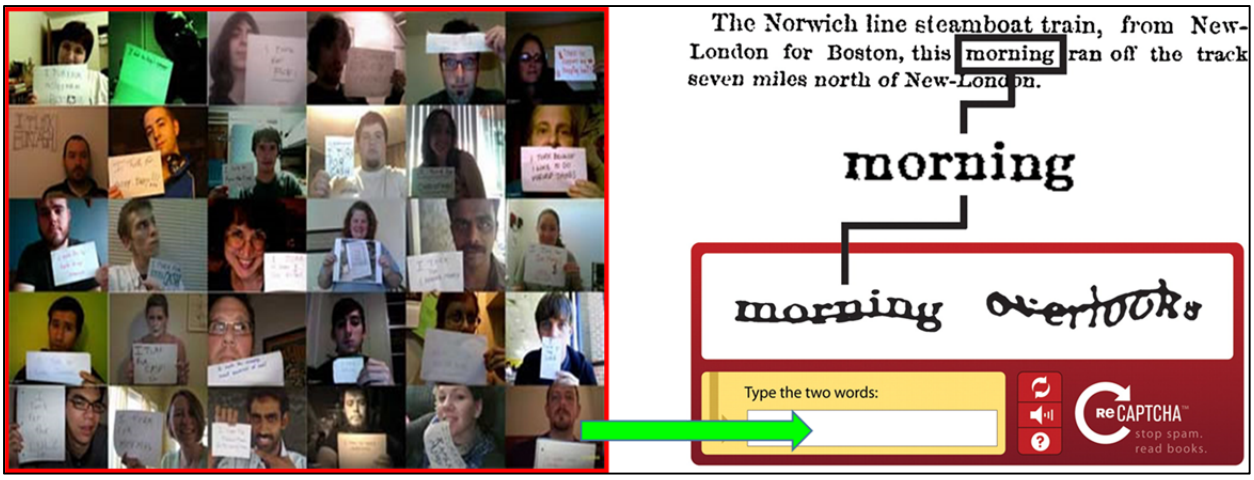


Figure 65 ReCAPTCHA

2.4 Majority Vote Classifier

MV and Maximum-a-Posteriori (MAP) decision algorithms determine the most likely crowd-sourced emotion label. Given the independent categorical variable $e, e \in E\{neutral, happy, sad, angry, anxious\}$, a count of c_e votes for each e, and the total count C_E for all emotions:

$$C_E = \sum_E c_e = c_{neutral} + c_{happy} + c_{sad} + c_{angry} + c_{anxious} \quad (2.1)$$

The most likely emotion $\hat{e}$ is

$$\hat{e} = \underset{e \in E}{argmax}\, \mathrm{p}(c_e|e)\, \mathrm{p}(e) \quad (2.2)$$

The MV estimate for $\mathrm{p}(c_e|e)$ is simple division of c_e by C_E.

$$\widehat{\mathrm{p}(c_e|e)} = \frac{c_e}{C_E} \quad (2.3)$$

If we assign equal likelihood to all emotions then $\mathrm{p}(e) = \frac{1}{5}, \forall e$, equation 3.2 becomes

$$\hat{e} = \underset{e \in E}{argmax} \left[\frac{C_e}{C_E} \right] \qquad (2.4)$$

For example, given the vote counts of anonymous assessments collected in response to "guess the emotion of the following speaker" of recording X_{ij} in Table 11.

Table 11 Crowd-Sourced Vote Collection Example 1

votes	$C_{neutral}$	C_{happy}	C_{sad}	C_{angry}	$C_{anxious}$	C_E
anonymous assessments		2		4	1	7

We can calculate an approximation $\hat{e}$ of the emotional truth of recording X_{ij} or $e_{ij}^{anonymous}(X_{ij})$ by using equation 2.4:

$$p(X_{ij}|neutral) = \left[\frac{0}{7}\right] = 0 \qquad p(X_{ij}|happy) = \left[\frac{2}{7}\right] = 0.143$$

$$p(X_{ij}|sad) = \left[\frac{0}{7}\right] = 0 \qquad p(X_{ij}|angry) = \left[\frac{4}{7}\right] = 0.571$$

$$p(X_{ij}|anxious) = \left[\frac{1}{7}\right] = 0.154$$

$$e_{ij}^{crowd}(X_{ij}) = \underset{e=neutral,happy,\ldots,anxious}{\text{argmax}} P(X_{ij}|\text{e}) = angry$$

2.4.1 Confidence Score

The ratio of the winning MV count over all votes is the confidence score.

$$e_{ij}^{confidence}(X_{ij}) = \frac{C_{\hat{e}}}{C_E} \qquad (2.5)$$

From the example in Table 11 the confidence score of the emotional truth *angry* is:

$$p(X_{ij}|angry) = \boldsymbol{e_{ij}^{confidence}(X_{ij})} = 0.571$$

2.4.2 Certainty Score

The confidence score gives an indication on confusability of the emotional truth that can be leveraged as an indication of expressiveness. However, for accuracy measurement, two utterances with 2/2 and 5/5 votes respectively for happy (100% confidence scores) should not have the same accuracy. There should be more certainty assigned to 5/5 votes. A new measure to reflect the number of votes is required.

ReCAPTCHA accuracies on human responses in word transcription 3.3.1 on page 64 are the only empirical data available on accuracy of crowd-sourced transcription known to this author. No data exists on the accuracy of a human's ability to determine the emotion of another human other than Steidl's 3/5 concurrence [43], ($Steidl_{3:60\%}$).

We assume ReCAPTCHA word transcription accuracies as an approximation to emotional truth labeling accuracies. Ahn states [93] *"67.87% of the words required only two human responses to be considered correct, 17.86% required three, 7.10% required four, 3.11% required five, and only 4.06% required six or more".* If 4.06% of the words required six or more human responses, then:

- Accuracy of 5 responses is 100% - 4.06% = 95.94% correct. Similarly,
- Accuracy of 4 responses is 100% - 4.06% - 3.11% = 92.83%
- Accuracy of 3 responses is 100% - 4.06% - 3.11% -7.11% = 85.72%
- Accuracy of 2 responses is 100% - 4.06% - 3.11% -7.11% - 17.86 = 67.86%

If we assume word accuracies for human responses on emotion labeling then we can estimate how "certain" we are of an emotional label by determining a factor for the number of responses ($certainty\_factor$), and multiplying this factor by the confidence score $e_{ij}^{confidence}$.

We generate a generalized linear regression model (glm) in R, assuming normal distribution [62]. An upper bound certainty of 100% for six responses in agreement from 6 humans (6/6) is assumed.

```
> c <- data.frame(label=c(0,2,3,4,5,6),acc=c(0,0.6786,0.8573,0.9284,0.9594,1))
> m <- glm(c$acc ~ c$label)
> summary(m)

Coefficients:
            Estimate Std. Error Pr(>|t|)
(Intercept)  0.13768    0.16512   0.834   0.4656
c$label      0.16982    0.03981   4.265   0.0236 *
```

Code Snippet 22 Calculation of Certainty Weights using OLS regression in R

$$\widehat{certainty\_factor} = 0.13768 + 0.16982\, x\, (\#\ human\ responses) \qquad (\ 2.6\ )$$

In Table 12 , we apply equation 2.6 to generate the glm approximation for incremental human responses in agreement (votes). Since the regression equation is inexact at the 0 boundary, we adjust the glm approximation of 0.13768 for 0 votes to 0. Similarly, since the *certainty_factor* should never be greater than 100%, we adjust the upper bound, which is 6 votes, to 1.

Table 12 Approximation of the certainty_factor

human responses (votes)	**RECAPTCHA certainty**	**glm approximation**	**certainty_factor**
0	0	0.13768	0
1		0.3075	0.3075
2	0. 6786	0.47732	0.47732
3	0.8573	0.64714	0.64714
4	0.9284	0.81696	0.81696
5	0.9494	0.98678	0.98678
6 or more	~ 1	1.1566	1

We now have a measure of certainty for the emotional truth label of utterance $\boldsymbol{X_{ij}}$:

$$e_{ij}^{certainty}(X_{ij}) = certainty\_factor[\text{votes}] \text{ x } e_{ij}^{confidence}(X_{ij}) \qquad (2.7)$$

Table 13 is an example to illustrate the calculation of $\boldsymbol{e_{ij}^{certainty}}$**.** A total of three votes have been collected; two for happy, and one for anxious.

Table 13 Crowd-Sourced Vote Collection Example 2

votes	$\boldsymbol{c_{neutral}}$	$\boldsymbol{c_{happy}}$	$\boldsymbol{c_{sad}}$	$\boldsymbol{c_{angry}}$	$\boldsymbol{c_{anxious}}$	$\boldsymbol{C_E}$
anonymous assessments		2			1	3

If we apply equation 2.4 to Table 13, we get

$$e_{ij}^{crowd}(X_{ij}) = \underset{e=neutral,happy,\ldots,anxious}{\text{argmax}} P(X_{ij}|\text{e}) = Happy$$

Applying equation 2.5 , $\boldsymbol{e_{ij}^{confidence}(X_{ij}) = \frac{2}{3} =}$ 0.666

Applying equation 2.7 we get:

$$e_{ij}^{certtainty}(X_{ij}) = certainty\_factor[C_E] \text{ x } 0.666$$
$$e_{ij}^{certtainty}(X_{ij}) = certainty\_factor[3] \text{ x } 0.666$$
$$e_{ij}^{certtainty}(X_{ij}) = 0.64714\text{x } 0.666 = 0.431$$

which reduces the confidence score to reflect the low vote count.

In contrast, the example in Table 11 $\boldsymbol{C_E = 7}$ which results in $certainty\_factor = 1$ giving $e_{ij}^{certtainty}(X_{ij}) = 0.571\ x\ 1 = 0.571$, which reflects the high vote count.

2.5 Crowd-Sourced Automatic Corpus Labels Collected

For the 8,249 emotions collected, 16,184 anonymous unsupervised crowd-sourced empathy votes $e_{ijka}^{relate}(X_{ka})$ were automatically collected during the last phase of the telephone dialogue, where the user is prompted with "Guess the emotion of the following speaker" (see section 1.3.2). The second source is the transcription $e_{ijt}^{transcribe}(X_{ij})$ captured from $transcriber_t$. 24,482 transcriptions were also collected on the 8,249 emotions from professional transcribers. The transcription algorithm assures no $transcriber_t$ assesses the same emotional recording X_{ij} twice. The transcriber listens to an emotional recording, and chooses an emotion from a drop-down list. Progress is recorded to allow multi-session transcription and to facilitate calculation of payment to the transcriber for work done.

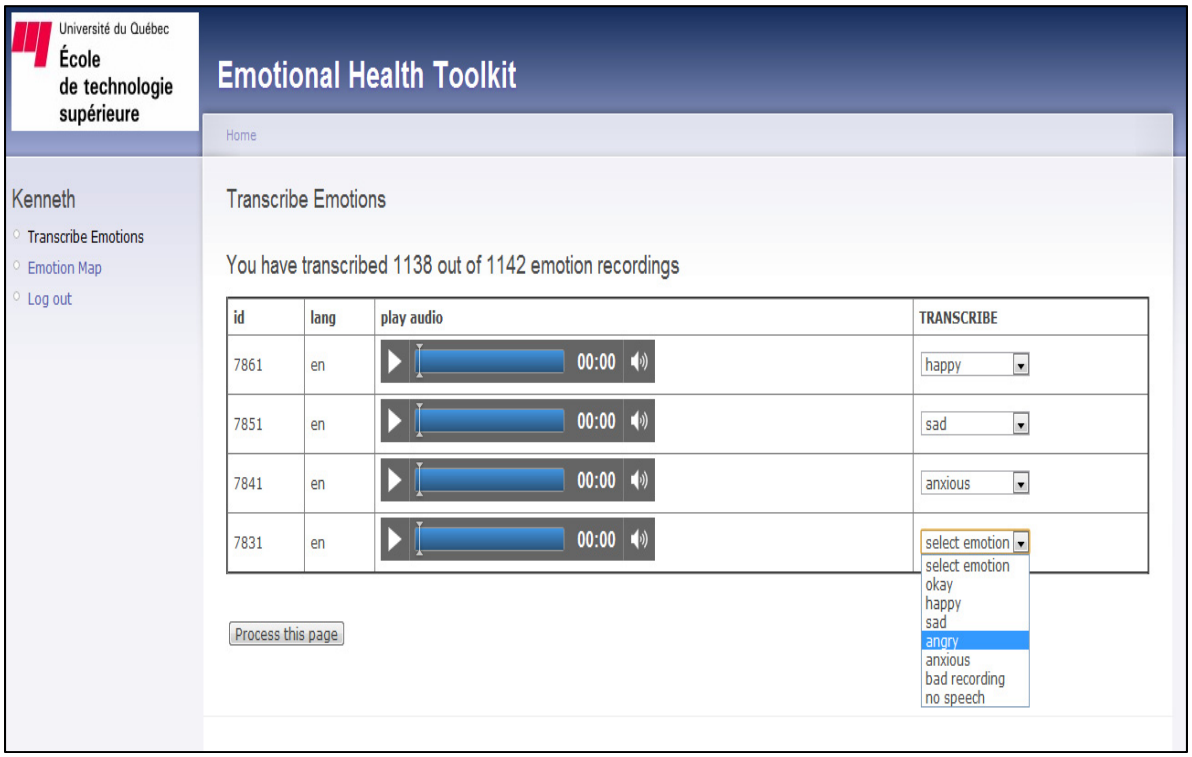


Figure 66 Transcription Tool

2.6 Corpus Label Frequencies

Table 14 Crowd-Sourced Emotion Label Frequencies

votes	A freq	T freq	Total	%
Missing	1281	569	480	5.8
1	1681	699	243	2.9
2	3136	706	51	.6
3	1382	5135	249	3.0
4	367	57	1379	16.7
5	143	205	3033	36.8
6	115	453	1524	18.5
7	60	405	928	11.2
8	44	20	200	2.4
9	15		92	1.1
10	13		22	.3
11	2		29	.4
12	6		4	.0
13	1		8	.1
14	2		1	.0
15	1		4	.0
17			1	.0
43			1	.0
Total	8249	8249	8249	100.0

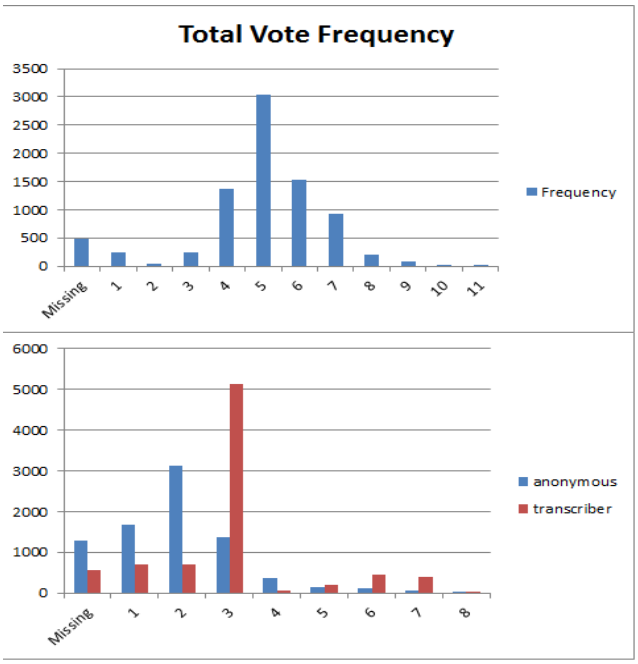


Table 14 provides the vote distribution across the corpus. The table heading "A freq" represents the frequency count of anonymous votes; "T freq" represents transcribers. For example, the table shows 1382 emotion recordings have 3 anonymous votes, and 5135 recordings have 3 transcriber votes. Counting anonymous plus transcriber votes gives the Total column and the Total Vote Frequency chart in Table 14. The skew of the anonymous vote distribution is because the accumulated frequency count was not considered when selecting an anonymous recording. The anonymous recording selection algorithm only ensured no recording was related to twice by the same patient. If overall frequency counts had been considered, the frequency distribution would be flattened.

We will compare the accuracy of the unsupervised anonymous MV classifier to the transcriber classifier once emotional ground truth has been established.

2.7 Fusing MV Classifiers to Establish Emotional Ground Truth

We need as many votes as possible to maximize the certainty of the emotional truth $e_{ij}^{truth}(X_{ij})$ of recording X_{ij} and to satisfy the minimum vote count of 5 for $Steidl_{3:60\%}$. From Table 14, 4.9% of emotional recordings have 5 or more anonymous votes, and 13.1% of recordings have 5 or more transcriber votes. 70.9% of recordings have at least 5 votes if we combine anonymous and transcriber sources. Combining self-reports collected during the prompt "please classify your mood" boosts the total to 87.5%. We leverage classifier fusing [94] to combine multiple MV classifiers. In general, the Fused Majority Vote Classifier formula is:

$$\hat{e} = p(e|c_{1e}, c_{2e} \dots, c_{ne}) = \underset{e \in E}{argmax} \sum_{z=1}^{n} \boldsymbol{w_z} \left[\frac{c_{ez}}{C_{Ez}}\right] \quad (2.8)$$
$$where \quad \sum_{z=1}^{n} \boldsymbol{w_z} = 1, \text{ and } C_{Ez} \neq 0.$$

2.8 Fused Classifier for Automatic Emotion Detection Algorithm Training

All vote sources can be leveraged to establish the emotional ground truth for emotion detection since there are no dependencies on vote sources by the emotion detection training algorithm. The emotion detection training corpus label for recording X_{ij} is computed by approximating the ground truth $e_{ij}^{truth}(X_{ij})$ by fusing the anonymous MV classifier, the transcriber MV classifier, and self-assessment in order to produce the Fused Majority Vote Classifier $e_{ij}^{crowdself}(X_{ij})$.

$$e_{ij}^{crowdself}(X_{ij}) = \underset{e \in E}{argmax} \left[w_1 \frac{c[\, e_{ij}^{relate}]}{C_{Erelate}} + w_2 \frac{c[\, e_{ij}^{transcribe}]}{C_{Etranscribe}} + w_3\, e_{ij}^{self}(X_{ij}) \right] \quad (2.9)$$
$$w_1 + w_2 + w_3 = 1;\ C_{Erelate}, C_{Etranscribe} \neq 0.$$

Equation 2.5 to compute the confidence score is still valid for the fused MV classifier.

2.8.1 Emotional Truth Example

In Table 15, votes have been collected from anonymous voters, transcribers, and a self-report.

Table 15 Crowd-Sourced Vote Collection Example 3

votes	**neutral**	**happy**	**sad**	**angry**	**anxious**	C_E	w_z
Self-assessment		1				1	0.2
anonymous assessments		2		4	1	7	0.4
Transcribers				5	1	6	0.4
$\sum C$	0	3	0	9	2	13	

The weights w_z in equation 2.8 are applied to the MV classifiers. We will set these weights later in section 2.8.2. For the purposes of this example, we set $w_1 = 0.4$, $w_2 = 0.4$, $w_3 = 0.2$ in equation 2.9.

We can calculate $e_{ij}^{crowdself}(X_{ij})$ using equation 2.9:

$$p(X_{ij}|neutral) = \left[0.4\frac{0}{7} + 0.4\frac{0}{6} + 0.2\frac{0}{7}\right] = 0$$

$$p(X_{ij}|happy) = \left[0.4\frac{2}{7} + 0.4\frac{0}{6} + 0.2\frac{1}{7}\right] = 0.143$$

$$p(X_{ij}|sad) = \left[0.4\frac{0}{7} + 0.4\frac{0}{6} + 0.2\frac{0}{7}\right] = 0$$

$$p(X_{ij}|angry) = \left[0.4\frac{4}{7} + 0.4\frac{5}{6} + 0.2\frac{0}{7}\right] = 0.562$$

$$p(X_{ij}|anxious) = \left[0.4\frac{1}{7} + 0.4\frac{1}{6} + 0.2\frac{0}{7}\right] = 0.123$$

The maximum fused MV is angry:

$$e_{ij}^{crowdself}(X_{ij}) = \underset{e=okay,sad,..,anxious}{\operatorname{argmax}} P(X_{ij}|\mathrm{e}) = angry \text{ with } e_{ij}^{confidence}(X_{ij}) = 0.562$$

There are 13 human responses; thus applying equation 2.7 produces:

$e_{ij}^{certainty}(X_{ij})$ = 0.562 x $certainty\_factor$ [6] = 0.562 x 100% = 0.562

2.8.2 Fused MV Classifier Weights calculation

In Multiple Classifier Systems (MCS) the combination weights w_{ij} for each $classifier_j$ is typically determined by minimizing the mean square error of the correlation matrix when compared to the ground truth [95]. ***This approach is not possible since there is no correlation reference for an MCS to approximate ground truth.***

2.8.2.1 Fused MV Classifier Intuitive approach

To determine the weights for each classifier, we compare an Intuitive approach to a Proportional approach. Intuitively, an equal transcriber and anonymous MV classifier weighting makes sense if there is no bias to the transcriber's empathic capability. We assign 50% less weight to self-assessment in order to avoid a dominant contribution, but enough weight to reinforce certainty and break MV score ties. Denote this approach 40-40-20 ($w_1 = 0.4, w_2 = 0.4, w_3$=0.2.). Plugging the weights into equation 2.9 gives:

$$e_{ij}^{crowdself}(X_{ij}) = \underset{e \in E}{argmax} \left[0.4 \frac{c[\, e_{ij}^{relate}]}{C_{Erelate}} + 0.4 \frac{c[\, e_{ij}^{transcribe}]}{C_{Etranscribe}} + 0.2\, e_{ij}^{self}(X_{ij}) \right] \quad (2.10)$$

Using 40-40-20 weighting, we calculate the emotional truth $e_{ij}^{crowdself}(X_{ij})$ and $e_{ij}^{certainty}(X_{ij})$ across the corpus. These weights will be compared to the MV Classifier Proportional weights in section 2.8.2.3.

2.8.2.2 Fused MV Classifier Proportional Weights approach

The Proportional approach assigns Classifier weights based on the overall proportion of votes. There is a dependence on $e_{ijt}^{transcribe}(X_{ij})$ and $e_{ijka}^{relate}(X_{ij})$ on $participant_j$. When a recording X_{ij} is listened to, there may be a bias towards voting for a certain emotion based familiarization with $participant_a$ from previously rated recordings. As such multilevel

variance should be taken into consideration in calculating the proportional weights [68]. The overall $mean$ is calculated considering multilevel variance. As described in section 1.7.13, the expected multilevel population average count L_{ij} given that vote count x_{1ij} is for ESM_{ij} in $participant_j$ is given by:

$$Ln(L_{ij}) = \gamma_0 + \gamma_1 x_{1ij} + U_{0j} \qquad (2.11)$$

This equation can be calculated with glmer() in the R package lme4 [81].

```
> ga <- glmer(avotes~1 + (1|idUsers)
,family = poisson, data=M)
> summary(ga)
Generalized linear mixed model fit by
the Laplace approximation
Formula: avotes ~ 1 + (1 | idUsers)
   Data: M
  AIC  BIC logLik deviance
 7263 7277  -3629     7259
Random effects:
 Groups  Name        Variance Std.Dev.
 idUsers (Intercept) 0.43096  0.65647
Number of obs: 8376, groups:idUsers,
130

Fixed effects:
            Estimate     Std.     Error
Pr(>|z|)
(Intercept)  0.45357    0.06054 <6.81e-
14
> exp(0.45357)
[1] 1.573921
```

```
> gt <- glmer(tvotes~1 + (1|idUsers)
,family = poisson, data=M)
> summary(gt)
Generalized linear mixed model fit by
the Laplace approximation
Formula: tvotes ~ 1 + (1 | idUsers)
   Data: M
  AIC  BIC logLik deviance
 4200 4214  -2098     4196
Random effects:
 Groups  Name        Variance Std.Dev.
 idUsers (Intercept) 0.17734  0.42112
Number of obs: 8376, groups:idUsers,
130

Fixed effects:
            Estimate     Std.     Error
Pr(>|z|)
(Intercept)  1.18704    0.03891 <2e-16
> exp(1.18704)
[1] 3.277366
```

Code Snippet 23 Calculation of Multilevel Vote Count Means in R

We apply equation 2.11 :

- For anonymous votes $Ln(L_{ij}) = 0.45357$; multilevel mean = $e^{0.45357} = 1.573921$
- For transcriber votes $Ln(L_{ij}) = 1.18704$; multilevel mean =$e^{1.18704} = 3.277366$
- For self-report votes, var(self-count) = 0, therefore mean $= 1$

Total = 1.573921 + 3.277366 + 1 = 5.851287

Calculating proportions gives:

$$w1 = \frac{1.573921}{5.851287} = 0.27\ ;\ w2 = \frac{3.277366}{5.851287} = 0.56\ ;\ w3 = \frac{1}{5.851287} = 0.17$$

Plugging the weights into equation 2.9 gives:

$$e_{ij}^{crowdself}(X_{ij}) = \underset{e\in E}{argmax}\left[0.27\frac{c[e_{ij}^{relate}]}{C_{Erelate}} + 0.56\frac{c[e_{ij}^{transcbe}]}{C_{Etranscribe}} + 0.17\, e_{ij}^{self}(X_{ij})\right] \quad (2.12)$$

Denote this $e_{ij}^{crowdself}(X_{ij})$ fused MCS approach 27-56-17.

2.8.2.3 40-40-20 versus 27-56-17 weighting

We want weighting that maximizes certainty across the corpus.

Table 16 compares the number of ESMs for $e_{ij}^{certainty}(X_{ij}) \geq X, X = 0..1$, for both sets of weights, there are more ESMs at higher certainty levels (0.5 to 0.9) with 27-56-17 weighting.

Table 16 ESM Certainty: 40-40-20 versus 27-56-17 Weights

certainty	40-40-20		27-56-17		difference	
	ESMs	**% ESMs**	**ESMs**	**% ESMs**	**ESMs**	**% ESMs**
0.0	8376	100.00%	8376	100.00%	0	0.00%
0.1	8376	100.00%	8376	100.00%	0	0.00%
0.2	8376	100.00%	8376	100.00%	0	0.00%
0.3	8368	99.90%	8362	99.80%	-6	-0.10%
0.4	7601	90.70%	7544	90.10%	-57	-0.60%
0.5	6712	80.10%	6809	81.30%	97	1.20%
0.6	4877	58.20%	5487	65.50%	610	7.30%
0.7	3916	46.80%	3957	47.20%	41	0.40%
0.8	3003	35.90%	3199	38.20%	196	2.30%
0.9	1343	16.00%	1510	18.00%	167	2.00%
1.0	636	7.60%	611	7.29%	-25	-0.31%

Code Snippet 24 indicates 91.57% concordance of $e_{ij}^{crowdself}(X_{ij})$ between 40-40-20 and 27-56-17 weighting. Sad & Angry correlate most; Happy & Anxious correlate least.

```
> confusionMatrix(df442$cs_amax, df1756$cs_amax)
Confusion Matrix and Statistics
          27-56-17
  40-40-20   ok hap  sad  angry anxious
        ok 3481  126  152   19   20
     happy  219 2067   19    4    6
       sad   23    2  814   17    4
     angry   13    1    8  522    2
   anxious   20    7   27   17  466

  Accuracy : 0.9157    95% CI : (0.9096, 0.9216) P-Value [Acc > NIR] : < 2.2e-16
Statistics by Class:
                     ok       happy    sad     angry    anxious
Pos Pred Value       0.9165   0.8929  0.94651  0.95604  0.86778
```

Code Snippet 24 Concordance of 40-40-20 versus 27-56-17 weighting in R

Interestingly, there is more concordance between the anonymous MV classifier $\frac{c\left[e_{ij}^{relate}\right]}{C_{Erelate}}$ with the transcriber MV classifier $\frac{c\left[e_{ij}^{transcribe}\right]}{C_{Etranscribe}}$ with 40-40-20 weighting, however maximization of certainty takes precedence.

Table 17 Majority Vote Concordance: 40-40-20 versus 27-56-17 Weights

certainty	**anonymous MV = transcriber MV concordance**		difference
	40-40-20	**27-56-17**	
0.0	60.07%	60.07%	0.00%
0.1	60.07%	60.07%	0.00%
0.2	60.07%	60.07%	0.00%
0.3	60.10%	60.14%	0.04%
0.4	61.03%	61.66%	0.63%
0.5	65.48%	64.46%	-1.02%
0.6	79.41%	72.56%	-6.85%
0.7	80.26%	81.29%	1.03%
0.8	91.49%	88.03%	-3.46%
0.9	92.32%	93.25%	0.93%
1.0	100.00%	100.00%	0.00%

The results from Table 16 and Table 17 are summarized in Figure 67. Both weight sets will be tested for emotion detection accuracy performance in chapter 3.

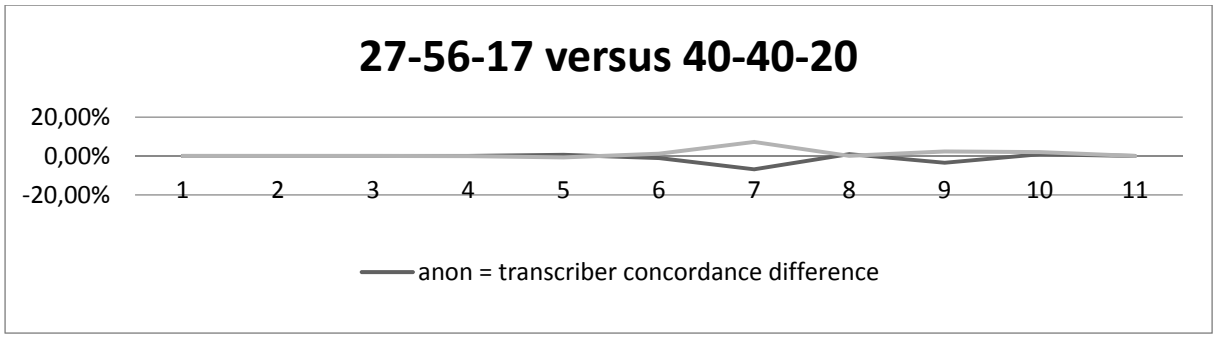


Figure 67 MV Concordance Differences of 40-40-20 versus 27-56-17

2.9 Fused Classifier for Statistical Analysis

A major aspect of emotional health is to compare self-assessment to the emotional ground truth; thus the approximation $e_{ij}^{crowd}(X_{ij})$ does not include the self-assessment $e_{ij}^{self}(X_{ij})$ in the fused MCS in order to respect the independence of the variables.

$$e_{ij}^{crowd}(X_{ij}) = \underset{e\in E}{argmax} \left[w_1 \frac{c[\, e_{ij}^{relate}]}{C_{Erelate}} + w_2 \frac{c[\, e_{ij}^{transcribe}]}{C_{Etranscribe}} \right] \qquad (2.13)$$
$$w_1 + w_2 = 1;\ C_{Erelate}, C_{Etranscribe} \neq 0.$$

From Code Snippet 23 we can calculate the proportional weighting for $e_{ij}^{crowd}(X_{ij})$:

Total = 1.573921 + 3.277366 = 4.851287

$$w1 = \frac{1.573921}{4.851287} = 0.33\ ;\ w2 = \frac{3.277366}{4.851287} = 0.67$$

Plugging the weights into equation 2.13 gives:

$$e_{ij}^{crowd}(X_{ij}) = \underset{e\in E}{argmax} \left[0.33 \frac{c[\, e_{ij}^{relate}]}{C_{Erelate}} + 0.67 \frac{c[\, e_{ij}^{transcribe}]}{C_{Etranscribe}} \right] \qquad (2.14)$$

2.10 Unsupervised Anonymous MV Classifier Accuracy

Unsupervised emotional truth corpus labeling requires automatic chunking of audio into an utterance with a single emotion, and unsupervised automatic emotional truth labeling. We have automatic chunking as described in section 2.1. We have developed optimized MV classifiers $e_{ij}^{crowdself}(X_{ij})$ and $e_{ij}^{crowd}(X_{ij})$ for emotion detection algorithm training and statistical analysis respectively.

The question is, given enough votes, is the MV classifier $e_{ij}^{relate}(X_{ij}) = \underset{e \in E}{argmax} \left[\frac{c\left[e_{ij}^{relate} \right]}{C_{Erelate}} \right]$ reliably accurate? This would give true unsupervised emotional truth corpus labeling. We compare the anonymous MV classifier $e_{ij}^{relate}(X_{ij})$ to the reference emotional truth classifier $e_{ij}^{crowdself}(X_{ij})$. There are 2132 recordings in the corpus with 3 or more anonymous votes $e_{ijka}^{relate}(X_{ka})$. The accuracy of $e_{ij}^{relate}(X_{ij})$ is 70.59% as calculated in Code Snippet 25.

```
> EMOa3 <- read.csv(file = "EMO_DATA_AUG2012/275617_atleast3anonVotes.csv")
> confusionMatrix(EMOa3$anonEMO, EMOa3$cs_amax)
Confusion Matrix and Statistics
          Reference
Prediction   ok   happy   sad   angry  anxious
        ok  717     136   132      27  25
     happy  166     464    10       4   4
       sad   18       2   108       8   2
     angry    6       2     6     105   1
   anxious   24      14    18      22 111
Overall Statistics
               Accuracy: 0.7059
                 95% CI: (0.6861, 0.7252)
```

Code Snippet 25 Concordance of Anonymous MV Classifier (C>3) versus Truth in R

There are 764 recordings with 4 or more anonymous relate votes. The accuracy of $e_{ij}^{relate}(X_{ij})$ is 70.03%, 95% CI: (0.6664, 0.7326). There are 399 recordings with 5 or more anonymous relate votes. The accuracy of $e_{ij}^{relate}(X_{ij})$ is 70.03%, 95% CI: (0.6664, 0.7326).

Table 18 Accuracy of the Anonymous MV Classifier

Minimum votes	Number of recordings	Accuracy
3	2132	70.59%
4	764	70.03%
5	399	69.42%
		70.01% (mean)

The 70% accuracy is insufficiently reliable to depend on unsupervised anonymous majority voting to label a corpus. However, the 70% accuracy does indicate a high degree of statistical power and as such does add to the certainty of the fused MV classifier. As a caveat, anonymous votes originating from participants includes patients who may have diminished capability to empathize with another human being. This is hypothesized to account for the lower accuracy.

2.11 Conclusion

We have developed optimized fused MV classifiers $e_{ij}^{crowdself}(X_{ij})$ and $e_{ij}^{crowd}(X_{ij})$ to approximate the emotional truth of an audio recording for emotion detection algorithm training and statistical analysis respectively. We have two sets of weights for the MV classifier $e_{ij}^{crowdself}(X_{ij})$; 40-40-20 and 27-56-17, which will both be tested in chapter 3.

We have analysed the feasibility of unsupervised emotional truth corpus labeling that requires automatic chunking of audio into an utterance with a single emotion, and unsupervised automatic emotional truth labeling. Automatic chunking is implemented and verified. At 70% accuracy, it is hypothesized that unsupervised automatic emotional truth labeling has a dependency on the empathic capability of the anonymous voters. Empathy differences between trial groups will be measured in CHAPTER 5.

CHAPTER 3

AUTOMATIC EMOTION DETECTION IN SPEECH

In order to monitor and analyze the emotional health of a patient, we must develop algorithms to accurately measure the emotional state of a patient in their natural environment. The preferred approach to emotional truth determination is automatic real-time emotion detection in speech as this will provide instantaneous results. This section describes the automatic emotion detection algorithm design and results.

Table 19 and Figure 68 describe the process to develop emotional state algorithms. Step 1, 2, and 4 (described in the Methodology section on page 16) have resulted in a labeled corpus of 8,249 emotions that enable automatic emotion detection experimentation.

Table 19 Emotional State Algorithm Development Process Steps

Step	Purpose	Description
1	Data Collection	Sample a patient's emotional state in their natural environment.
2	Data Collection	Capture the emotional state to a secure cloud database.
4	Data Collection	Collect emotional states and label the emotional truth to establish a corpus for algorithm development.
5	Measure Emotional Health	Develop emotion measurement algorithms.

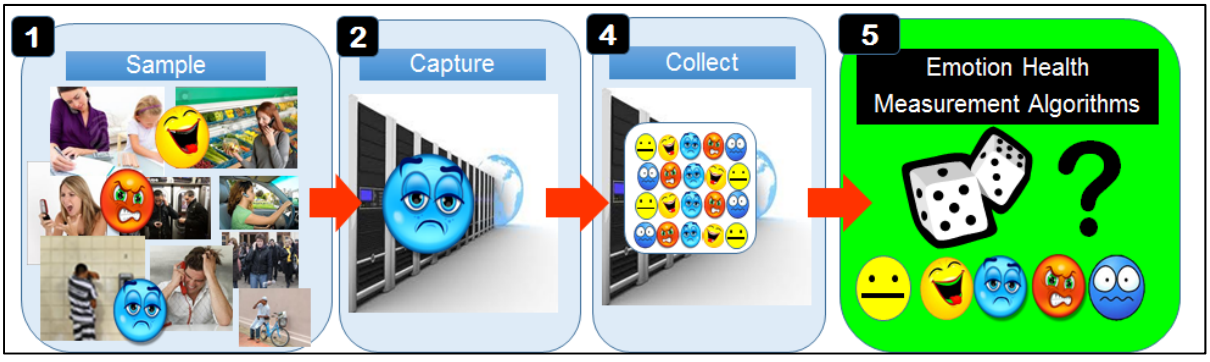


Figure 68 Emotion Classification Algorithm Development Process Flow

3.1 Automatic Emotion Detection to Approximate Emotional Truth

We attempt to automatically calculate $e_{ij}^{detect}(X_{ij})$ as an approximation of $e_{ij}^{truth}(X_{ij})$ through automatic acoustical emotion detection. Automatic emotion detection in speech consists of extracting features from speech and then classifying the features to an emotion. There are two distinct phases in automatic emotion detection: The acoustical model training phase, and the run-time automatic emotion detection phase. Each stage of the training phase will be described in this chapter. An accurate $e_{ij}^{detect}(X_{ij})$ would enable real-time, automatic, and complete emotional health measurements on $participant_j$ during call c_{ij}; a capability offline MV classifiers $e_{ij}^{crowdself}(X_{ij})$ and $e_{ij}^{crowd}(X_{ij})$ cannot perform.

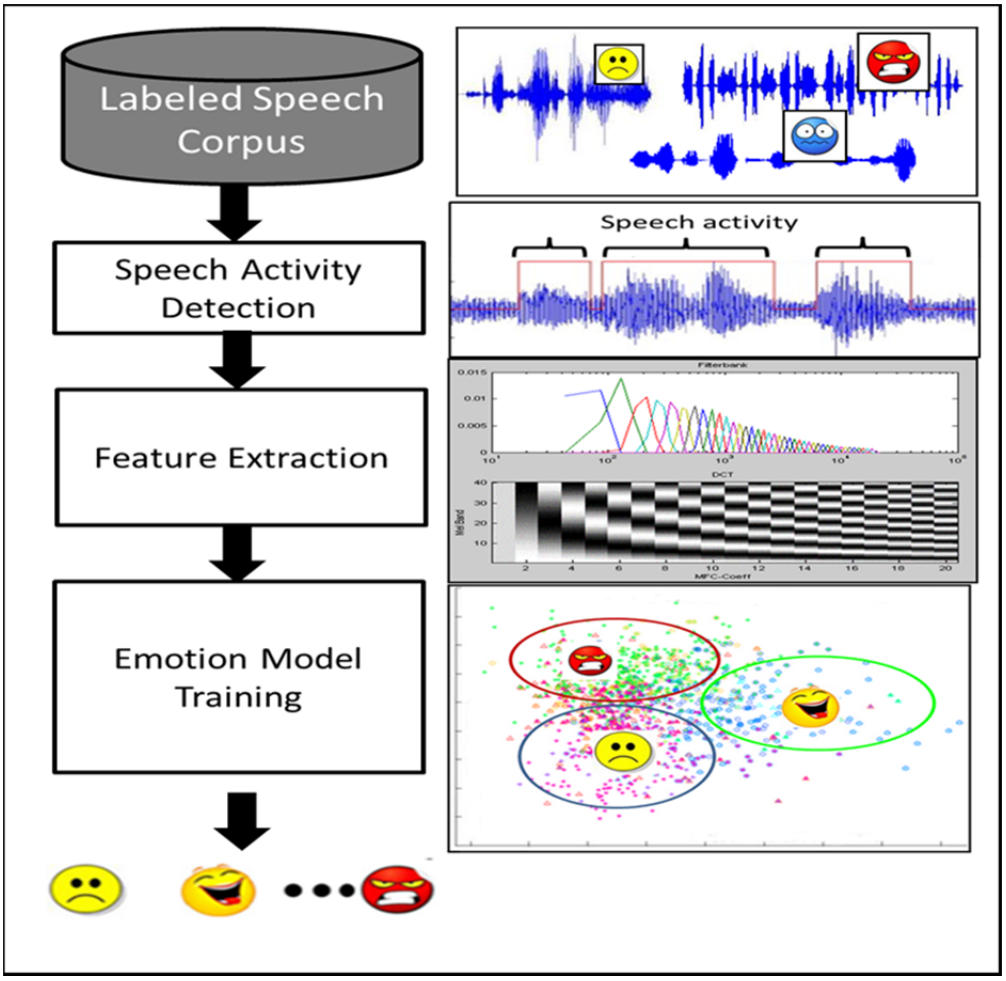


Figure 69 Emotion Model Training

Figure 69 depicts the stages of constructing emotion models for the emotion detector. The corpus, consisting of emotion recordings and their corresponding labels computed by the MV Classifier $e_{ij}^{crowdself}(X_{ij})$, are processed to compute the emotion models for Neutral, Happy, Sad, Angry and Anxious. The stages consist of speech activity detection where silence and non-speech is removed from the recording, feature extraction and calculation of MFCCs, and emotion model training.

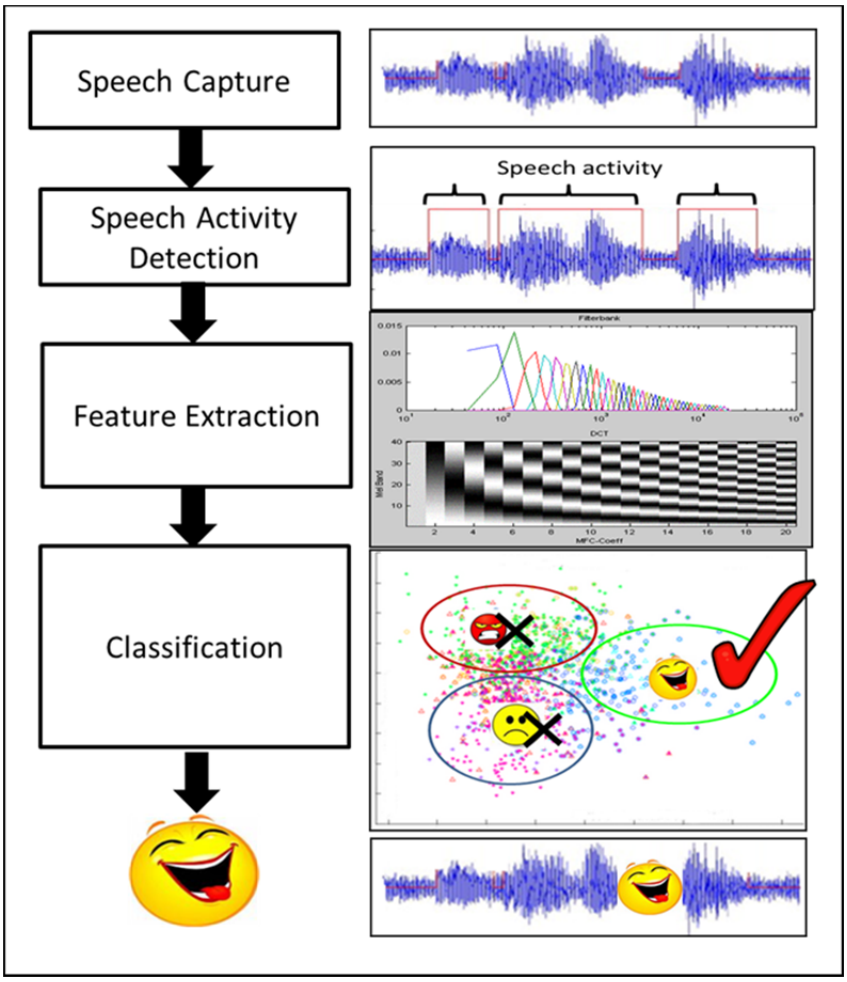


Figure 70 Run-Time Emotion Detection

Figure 70 depicts emotion detection of a captured speech recording. Speech activity detection and feature extraction are identical to the emotion model training phase. The extracted features are mapped to the most likely emotion model in the final classification stage.

3.2 State-of-the-art in emotion detection

Nwe et al. [42] conducted experiments to measure the performance of human classification of utterances into six classes: (Anger, Dislike, Fear, Joy, Sadness, Surprise). The average performance was 65.7%. The language of the utterances presented to the human subject was neither his mother tongue nor any other language that he has any knowledge to perceive linguistically; thus assuring only acoustic features were considered.

There is considerable research activity in emotion detection in speech. Approaches and results vary widely. The reader is directed to a survey of audio and visual emotion detection methods performed by Zeng et al. [96] in 2009.

The Open Performance Sub-Challenge [34] Prize was awarded to Pierre Dumouchel et al. [44] for their victory in this sub-challenge: they had managed to obtain the best result (70.29% UA recall) in the two-class task, significantly ahead of their eight competitors. The best result in the five-class task (41.65% UA recall) was achieved by Marcel Kockmann et al. [97].

The five-class emotion set for the competition [34] was (Angry, Emphatic, Neutral, Positive and Rest). This problem is similar to the emotion set for this thesis (Happy, Sad, Angry, Neutral, and Anxious) which has been tailored towards mood and anxiety disorders.

Kockmann et al. [34] extracted 13 MFCCs including C0 log energy, its first derivative $\Delta MFCC(n)$ and second derivative $\Delta^2 MFCC(n)$. Kockmann also used the third derivative $\Delta^3 MFCC(n)$ and discarded non-voiced frames. Other features like Shifted Delta Cepstra and Syllable Contours were experimented, but Kockmann et al concludes: "appropriate feature type still has to be found". EM and MAP adaptation was used to train the Gaussian Model Mixtures (GMM) Universal Background Models (UBM), and Joint Factor Analysis (JFA) was used to "cope with the problem of speaker and session variability in GMM-based speaker verification". However, a result of 40.8% recall was achieved using

the GMM-UBM approach alone. Instead of frame-based full log-likelihood evaluation, approximate fast linear scoring based on utterance statistics was used.

Dumouchel et al. [34] achieved 39.40% UA recall on the 5-class problem removing silence a priori, with similar features as Kockmann et al (12 MFCC + log energy, $\Delta MFCC$ and $\Delta^2 MFCC$, GMMs λ_i trained using EM, and ML, $\mathrm{E} = \underset{\mathrm{i=1..5}}{\arg\max} \log \mathrm{P}(X|\lambda_i)$ scoring.

Phonemes can be trained on the acoustic features in order to extract information on language content, speech rate, and pauses. A proven approach to phoneme training and recognition is based on GMMs & 3-state Hidden Markov Models [98] (HMM). Acoustic features can be augmented with the probability of a word occurring during the expression of emotions in a Support Vector Machine (SVM) vector [99]. Once the features have been selected, one can look at emotion classification model training. Bayes probabilistic classification is a good approach [98], as are SVM and K-Nearest Neighbors [6].

3.3 Speech Activity Detection

The Speech Activity Detector [100] removes silence and non-speech from the recording prior to feature extraction, model training, and test. Experiments were performed to adjust parameters with the goal of ensuring no valid speech recordings were discarded (e.g. the response utterance “ok” can be as short as 0.2 seconds), and the GMM emotion detector’s accuracy was maximized. Nominal signal level was set to -45 dB and noise level was set to -50 dB to perform in both noisy cellular phone and landline phone conditions.

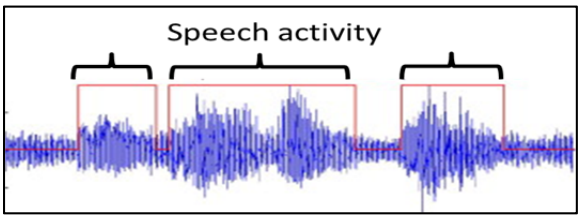


Figure 71 Speech Activity Detection

The minimum utterance duration was 0.05 seconds to capture short utterance like “ok”. The intra-speech silence was set at 0.2 seconds.

3.4 Feature Extraction

MFCCs are calculated from the log filterbank amplitudes using ***HCopy()*** command in the Hidden Markov Models Toolkit [101] (HTK). The human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves recognition performance. The Fourier transform based triangular filters are and equally spaced along the mel-scale which is defined by $Mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$

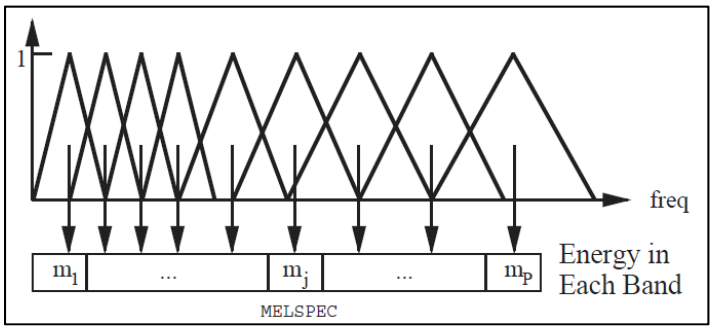


Figure 60 Mel-Scale Filter Bank

MFCCs are calculated from the log filter bank amplitudes using the Discrete Cosine Transform $c_i = \sqrt{\frac{2}{N}} \sum_j^N m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right)$ where N is the number of filter bank channels.

A sequence of MFCC feature vectors $X = \{ x_1, x_2, \ldots, x_T \}$ where x_i consists of 60 features including MFCCs + log energy C0, the 1st derivative $\Delta MFCC(n)$ and the second derivative $\Delta^2 MFCC(n)$ are extracted from the speech recording using a 25 millisecond Hamming window and a frame advance of 10 milliseconds. Key features are the fundamental frequency F0 and the normalized energy C0 [44].

3.5 Emotion Detection Algorithm

Our approach to automatic emotion detection in speech is inspired from Dumouchel *et al*. [44] and consists of extracting MFCCs and energy features from speech and then classifying these acoustic features to an emotion. A large GMM referred to as the UBM, which plays the role of a prior for all emotion classes, was trained on the emotional corpus of 8,376 speech recordings using the EM algorithm. After training the UBM, we adapted it to the acoustic features of each emotion class using the MAP algorithm. As in Reynolds *et al*. [102] we used MAP adaptation rather than the Maximum Likelihood Linear Regression (MLLR) algorithm [103] because we had very limited training data for each emotion class (which increased the difficulty of separate training of each class GMM).

The probability of observing a feature vector x_t from a given GMM ($p(x_t|\lambda) = \sum_{i=1}^{C} w_i\, p_i(x_t)$ or alternatively $p(x_t|\lambda) = \sum_{i=1}^{C} w_i\, \aleph\{x_t: \mu_i, \Sigma_i\}$) is a weighted combination of $\boldsymbol{C}$ Gaussian densities $\boldsymbol{p_i}(x_t)$**,** where each Gaussian is parameterized by a mean vector $\boldsymbol{\mu_i}$ of dimension $\boldsymbol{d}$ and a covariance matrix $\boldsymbol{\Sigma_i}$ is given by**:**

$$p_i(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_t-\mu_i)'(\Sigma_i)^{-1}(x_t-\mu_i)} \qquad (\ 3.1\ )$$

The mixture weights $\boldsymbol{w_i}$ must satisfy the condition $\sum_{i=1}^{C} w_i = 1$. Each emotion class $\boldsymbol{e_m}$ is represented by a single GMM. Each GMM is trained on the data from the same emotion class using the expectation-maximization algorithm [42]. The feature vectors x_t are assumed to be independent; therefore the log likelihood for each emotion model $\boldsymbol{e_m}$ is:

$$\log p(X|e_m) = \sum_{t=1}^{T} log\, p(x_t|e_m) \qquad (\ 3.2\ )$$

where T is the length of the utterance.

There is limited data for each class in the corpus. The K-fold cross-validation algorithm [104] (K=10) was used for model training and test due to the small corpus size. To compensate further, the MAP adaptation approach was used to build the GMM models. One large GMM named UBM was trained. The UBM GMM was then adapted to each emotion class. UBM GMM MAP adaptation is summarized in Figure 61.

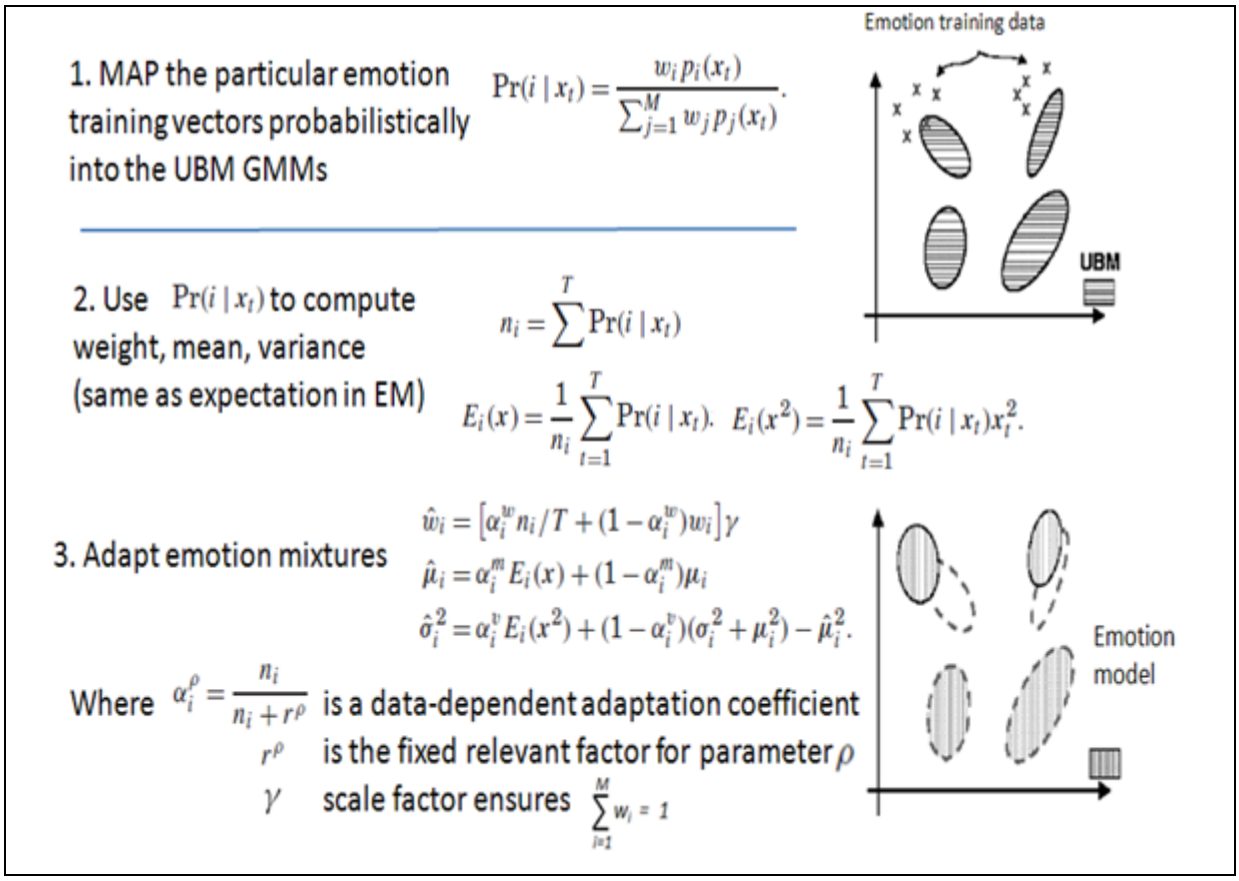


Figure 72 MAP Adaptation (Summarized from Reynolds et al)

Reynolds et al. [105] describe the advantages of MAP:

- Data-dependent adaptation coefficient allows a mixture-dependent adaptation of parameters. This approach is more robust for limited training data.

- Not all Gaussians in the UBM are adapted during speaker model training. As such, storage can be reduced by storing only the difference between the speaker model and the UBM.
- The log-likelihood ratio $\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{UBM})$ is a faster scoring method than computing $(X) = \log p(X|\lambda_{hyp}) - \log\{\frac{1}{B}\sum_{b=1}^{B} p(X|\lambda_b)$ (B is the number of Background models).

The Naïve Bayes theorem is applied with equal emotion class weights in order to calculate the maximum likelihood that an utterance $\boldsymbol{X}$ corresponds to the emotion $\boldsymbol{e}$. The posterior distribution of each class e given the utterance X can be simplified as follows:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}} \Pr(e|X) = \underset{e \in E}{\operatorname{argmax}} \Pr(X|e)\Pr(e)$$
$$e \in E\{Neutral, Happy, Sad, Angry, Anxious\} \qquad (3.3)$$

For each emotion $\boldsymbol{e}$, $\Pr(e)$ can be calculated from the frequency of occurrence of each emotion in the speech training data.

Table 20 Pr(e) Calculation

emotion label	**recordings**	**Pr(e)**
okay	3757	47%
happy	2205	27%
sad	1023	13%
angry	566	7%
anxious	510	6%

With the small number of negative emotions, Pr(e) overly biased towards positive emotions as such, Pr(e) was removed from the equation, and all experiments were run with equal emotion weighting, reducing equation 3.3 to:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}} \Pr(e|X) = \underset{e \in E}{\operatorname{argmax}} \Pr(X|e) \qquad (3.4)$$

3.6 GMM Model Training

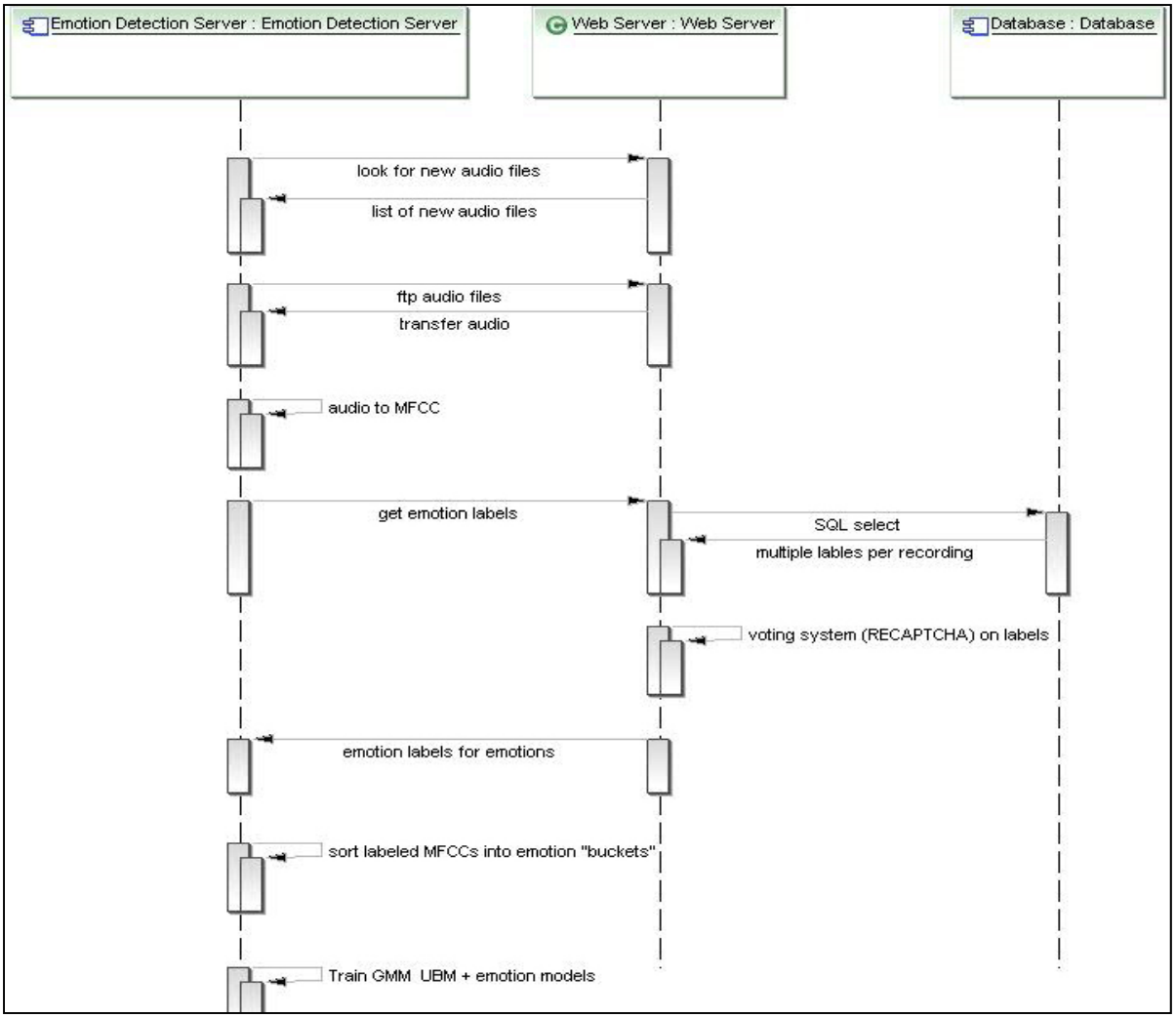


Figure 73 Emotion Detector Training Sequence Diagram

The sequence diagram in Figure 73 depicts the sequence of events, encapsulated in Perl and PHP scripts, to train emotional models. A Cron daemon periodically looks for new ESMs on the www.emotiondetect.com , www.emosub.com , and www.emotoolkit.com web servers; which calls participants and collects ESMs. The MV Classifier $e_{ij}^{crowdself}(X_{ij})$ determines the highest probability emotional label. The emotion labels are used to sort the corresponding audio into emotion subdirectories or "buckets". The GMM training algorithm then computes the emotional models.

3.6.1 HTK GMM Training

The ***gmm_trn()*** from the HTK toolkit [101] is embedded in a Perl script to calculate GMMs from the labeled MFCCs. A first pass generates the UBM, and a second pass generates the emotion models adapted from the UBM.

There are many parameter options in ***gmm_trn()***. The number of iterations to converge the algorithm, the number of Gaussians in the mixture, and the Adaptation algorithm are the key parameters. Adaptation algorithms available in HTK version 3.4 are MAP, MLLR, and MAP tree algorithms. Models were generated for 128, 256, and 512 Gaussians using MAP and MLLR adaptation. Results are summarizes in the next section.

3.6.2 Parallel Processing

The Perl library Proc::ParallelLoop [106] allowed emotion GMM training to run on 5 processors on the training computer's Central Processing Unit (CPU) (one for each emotion model) to reduce compute time.

MAP adaptation with parameters set for a minimum of 25 iterations, 10% standard deviation threshold, and 512 Gaussian mixtures, required 14 hours to compute 1 world GMM and 10 GMMs per emotion (K-fold training) sequentially versus 20 hours in parallel on an Intel i7 K875 8-processor core running at 2.93 GHz with 16 GBytes of memory.

3.7 GMM Emotion Detection

The computed emotional models are then uploaded to each web server (www.emotiondetect.com, www.emosub.com, and www.emotoolkit.com). The HTK command ***gmm_llk()*** [101] is executed in a PERL script on the web server to compute the scores of $\Pr(X_{ij}|e)$ for each emotion GMM. The highest score is selected as $e_{ij}^{detect}(X_{ij})$.

3.8 Experimentation and Accuracy of edetect(X)

Emotion detection experiments were conducted with 128, 256, and 512 Gaussian mixtures per emotion model; MAP and MLLR adaptation; UBM; 27-56-17 and 40-40-20 fused MCS w1-w2-w3 weighting to compute $e_{ij}^{crowdself}(X_{ij})$. $e_{ij}^{certainty}(X_{ij})$ thresholding was abandoned in order to maximize the corpus set. Emotion detection results $e_{ij}^{edetect}(X_{ij})$ are compared to $e_{ij}^{crowdself}(X_{ij})$. The function ***confusionMatrix()*** from the R package caret [107] computes the experimental results. A heat map is also provided where each cell's greyscale is weighted by the corresponding confusion matrix's percentage (black is high, white is low). Table 21 describes how to interpret the confusion matrix, and Table 22 describes the measures computed.

Table 21 Confusion Matrix Definition

		Reference	
		Emotion 1	**Emotion 2**
Predicted	**Emotion 1**	True Positive emotion 1 (TP_1)	False Positive emotion 1 (FP_1)
	Emotion 2	False Positive emotion 2 (FP_2)	True Positive emotion 2 (TP_2)

Table 22 Confusion Matrix Results Interpretation

Measure	**Formula**	**description**
Precision	$\frac{TP_e}{TP_e + FP_e}$	percentage of correct positive predictions for emotion e.
Accuracy	$\frac{\sum TP}{\sum TP + \sum FP}$	Total percentage of predictions correct.
95% CI	Confidence Interval calculated with the binomial exact test and a one-sided test to see if the accuracy is better than the "no information rate" which is taken to be the largest class percentage in the data.	
Unweighted Kappa	$\kappa = \frac{\text{Accuracy} - P_{chance}}{1 - P_{chance}}$	Agreement adjusted for that expected by chance [108].
P-value	McNemar's test evaluates changes in related or paired binomial attributes, whether changes in one direction is significantly greater than the opposite direction [109].	

Figure 74 MAP-UBM: 27-56-17 Classifier Concordance Heat Map

Table 23 MAP-UBM: 27-56-17 Classifier Precision and Accuracy

Accuracy : 41.92% 95% CI : (0.408, 0.430) No Info Rate : 0.466 Kappa : 0.2288 P-Value : <2e-16	

	okay	happy	sad	angry	nervous
okay	64%	21%	8%	4%	4%
happy	39%	47%	4%	5%	4%
sad	45%	15%	29%	4%	7%
angry	34%	27%	6%	23%	9%
nervous	36%	26%	11%	11%	17%

	okay	happy	sad	angry	nervous
Precision	63.8%	47.3%	29.0%	23.2%	16.7%

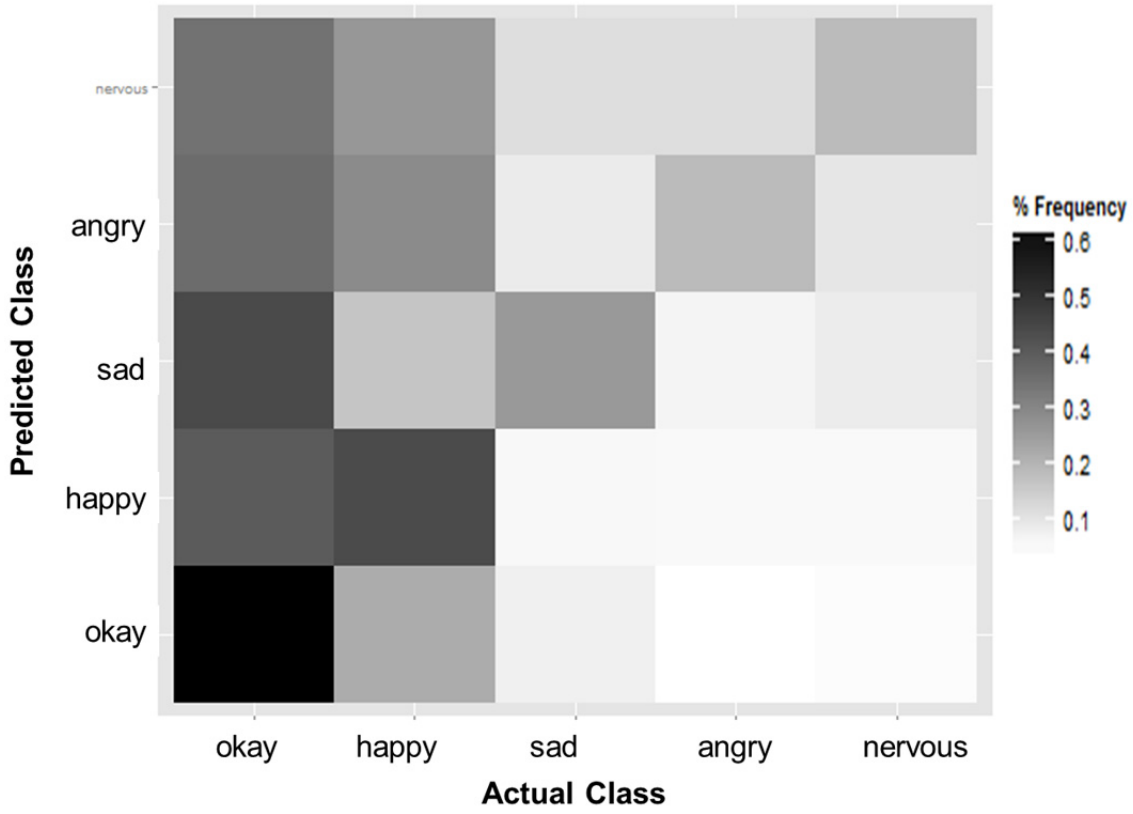


Figure 75 MLLR-UBM: 27-56-17 Classifier Concordance Heat Map

Table 24 MLLR-UBM: 27-56-17 Classifier Precision and Accuracy

Accuracy : 35.13%
95% CI : (0.341, 0.362)

No Info Rate : 0.466
Kappa : 0.1913
P-Value : <2e-16

	okay	happy	sad	angry	nervous
okay	68%	22%	6%	2%	3%
happy	42%	47%	4%	4%	3%
sad	47%	16%	26%	5%	7%
angry	38%	30%	7%	18%	8%
nervous	36%	27%	10%	10%	18%

	okay	happy	sad	angry	nervous
Precision	67.8%	46.7%	26.2%	17.8%	17.9%

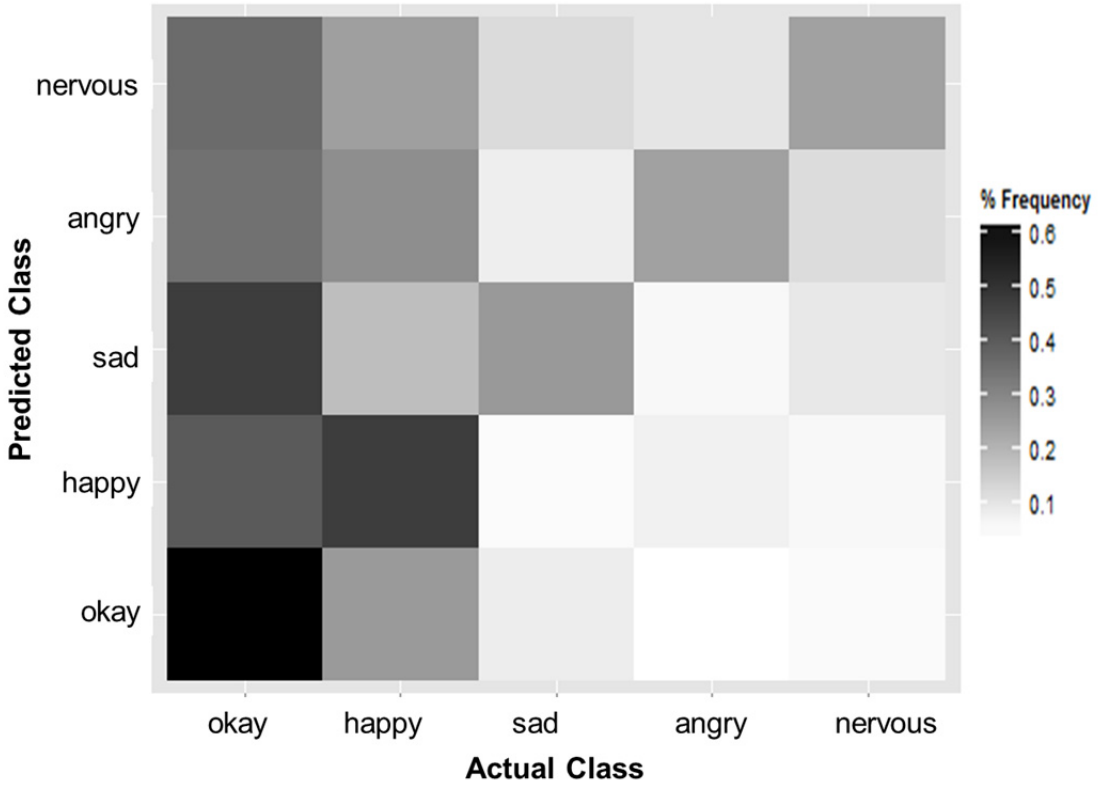


Figure 76 MAP-UBM: 40-40-20 Classifier Concordance Heat Map

Table 25 MAP-UBM: 40-40-20 Classifier Precision and Accuracy

Accuracy :	41.13%
95% CI :	(0.400,0.422)
No Info Rate :	0.4711
Kappa :	0.2136
P-Value :	<2e-16

	okay	happy	sad	angry	nervous
okay	63%	24%	6%	3%	4%
happy	40%	47%	4%	6%	4%
sad	47%	17%	25%	4%	8%
angry	34%	27%	6%	23%	10%
nervous	35%	23%	10%	8%	23%

	okay	happy	sad	angry	nervous
Precision	62.6%	47.1%	16.5%	22.9%	22.8%

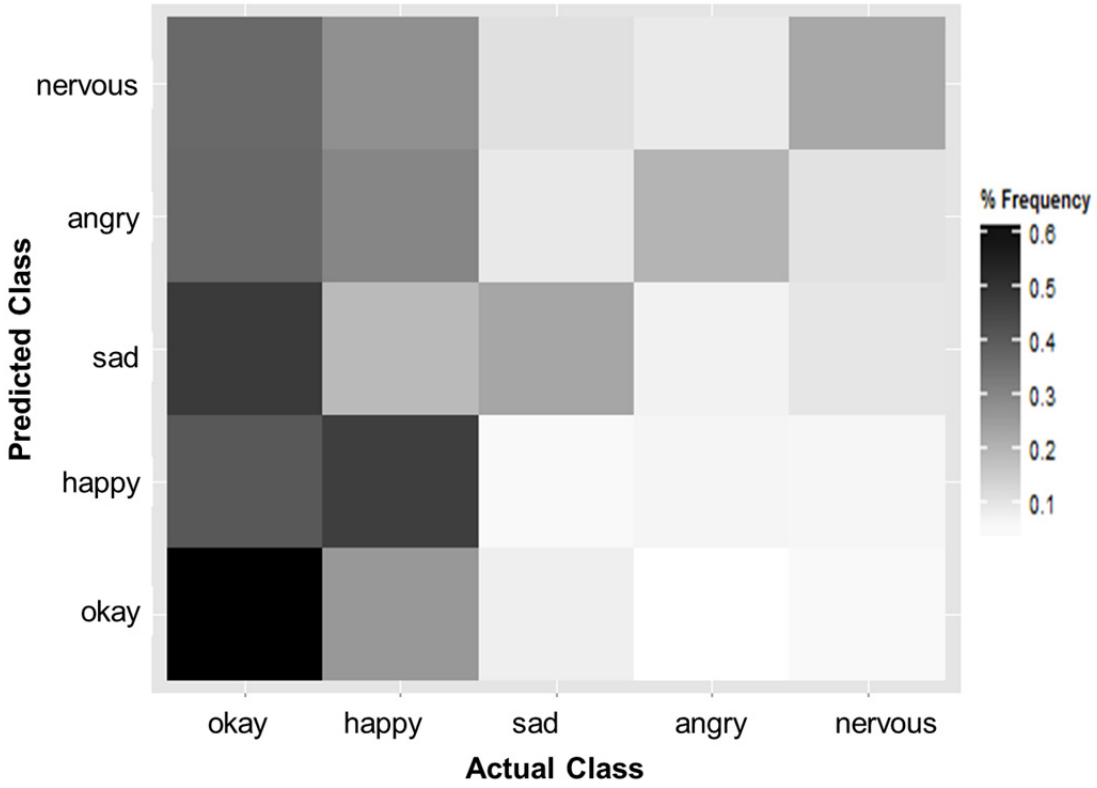


Figure 77 MLLR-UBM: 40-40-20 Classifier Concordance Heat Map

Table 26 MLLR-UBM: 40-40-20 Classifier Precision and Accuracy

Accuracy : 35.08%
95% CI : (0.340, 0.361)

No Info Rate : 0.4711
Kappa : 0.1838
P-Value : <2e-16

	okay	happy	sad	angry	nervous
okay	65%	25%	5%	2%	3%
happy	41%	48%	3%	4%	4%
sad	49%	17%	22%	5%	7%
angry	37%	30%	7%	19%	8%
nervous	37%	27%	9%	6%	21%

	okay	happy	sad	angry	nervous
Precision	64.9%	48.2%	21.9%	18.6%	21.4%

3.8.1.1 Self-assessment compared to Emotional Truth (eCROWD)

Self-assessments $e_{ij}^{self}(X_{ij})$ are compared to $e_{ij}^{crowd}(X_{ij})$ to provide contrast for the emotion detector results. We do not use $e_{ij}^{crowdself}(X_{ij})$ in order to avoid bias.

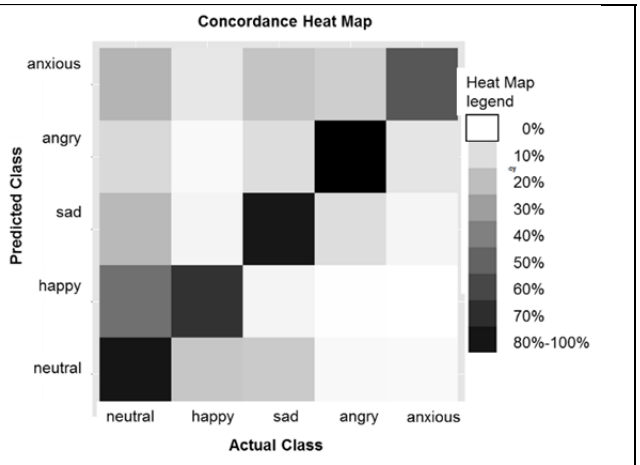
Table 27 Self-Assessment Concordance

Pre/Act	neutral	happy	sad	angry	anxious
neutral	66%	15%	14%	3%	2%
happy	38%	57%	3%	1%	1%
sad	18%	4%	66%	9%	3%
angry	10%	2%	9%	72%	7%
anxious	20%	7%	16%	13%	45%

Accuracy : 61.66%
95% CI : (0.6056, 0.6276)

No Information Rate : 0.4616
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4415
Mcnemar's Test P-Value : < 2.2e-16



	Class:okay	Class:happy	Class:sad	Class:angry	Class:anxious
Precision	66.44%	56.59%	66.03%	71.57%	45.19%

3.8.1.2 Results summary and analysis

Table 28 summarizes results for the 4 emotion detectors along with the self-assessments.

Table 28 Accuracy of Emotion Detection Algorithms and Self-Assessment

		Precision				
512 Gaussian detector	**Accuracy**	**Neutral**	**Happy**	**Sad**	**Angry**	**Anxious**
27-56-17 MAP	**62.58%**	**85.15%**	**70.02%**	**36.86%**	**45.03%**	**71.97%**
27-56-17 MLLR	51.88%	86.02%	67.04%	31.36%	30.49%	59.45%
40-40-20 MAP	61.60%	85.08%	68.44%	33.48%	45.83%	69.09%
40-40-20 MLLR	50.94%	84.51%	65.01%	27.58%	31.56%	59.58%
SELF-Assessment	61.66%	66.44%	56.59%	66.03%	71.59%	45.19%

As predicted in section the weights 27-56-17 is more accurate than the intuitive 40-40-20 weighting for both MAP (+0.98%) and MLLR (+0.94%).

Precision for emotion classification is directly proportional to the number of speech samples collected, and it is speculated that collecting more samples will improve precision.

Table 29 Proportion of emotional speech samples collected

total	okay	happy	sad	angry	nervous
8041	3788	2313	859	545	536
	47.1%	28.8%	10.7%	6.8%	6.7%

3.9 Conclusion

After experimentation with 128, 256, and 512 Gaussian mixtures per emotion model; MAP-UBM and MLLR-UBM adaptation; and weights of 40-40-20 ($w_1 = 0.4, w_2 = 0.4, w_3$=0.2.) and 27-56-17 for the labeling of the emotion training corpus by the MV classifier $e_{ij}^{crowdself}(X_{ij})$; we have an automatic emotion detector $e_{ij}^{detect}(X_{ij})$ that is 41.92% accurate (UA recall) with precision of Neutral=64%, Happy=47%, Sad=29%, Angry=23%, Anxious=16%.

Is the $e_{ij}^{detect}(X_{ij})$ performance at 41.92% good enough for reliable detection of emotional truth? As discussed in section 3.2:

- The 5-class winner from INTERSPEECH 2009 had an accuracy was 41.65% [97].
- Performance of human classification of utterances into six classes was 65.7% [42].
- Emotion labeler agreement in most cases is 3 out of 5. This equates to 60% [43].
- The commercialization threshold for automatic classification systems is 80% [41].
- The concordance of $e_{ij}^{self}(X_{ij}) == e_{ij}^{\widehat{truth}}(X_{ij})$ was 61.66% (section 3.8.1.1) .

41.92% emotion truth accuracy is not sufficiently reliable for clinical patient monitoring or to establish clinical efficacy through statistical analysis. In chapter 4 we will fuse $e_{ij}^{detect}(X_{ij})$ with $e_{ij}^{crowdself}(X_{ij})$ in order to maximize emotional truth. In section 5.13.2 we explore the confusability of emotional truth and discover that emotional truth accuracy is not a black and white measurement; some people have flat affect which confuses emotional truth. This confusability provides insight on their expressiveness/affect.

CHAPTER 4

PSEUDO REAL-TIME EMOTIONAL TRUTH MEASUREMENT

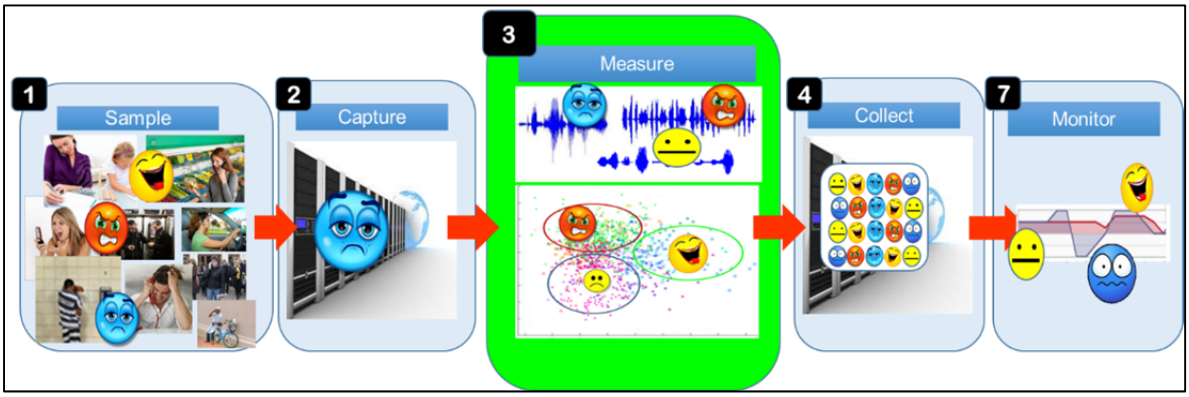


Figure 78 Real-Time Emotional Truth Measurement

For online monitoring of patients, the preferred approach to emotional truth measurement is automatic real-time emotion detection in speech to enable real-time emotional health results. As depicted in Figure 78, real-time monitoring consists of experience sampling; capturing emotional health indicators; completing emotional health sampling by measuring emotional truth, expressiveness, self-awareness, and empathy; collecting daily samples; and monitoring samples over time to detect patient trends.

As presented in the introduction, the 80% accuracy benchmark [41] is a good threshold for viable commercialization of automatic classification systems for deterministic classification problems like image recognition and speech recognition, but may not be the right benchmark for confusable classification like emotion detection.

Nwe et al. [42] determined that human accuracy is 65.7% for six-class emotion classification. In chapter 3, the five-class detector $e_{ij}^{detect}(X_{ij})$ achieved 41.92% emotion truth accuracy which is not reliable for clinically reliable patient monitoring. In section 5.13.2 we discover that emotional truth accuracy is not a black and white measurement and in some cases nondeterministic; some people have flat affect which confuses emotional truth.

The critical real-time aspects of patient monitoring is detecting threshold conditions which require immediate professional intervention. Missing calls over multiple days, as analyzed in section 5.14, is one threshold which should be considered as an intervention trigger condition. Multiple days in a negative state is also a logical condition for a professional to intervene. The key here is "multiple days" indicating a real-time lag of a day or two would be an acceptable compromise if emotional truth accuracy can be improved. This improvement can be incremental, as depicted in Figure 79. The sequence of events proceeds as described for Figure 78. The initial measurements are based on $e_{ij}^{detect}(X_{ij})$. We then incrementally improve accuracy over time by fusing new reinforcing data from anonymous votes and transcriber votes become available. This incremental accuracy improvements across the entire data collection are recalculated each night by a CRON daemon. The certainty score provides an indication of the accuracy.

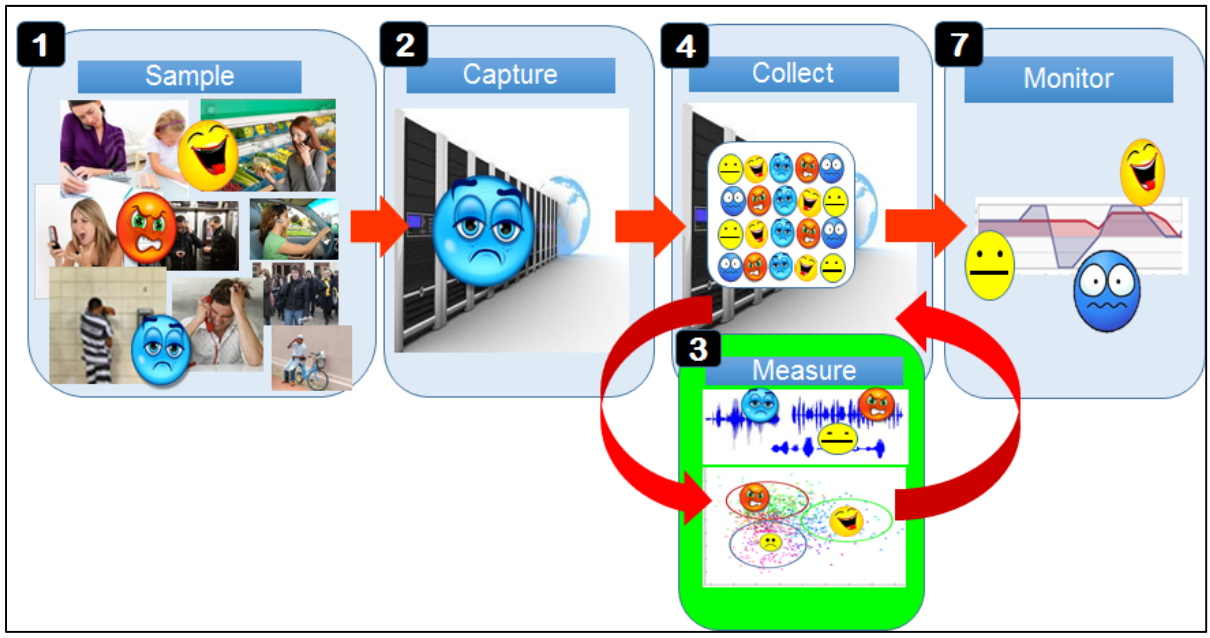


Figure 79 Pseudo-Real-Time Emotional Truth Measurement

4.1 Accuracy-Optimized Pseudo Real-Time Emotion Classifier

The crowd-sourced fused MV classifier in equation 2.9 is fused to the acoustic detector $e_{ij}^{detect}(X_{ij})$ to produce a new pseudo-real-time classifier $\hat{e}(X_{ij})$ in equation 4.1:

$$e_{ij}^{crowdselfdetect}(X_{ij}) = \underset{e\in E}{argmax}\left[w_1 \frac{c[\, e_{ij}^{relate}]}{C_{Erelate}} + w_2 \frac{c[\, e_{ij}^{transcribe}]}{C_{Etranscribe}} + w_3\, e_{ij}^{self} + w_4\, e_{ij}^{detect} \right] \quad (4.1)$$

$$w_1 + w_2 + w_3 + w_4 = 1;\ C_{Erelate}, C_{Etranscribe} \neq 0.$$

This emotion classifier is run each night by a CRON daemon and improves accuracy as new votes become available.

4.2 Determining the Weights for the Real-Time Emotion Classifier

For patient monitoring, we need accurate emotional truth as well as an unbiased means of comparing emotional truth to self-assessment $e_{ij}^{self}(X_{ij})$ in order to measure empathy and self-awareness. Chapter 3 section 3.8 and equation 2.9 in chapter 2 determined optimal weights of $w_1 = 0.27, w_2 = 0.56, w_3 = 0.17$ for relate, transcriber, and self-assessment classifiers respectively to calculate $e_{ij}^{crowdself}(X_{ij})$. In section 2.8.2.2 and in equation 2.12, we established a proportional weighting for the fused emotion classifier of 27% for the anonymous MV classifier, 56% for the transcriber MV classifier, and 17% for self-assessment. Applying equation 2.12 to label the corpus for emotion detection training and test produced the best $e_{ij}^{detect}(X_{ij})$ accuracy at 42% (section 0). From
Table 28, $e_{ij}^{self}(X_{ij})$ is 61.66% accurate as compared to $e_{ij}^{crowd}(X_{ij})$, the fused transcriber and anonymous vote classifier; which is more reliable than $e_{ij}^{detect}(X_{ij})$ at 42%.

However, including $e_{ij}^{self}(X_{ij})$ in a fused classifier, and comparing this to $e_{ij}^{self}(X_{ij})$ biases the classifier (same term on both sides of the equation). On the other hand, there is statistical truth in the reliability of $e_{ij}^{self}(X_{ij})$ and it should be considered in the event of ties between emotional truth scores. In the event of a tie, the certainty score will be low giving an indication of the reliability of the emotional truth. Therefore the ranking order is:

$$w_2 > w_1 > w_4 > w_3 \qquad (4.2)$$

In the following examples we set the weights as $w_2 = 0.4, w_1 = 0.3, w_4 = 0.2,\ w_3 = 0.1$ to simplify the examples and to respect the rankings of equation 4.2. The purpose of the examples is to demonstrate that emotional truth can be automatically calculated and improved as more data becomes available; and an indicator of reliability can be provided thought the certainty score. A more empirical approach to setting the weights will require further analysis. Equation 4.1 then becomes equation 4.3.

$$e_{ij}^{crowdselfdetect}(X_{ij}) = \underset{e \in E}{argmax} \left[0.3 \frac{c[e_{ij}^{relate}]}{C_{Erelate}} + 0.4 \frac{c[e_{ij}^{transcribe}]}{C_{Etranscribe}} + 0.1 e_{ij}^{self} + 0.2\, e_{ij}^{detect} \right] \qquad (4.3)$$

4.3 Pseudo Real-Time Emotion Classifier Example

The Pseudo Real-Time Emotion Classifier was tested during an emotion trial to measure the validity of the automatic emotion detector in detecting mood predictive of performance on the Iowa gambling task conducted by Ogura et al. [110] from January through March of 2013. What follows is a typical example of pseudo real-time emotion classification.

Suppose $patient_2$ is called and the 10th emotionally-charged utterance $X_{10,2}$ is recorded. $patient_2$ self-assesses himself as neutral ($e_{10,2}^{self}(X_{10,2}) = Neutral$). The automatic emotion detector immediately executes, and classifies the speech recording as Happy ($e_{10,2}^{detect}(X_{10,2}) = Happy$). This calculation is depicted in Figure 80. The score of 0 for Happy is highest.

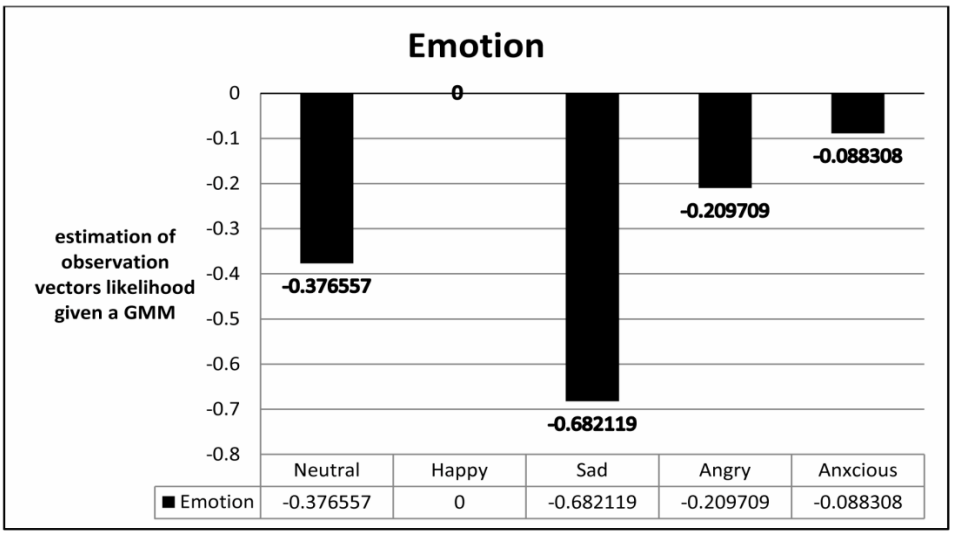


Figure 80 Real-Time Acoustic Emotion Classifier determines Happy

Vote sources for this example are currently as follows:

Table 30 Real-Time Vote Collection

votes	**neutral**	**happy**	**sad**	**angry**	**anxious**	C_E	w_z
Self-Assessment	1					1	0.1
Anonymous						0	0.3
Transcribers						0	0.4
Acoustic eDetect		1				1	0.2
$\sum C$	1	1	0	0	0	2	

Applying equation 4.3 produces Happy as the emotional truth:

$$p(X_{10,2}|Happy) = \left[0.2\frac{1}{1} + 0.1\frac{0}{1}\right] = 0.2$$

$$p(X_{10,2}|Neutral) = \left[0.2\frac{0}{1} + 0.1\frac{1}{1}\right] = 0.1$$

$$e_{10,2}^{confidence}(X_{10,2}) = 0.2$$

$$e_{10,2}^{certainty}(X_{10,2}) = certainty\_factor[2] \text{ x } 0.2 = 0.47732 \text{ x } 0.2 = 0.095464$$

During the course of the day, utterance $X_{10,2}$ is randomly selected during the empathic leg of ESM calls for $patient_8$ on her 22nd ESM, and $patient_{77}$ during his 35th ESM. The votes were $e^{relate}_{22,8,10,2}(X_{10,2}) = Happy$ and $e^{relate}_{35,77,10,2}(X_{10,2}) = Neutral$. The CRON daemon is run that night producing a new calculation for the emotional truth.

Table 31 Vote Collection + 1 day

votes	neutral	happy	sad	angry	anxious	C_E	w_z
Self-Assessment	1					1	0.1
Anonymous	1	1				2	0.3
Transcribers						0	0.4
Acoustic eDetect		1				1	0.2
$\sum C$	2	2	0	0	0	4	

Applying equation 4.3 produces Happy as the emotional truth:

$p(X_{10,2}|Happy) = \left[0.3\frac{1}{2} + 0.2\frac{1}{1} + 0.1\frac{0}{1}\right] = 0.35$

$p(X_{10,2}|Neutral) = \left[0.3\frac{1}{2} + 0.2\frac{0}{1} + 0.1\frac{1}{1}\right] = 0.25$

$e^{confidence}_{10,2}(X_{10,2}) = 0.35$

$e^{certainty}_{10,2}(X_{10,2}) = certainty\_factor[4]$ x $0.35 =$ 0.81696 x 0.35 = 0.285936

The emotional truth is still happy, with an increase in certainty from 9.5464% to 28.5936%.

Every 3rd day, five hired professional transcribers incrementally screen emotions collected since their last session. Three transcribers rate as happy, one as neutral, 1 as anxious.

Table 32 Vote Collection + 3 days

votes	neutral	happy	sad	angry	anxious	C_E	w_z
Self-assessment	1					1	0.1
Anonymous	1	1				2	0.3
Transcribers	1	3			1	5	0.4
eDetect		1				1	0.2
$\sum C$	3	5	0	0	1	9	

Applying equation 4.3 produces Happy as the emotional truth:

$$p(X_{10,2}|Happy) = \left[0.3\frac{1}{2} + 0.4\frac{3}{5} + 0.2\frac{1}{1} + 0.1\frac{0}{1}\right] = 0.59$$
$$p(X_{10,2}|Neutral) = \left[0.3\frac{1}{2} + 0.4\frac{1}{5} + 0.2\frac{0}{1} + 0.1\frac{1}{1}\right] = 0.22$$
$$p(X_{10,2}|Anxious) = \left[0.3\frac{0}{2} + 0.4\frac{1}{5} + 0.2\frac{0}{1} + 0.1\frac{0}{1}\right] = 0.08$$

$$e_{10,2}^{confidence}(X_{10,2}) = 0.59$$
$$e_{10,2}^{certainty}(X_{10,2}) = certainty\_factor[9] \text{ x } 0.2 = 1 \text{ x } 0.35 = 0.59$$

The emotional truth is still happy, with an increase in certainty from 28.5936% to 59%. The certainty score matches the confusability score.

4.4 Conclusion

Professional intervention can be triggered on patient trends such as missed call rates and multiple ESMs containing negative emotions. Trend detection windows would logically be over a period of at least two consecutive days.

Emotional trends are dependent on emotional truth accuracy. Accuracy can be incrementally improved over time, as new data becomes available. The pseudo real-time classifier can provide a preliminary accuracy of 42% (UA). Accuracy can be maximized within a few days, as demonstrated during the emotion trial to measure the validity of the automatic emotion detector in detecting mood predictive of performance on the Iowa gambling task conducted by Ogura et al. [110], which should provide ample time to trigger professional intervention on negative emotions.

Emotional truth accuracy is not a black and white measurement and in some cases nondeterministic; people have flat affect which confuses emotional truth (a statement that

will be proven in section 5.14). Certainty scores and confusability scores provide a good indication of emotional truth accuracy.

CHAPTER 5

STATISTICAL ANALYSIS OF TRIAL DATA

Figure 81 Statistical Analysis on the Trial Data Collection

In Chapters 2 through 4 we sampled, captured, measured, and collected emotional data. The goal of this chapter is to provide evidence, through statistical analysis[17], that capturing and measuring Emotional Health in speech can provide a mechanism for Medical Doctors and Psychiatrists to measure the effectiveness of psychotropic medication, and to provide evidence of psychotherapy effectiveness in mental health and substance abuse treatment programs.

Statistical analysis is the determination of correlation between variables describing the population based on correlation calculated on sampled data. The software engineering approach is to exhaustively explore for all possible significant differences over all variables. The social sciences approach is to formulate a hypothesis and either confirm or reject the hypothesis.

---

[17] Refer to section 2.8 *Step 6: Emotional Health Statistical Analysis* for details on the multilevel statistical analysis methodology applied in this chapter.

Note only differences in means are analyzed. Longitudinal analysis of emotions versus date and time of day indicated that there is not enough data in the corpus; goodness of fit tests failed.

Social Science Hypothesis

We sampled emotional health from one hundred and thirteen (113) participants including three groups: Opioid Addicts undergoing Suboxone® treatment (SUBX) at Dr. Charles Moehs MD MPH clinic (Occupational Medicine Associates of Northern New York) N = 36 [13 men; Expressions = 1054] with an average SUBX continued maintenance period of 1.66 years (Standard Deviation (SD) = 0.48); General Population (GP), N = 44 [15 men; Expressions = 2440]; and Alcohol Anonymous (AA), N = 33 [29 men; Expressions = 3848].

In the Introduction, we presented research findings that mood disorder and anxiety are directly associated with substance abuse [15]. The known pharmacological profile of SUBX [19] is flat affect and lower happiness. Opioid addicts on methadone are less reactive to mood induction. Methadone blunts both elative and depressive emotional reactivity [12]. Patients on opioids, including SUBX and methadone, experience a degree of depression and are in some cases prescribed anti-depressant medication [19]. Scott's concluded that most chemically-dependent individuals cannot identify their feelings (low self-awareness) and do not know how to express them effectively (low empathy) [2].

The null hypothesis would be that there are no differences in happiness, self-awareness, empathy, or affect between the SUBX group and the General Population.

Software engineering approach to statistical analysis

In this thesis, a software engineering approach to exhaustively explore all possible correlations between variables was actually taken. We explored the following questions:

1) Are there differences in emotional truth, self-assessment, self-awareness, and empathy across groups (General Population, AA members, and SUBX patients)? Does gender (Male, Female) have an effect? Does language (English, French) have an effect? Do emotional health indicators vary with the time of day?

2) Does length of the response vary with emotion or group? Does the confidence score (confusability) of the emotional label vary with emotion or group?

3) Are there differences in call completion rates? Which group would be more likely to continue in data collections?

5.1 Transformation of Variables into R

In order to analyze for statistical significance, we must first transform the variables captured and measured into R factors and outcome variables.

In section 0 we established the i^{th}experience sample for $patient_j$ as:

$$ESM_{ij} = \{captured\ parameters\} + \{measured\ parameters\}$$

$$ESM_{ij} = \{patient_j, c_{ij}^{calltype}, c_{ij}^{time}, c_{ij}^{state}, X_{ij}, \quad e_{ij}^{self}(X_{ij}), E^{relate}, c_{ij}^{duration}\} +$$
$$\{e_{ij}^{truth}(X_{ij}), e_{ij}^{self-aware}(X_{ij}), E^{empathy}, length_{ij}(X_{ij}), e_{ij}^{confidence}(X_{ij})\}$$

These variables are transformed into R factors and outcome variables.

5.2 R Factors (Explanatory Variables)

We have grouping factors associated with $patient_j$ including the group (General Population, AA members, and SUBX patients), gender, and language (English, French).

Table 33 R grouping variables (factors)

R Variable	Description
p	$patient_j$ (micro-level grouping variable)
group	(GP, AA, and SUBX)
language	(English, French) note: only the general population group had French
gender	(Male, Female)

5.3 R Outcome Variables

In chapter 2 we developed approximations for the emotional truth, $e_{ij}^{truth}(X_{ij})$ as $e_{ij}^{crowd}(X_{ij})$ and $e_{ij}^{crowdself}(X_{ij})$; and in chapter 3 the automated acoustic detector $e_{ij}^{detect}(X_{ij})$. In chapter 4 the pseudo real-time classifier $e_{ij}^{crowdselfdetect}(X_{ij})$ was introduced.

$e_{ij}^{crowd}(X_{ij})$ is the approximation of $e_{ij}^{truth}(X_{ij})$ used in this section to analyze statistical significance of emotional truth, empathy, and self-awareness to ensure complete independence of measurement; as it is the only estimator that does not include the self-report $e_{ij}^{self}(X_{ij})$. $e_{ij}^{detect}(X_{ij})$ is not sufficiently accurate. $e_{ij}^{crowdselfdetect}(X_{ij})$ is the best approximation of $e_{ij}^{truth}(X_{ij})$, but is only applicable to new data collected – not the existing corpus. The R outcome variables are summarized in Table 34.

.

Table 34 R Outcome Variables

Variable	R Variable	Description
c_{ij}^{state}	callcomplete	$c_{ij}^{state} == call\ completed$
c_{ij}^{time}	date	Timestamp of the sample (stripped out seconds)
c_{ij}^{time}	hour	Hour part of Timestamp
$e_{ij}^{crowd}(X_{ij})$	eCROWD	The emotional truth $e_{ij}^{truth}(X_{ij})$ of recording X_{ij} measured by crowd-sourcing
$e_{ij}^{certainty}(X_{ij})$	eCERTAINTY	Certainty $[e_{ij}^{crowd}(X_{ij}) == e_{ij}^{truth}(X_{ij})]$
$e_{ij}^{confidence}(X_{ij})$	eCONFIDENCE	Confidence $[e_{ij}^{crowd}(X_{ij}) == e_{ij}^{truth}(X_{ij})]$ as a measure of expressiveness
$e_{ij}^{crowdself}(X_{ij})$	eCROWDSELF	The emotional truth $e_{ij}^{truth}(X_{ij})$ of recording X_{ij} measured by crowd-sourcing and self
$length_{ij}(X_{ij})$	eLENGTH	The length of recording X_{ij} as a measure of expressiveness
$e_{ij}^{self-aware}(X_{ij})$	eSELFAWARE	Concordance measure $[e_{ij}^{self}(X_{ij}) == e_{ij}^{truth}(X_{ij})]$
$e_{ijka}^{empathy}(X_{ka})$	eEMPATHY	Concordance measure of ability to determine the emotions of others $[e_{ijka}^{relate}(X_{ka}) == e_{ij}^{truth}(X_{ka})]$

5.4 R Data types

Table 35 explores the R variables from Table 33 and Table 34. Data types are nominal, ordinal, continuous or logical. Variables are either dependent (D) or independent (I).

Table 35 R Data Types

R Variable	**Data type**	**I/D**	**Possible values**
p	nominal	I	[1,∞]
group	nominal	I	GP=1, AA=2, SUBX=3
gender	nominal	I	Male=1, female=2
language	nominal	I	English=1, French=2
callcomplete	logical	D	TRUE, FALSE
eCROWD	nominal	D	OK=1, happy=2, sad=3, angry=4, anxious=5
eCERTAINTY	continuous	D	[0,1] binomial
eCONFIDENCE	continuous	D	[0,1] binomial
eCROWDSELF	nominal	D	OK=1, happy=2, sad=3, angry=4, anxious=5
eLENGTH	continuous	D	[0,10] seconds
eSELFAWARE	nominal	D	OK=1, happy=2, sad=3, angry=4, anxious=5
eEMPATHY	nominal	D	OK=1, happy=2, sad=3, angry=4, anxious=5

Emotional nominal (categorical) variables eCROWD, eSELFAWARE, and eEMPATHY are further divided into binomial variables to enable binomial logistical regression. For example, eCROWD expands to the five binomials.

Table 36 R Binomials Derived from the Emotion Set

R Variable	Data type	values
okCROWD	logical	TRUE, FALSE
happyCROWD	logical	TRUE, FALSE
sadCROWD	logical	TRUE, FALSE
angryCROWD	logical	TRUE, FALSE
anxCROWD	logical	TRUE, FALSE

5.5 Data summary

This section provides a brief visualization of key factors and dependent variables. There are a total of 8,376 ESMs. Frequencies in Figure 82 describe and graph the ESMs collected per trial participant. ESM frequencies are skewed towards a Poisson distribution. The median indicates that half the participants contributed less than 36.5 ESMs each. A few participants provided in excess of 400 ESMs.

ESM frequencies

N	Valid	130
	Missing	0
Mean		64.43
Median		36.50
Std. Deviation		85.936
Variance		7385.022
Range		458
Minimum		1
Maximum		459
Percentiles	10	5.20
	20	14.00
	25	16.00
	30	20.30
	40	28.40
	50	36.50
	60	50.00
	70	59.00
	75	67.25
	80	84.80
	90	177.60

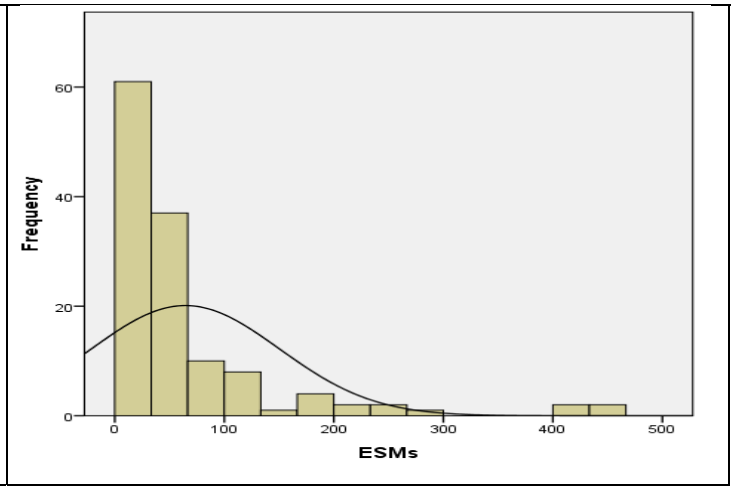


Figure 82 Participant ESM Frequencies

Figure 83 Histograms of Regression Factors

Figure 83 provides histograms of factor frequencies. Of 8,376 ESMs, 7,342 ESMs are associated within the three population groups (GP, AA, and SUBX). 1,034 ESMs were

collected outside of the 3 groups through a demonstration phone number. Of the 8,376 ESMs, 5,513 ESMs came from male participants, 2,576 ESMs from females, and 287 from the demo-line with gender unknown. 6,795 ESMs come from English speakers, and 1,581 from French speakers. Figure 84 depicts the frequencies of speech duration. Most speech captured was less than five seconds in duration.

```
     eLENGTH
Min.   : 0.09575
1st Qu.: 2.11575
Median : 2.97575
Mean   : 3.79197
3rd Qu.: 4.62975
Max.   :20.30570
```

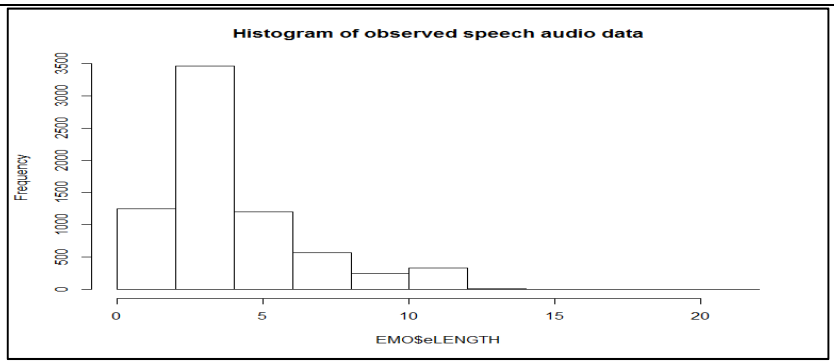
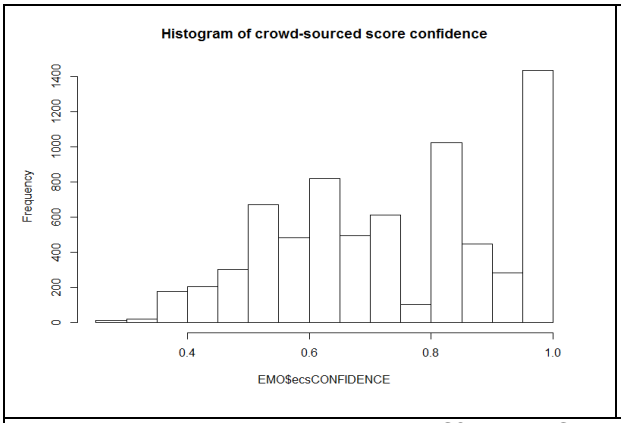
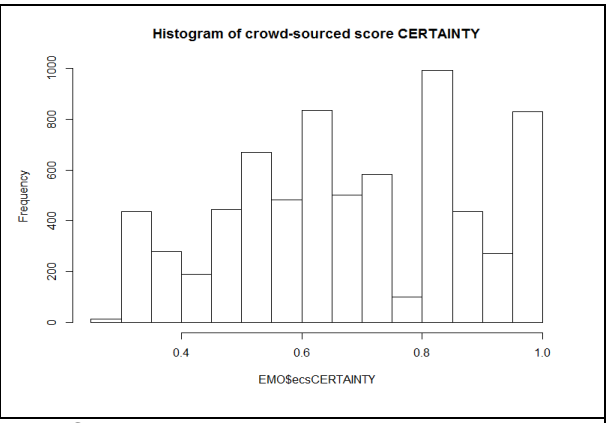


Figure 84 Speech Duration Histogram






```
  ecCONFIDENCE     ecCERTAINTY
Min.   :0.2860   Min.   :0.0000
1st Qu.:0.6120   1st Qu.:0.5430
Median :0.7770   Median :0.6610
Mean   :0.7637   Mean   :0.6469
3rd Qu.:0.9900   3rd Qu.:0.8240
```

Figure 85 Confidence and Certainty Histogram

Figure 85 depicts the frequencies of eCROWD confidence and certainty levels. Most of the speech has a labeled confidence of and certainty of more than 40%.

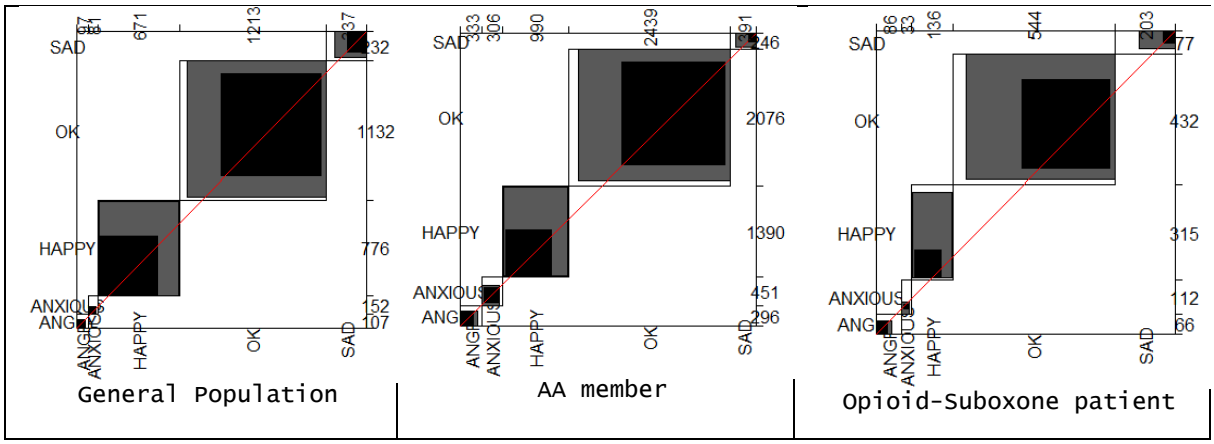


Figure 86 Emotional Truth Concordance

Figure 86 depicts emotional truth concordance across groups. The grey area for each emotion represents emotional truth; and the black represents self-reported emotion. Visually, the General Population is the most "in tune" with their own emotions, and the Opioid addicts are least aware of how they are feeling. The most glaring discrepancy is self-awareness of happiness for Opioid addicts.

5.6 Emotional Health Means

Emotional truth, self-awareness, and empathy categorical variables are dependent discrete-choice outcome variables (a.k.a. unordered polytomous variables). The standard statistical model for discrete-choice is logistical regression; where each binomial choice is split out from the multinomial category and independent logistical regressions are performed on each binomial (e.g. emotional truth variable eTRUTH split into binomials happyTRUTH, sadTRUTH, angryTRUTH, anxiousTRUTH, and neutralTRUTH variables).

We cannot statistically compare means between each emotion since each mean is calculated in a separate analysis. However, the contrast of emotion means, calculated using equation 1.34 for each emotion, is interesting and is presented. An attempt to perform multinomial regression analysis is performed in chapter 6.

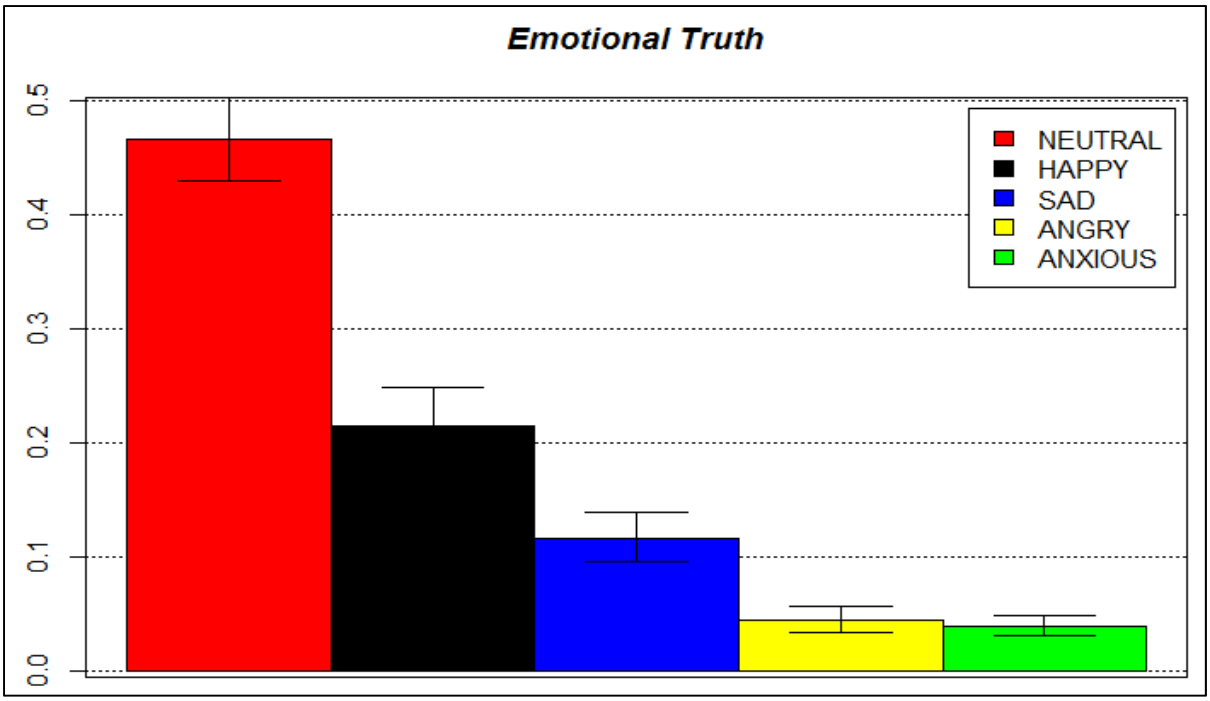


Figure 87 Emotional Truth Means

Table 37 Emotional Truth Means

Emotional Truth	Probability	95% confidence interval	p-value Pr(>|z|)
Neutral	46.6%	43.0% - 50.3%	0.0663
Happy	21.5%	18.5%-24.9%	<0.0001
Sad	11.6%	9.6%-13.9%	<0.0001
Angry	4.4%	3.4%-5.6%	<0.0001
Anxious	3.9%	3.1%-4.8%	<0.0001
Total	88.00%		

Figure 87 and Table 37 provide the multilevel null means for emotional truth. The total of 88% is less than 100%, since each logistical regression analysis was performed separately (see chapter 6 for analysis and discussion). The ratio 3.42:1 of positive emotions (Neutral + Happy = 68.1%) to negative emotions (Sad + Angry + Anxious = 19.9%) approaches Lyubomirsky et al. [23] 80% (4:1) threshold.

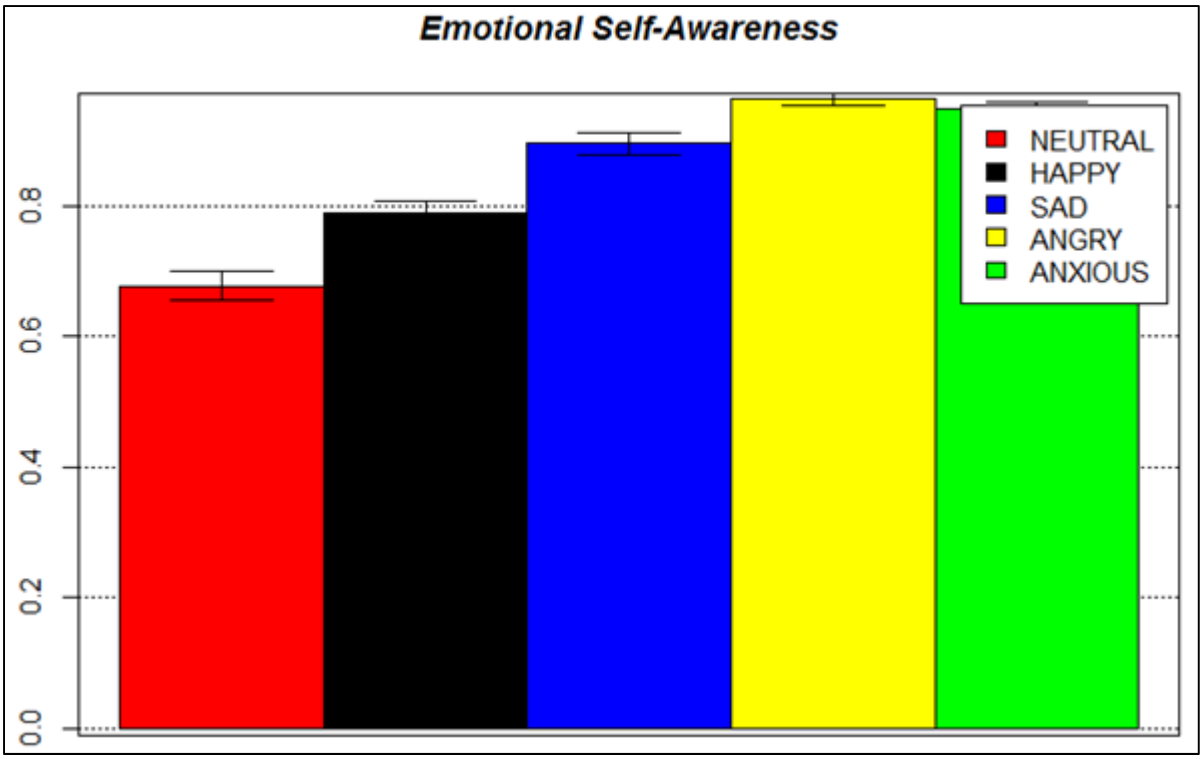


Figure 88 Self-Awareness Means

Table 38 Self-Awareness Means

Self-Awareness	Probability	95% confidence interval	p-value Pr(>|z|)
Neutral	67.8%	65.6% - 69.9%	<0.0001
Happy	78.8%	76.6%-80.8%	<0.0001
Sad	89.6%	87.8%-91.2%	<0.0001
Angry	96.4%	95.5%-97.1%	<0.0001
Anxious	94.9%	93.7%-96.0%	<0.0001

Figure 88 and Table 38 provide the multilevel null means for self-awareness. People seem to be highly aware of their Anger and Anxiety.

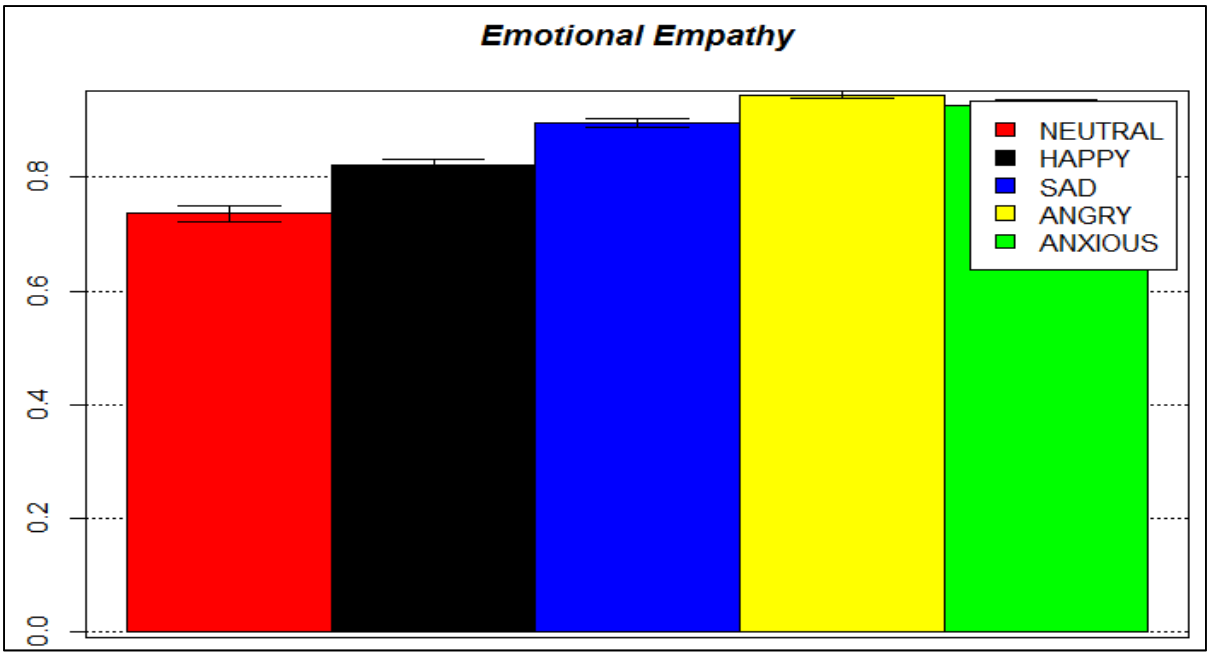


Figure 89 Empathy Means

Table 39 Empathy Means

Empathy	Probability	95% confidence interval
Neutral	73.7%	72.3% - 75.0%
Happy	82.0%	80.8%-83.2%
Sad	89.7%	88.8%-90.5%
Angry	94.6%	93.9%-95.2%
Anxious	92.7%	91.6%-93.7%

Figure 89 and Table 39 provide the multilevel null means for empathy towards other people. People seem to more empathetic towards Anger and Anxiety.

5.7 Statistical Analysis for Emotional Health Effects

The procedures for calculating the statistical significance for continuous and discrete-choice outcome variable is described in section 1.7.

For each $emotion \in (neutral, happy, sad, angry, anxious)$, an R function is executed to explore for significant effects (p<0.05) and trends (p<0.1) against categorical factors group3, language, and gender. Homoscedasticity, distribution normality, and goodness-of-fit tests ensure validity or results. This process is repeated for each $category \in \{eCROWD, eSELFAWARE, eEMPATHY\}$. Analysis of expressiveness is through continuous outcome variables eLENGTH and ecCONFIDENCE. To avoid output noise, the R output is pruned for worthy results and summarized in the next sections.

5.8 Happiness Effects

Table 40 summarizes the effects (p<0.05) and trends (p<0.1) within happiness truth, self-assessment, self-awareness, and empathy across group, gender, language. Each effect is explained in detail in the following pages. Detailed happiness regression analysis can be found in appendix D.

Table 40 Happiness Effects (p<0.05) and Trends (p<0.1)

Happiness Health indicator	**Fixed effect**	**Probability**	**95% confidence interval**	**p-value Pr(>|z|)**
	happy<CROWD|SELF|SELFAWARE|EMPATHY> ~ (1 | p)			
Self-Assessment	2-level null model	31.0%	27.3%-35.0%	<0.0001
Emotional Truth	2-level null model	21.5%	18.5%-24.9%	<0.0001
Self-Awareness	2-level null model	78.8%	76.6%-80.8%	<0.0001
Empathy	2-level null model	82.0%	80.8%-83.2%	<0.0001
Effect	**Formula: happyCROWD ~ group3 + (1 | p) …base-level = GP**			
Emotional Truth	GP	24.7%	19.2%-31.0%	<0.0001
Emotional Truth	SUBX	15.2%	9.7%-22.9%	0.0171
Effect	**Formula: happyCROWD ~ group3 + (1 | p) …base-level = AA**			
Emotional Truth	AA	24.0%	18.2%-31.0%	<0.0001
Emotional Truth	SUBX	15.2%	9.5%-23.3%	0.0310
Trend	**Formula: happySELFAWARE ~ group3 + (1 | p)**			
Self-Awareness	GP	78.8%	76.6%-80.8%	<0.0001
Self-Awareness	SUBX	75.3%	68.4%-81.1%	0.0656

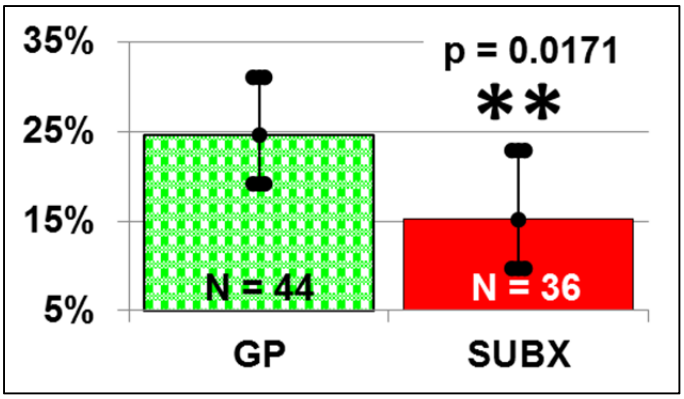


Figure 90 SUBX less Happy than GP

SUBX patients have a significantly lower probability of being happy (p=0.0171) (15.2%; CI: 9.7%-22.9%) than the GP (24.7%; CI: 19.2%-31.0%).

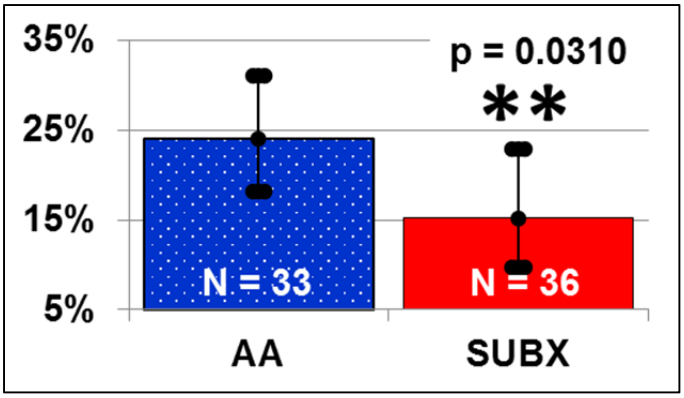


Figure 91 SUBX less Happy than AA

SUBX patients have a significantly lower probability of being happy (p=0.0310) (15.2%; CI: 9.5%-23.3%) than AA members (24.0%; CI: 18.2%-31.0%).

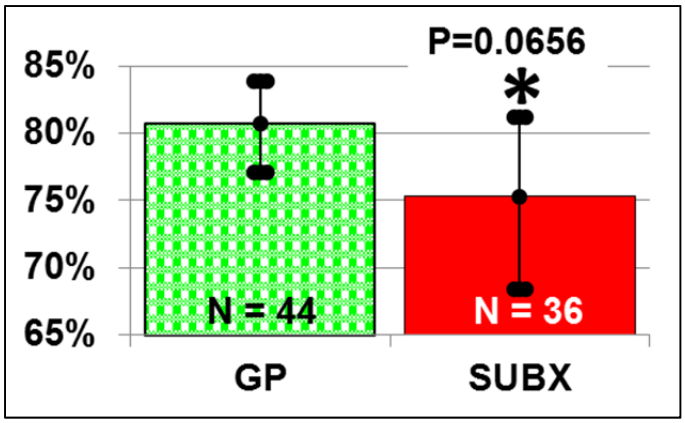


Figure 92 Trend that SUBX less Self-Aware of Happiness than GP

There is a trend that SUBX patients have a lower probability of being self-aware of their happiness (p=0.0656) (75.3%; CI: 68.4%-81.1%) than AA members (78.8%; CI: 76.6%-80.8%).

There is no difference in happiness between the GP and AA members. There is no difference in happiness self-awareness between AA members and the General Population or Opioid-Suboxone patients. There are no emotional health differences across gender or language.

5.9 Sadness Effects

Table 41 summarizes the effects ($p<0.05$) and trends ($p<0.1$) within happiness truth, self-assessment, self-awareness, and empathy across group, gender, language. Each effect is explained in detail in the following pages. Detailed multilevel sadness regression analysis can be found in appendix E.

Table 41 Sad Effects (p<0.05) and Trends (p<0.1)

Sadness Health indicator	Fixed effect	Probability	95% confidence interval	p-value Pr(>|z|)
	Formula: sad< SELF|CROWD|SELFAWARE|EMPATHY> ~ (1 | p)			
Self-assessment	2-level null model	5.2%	4.1%-6.5%	<0.0001
Emotional Truth	2-level null model	11.6%	9.6%-13.9%	<0.0001
Self-Awareness	2-level null model	89.6%	87.8%-91.2%	<0.0001
Empathy	2-level null model	89.7%	88.8%-90.5%	<0.0001
Trend	**Formula: sadCROWD ~ group3 + (1 | p) …base-level = GP**			
Emotional Truth	GP	12.6%	9.2%-16.9%	<0.0001
Emotional Truth	AA	8.3%	5.1%-13.2%	0.0766
Effect	**Formula: sadCROWD ~ gender + (1 | p)**			
Emotional Truth	Female	14.7%	11.2-19.0	<0.0001
Emotional Truth	Male	8.8%	5.9-12.8	0.0061
Trend	**Formula: sadSELFAWARE ~ group3 + (1 | p) …base-level = GP**			
Self-Awareness	GP	89.6%	87.8%-91.2%	<0.0001
Self-Awareness	SUBX	85.3%	78.6%-90.2%	0.0817
Effect	**Formula: sadSELFAWARE ~ group3 + (1 | p) …base-level = AA**			
Self-Awareness	AA	91.3%	88.4%-93.6%	<0.0001
Self-Awareness	SUBX	85.3%	78.3%-90.3%	0.0127
Trend	**Formula: sadSELFAWARE ~ gender + (1 | p)**			
Self-Awareness	Female	87.5%	84.1%-90.3%	<0.0001
Self-Awareness	Male	91.0%	87.4%-93.6%	0.0535

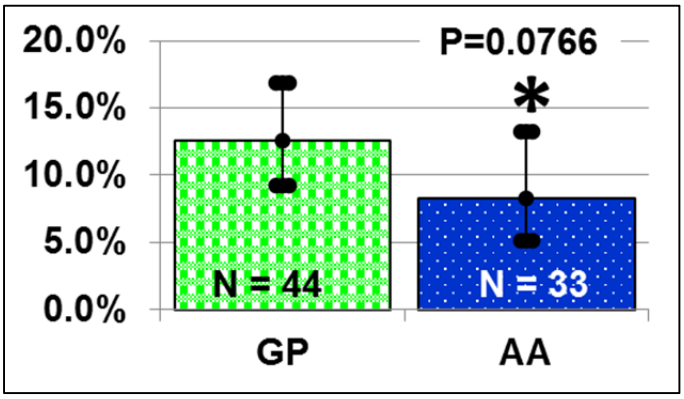


Figure 93 AA less Sad than GP

There is a trend that AA Members have a lower probability of being sad (p=0.0766) (8.3%; CI: 5.1-13.2) than the GP (12.6%; CI: 9.2-16.9).

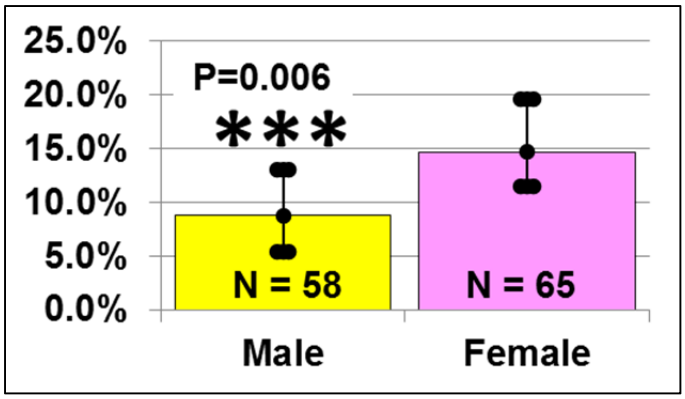


Figure 94 Males less Sad than Females

Males have a lower probability of being sad (p=0.006) (8.8%; CI: 5.9-12.8) than Females (14.7%; CI: 11.2-19.0).

.

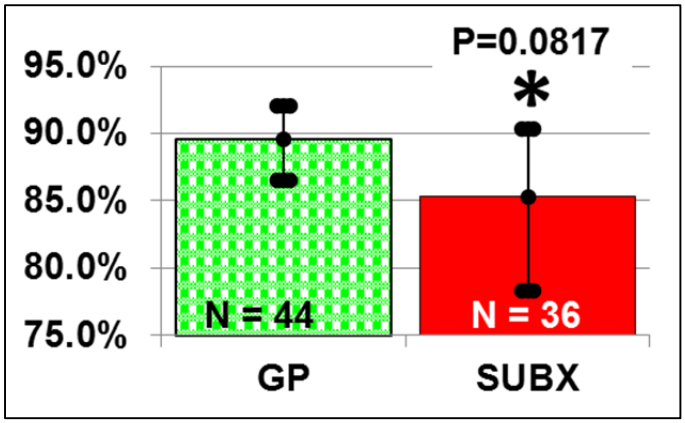


Figure 95 Trend that SUBX less Self-Aware of Sadness than GP

There is a trend that SUBX patients are less self-aware of their sadness (p=0.0817) (85.3%; CI: 78.6-90.2) than the GP (89.6%; CI: 87.8-91.2)

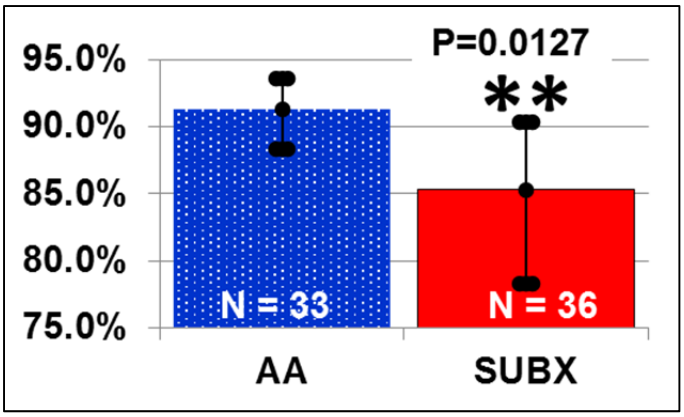


Figure 96 SUBX less Self-Aware of Sadness than AA

SUBX patients are less self-aware of their sadness (p=0.0127) (85.3%; CI: 78.3-90.3) than AA Members (91.3%; CI: 87.4-93.6).

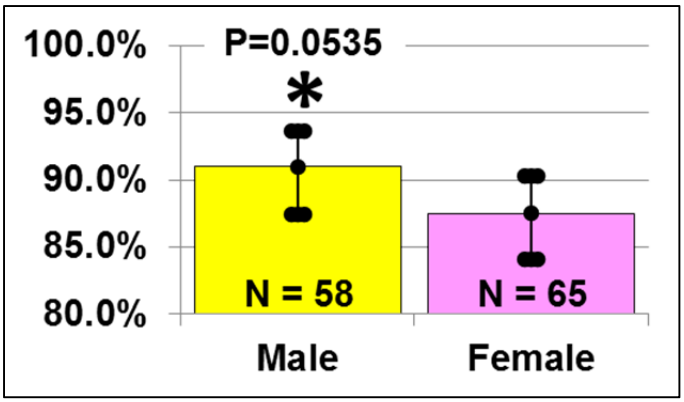


Figure 97 Trend that Males are more Self-Aware of Sadness than Females

There is a trend that Males are more self-aware of sadness (p=0.0535) (91.0%; CI: 87.4-93.6) than females (87.5%; CI: 84.1-90.3).

There are no significance differences of empathy of sadness across group, gender or language. There are no effects across language.

5.10 Anxiety Effects

Table 42 summarizes the effects (p<0.05) and trends (p<0.1) within happiness truth, self-assessment, self-awareness, and empathy across group, gender, language. Each effect is explained in detail in the following pages. Detailed multilevel sadness regression analysis can be found in appendix F.

Table 42 Anxious Effects (p<0.05) and Trends (p<0.1)

Anxiety Health indicator	Fixed effect	Probability	95% confidence interval	p-value Pr(>|z|)
	Formula: anx<CROWD|SELF|SELFAWARE|EMPATHY> ~ (1 | p)			
Self-Assessment	2-level null model	5.9%	4.6%-7.5%	<0.0001
Emotional Truth	2-level null model	3.9%	3.1%-4.8%	<0.0001
Self-Awareness	2-level null model	94.9%	93.7%-96.0%	<0.0001
Empathy	2-level null model	92.7%	91.6%-93.7%	<0.0001
Effect	**Formula: anxCROWD ~ group3 + (1 | p) …base-level = AA**			
Emotional Truth	AA	4.8%	3.2%-7.3%	<0.0001
Emotional Truth	SUBX	2.2%	1.1%-4.5%	0.0282
Effect	**Formula: anxSELFAWARE ~ group3 + (1 | p) …base-level = GP**			
Self-Awareness	GP	95.8%	93.8%-97.1%	<0.0001
Self-Awareness	SUBX	91.8%	86.0%-95.3%	0.0190
Effect	**Formula: anxSELFAWARE ~ group3 + (1 | p) …base-level = AA**			
Self-Awareness	AA	95.6%	93.4%-97.1%	<0.0001
Self-Awareness	SUBX	91.8%	85.8%-95.4%	0.0332
Effect	**Formula: anxEMPATHY ~ group3 + (1 | p) …base-level = GP**			
Empathy	GP	93.5%	91.8%-94.8%	<0.0001
Empathy	AA	90.4%	86.7%-93.1%	0.0215
Effect	**Formula: anxEMPATHY ~ group3 + (1 | p) …base-level = AA**			
Empathy	AA	90.4%	87.8%-92.5%	<0.0001
Empathy	SUBX	93.5%	90.3%-95.7%	0.0484
Trend	**Formula: anxEMPATHY ~ gender + (1 | p) …base-level = AA**			
Empathy	Female	93.7%	92.1%-94.9%	<0.0001
Empathy	Male	91.8%	89.2%-93.9%	0.0820

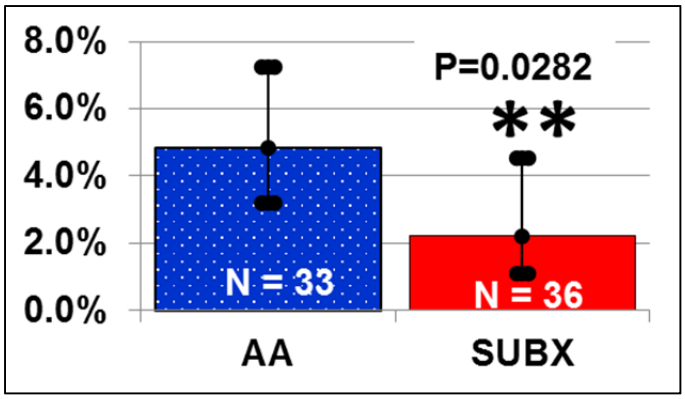


Figure 98 SUBX less Anxious than AA

AA Members have over twice the probability of being anxious (4.8%; CI: 3.2%-7.3%) than SUBX patients (p=0.0282) (2.2%; CI: 1.1%-4.5%)

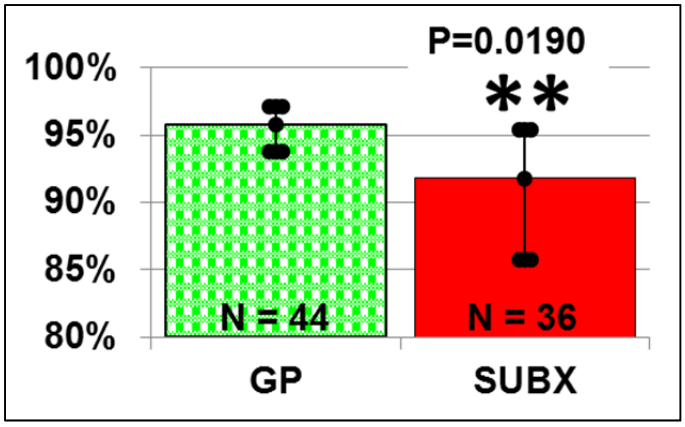


Figure 99 SUBX less Self-Aware of Anxiety than GP

SUBX patients are less self-aware of their anxiety (p=0.0190) (91.8%; CI: 86.0%-95.3%) than the GP (95.8%; CI: 93.8%-97.1%)

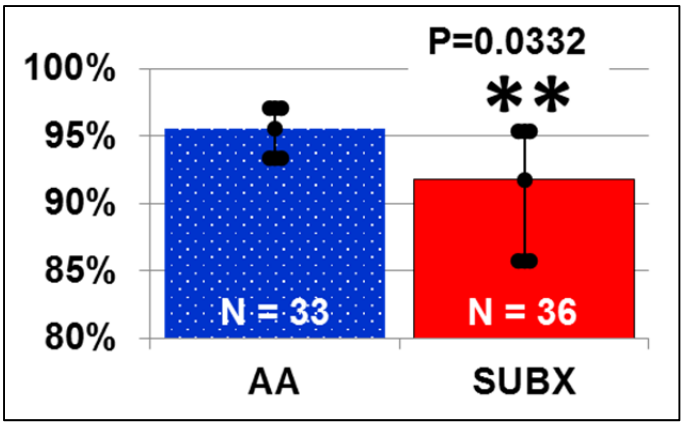


Figure 100 SUBX less Self-Aware of Anxiety than AA

SUBX patients are less self-aware of their anxiety (p=0.0332) (91.8%; CI: 85.8%-95.4%) than AA members (95.6%; CI: 93.4%-97.1%)

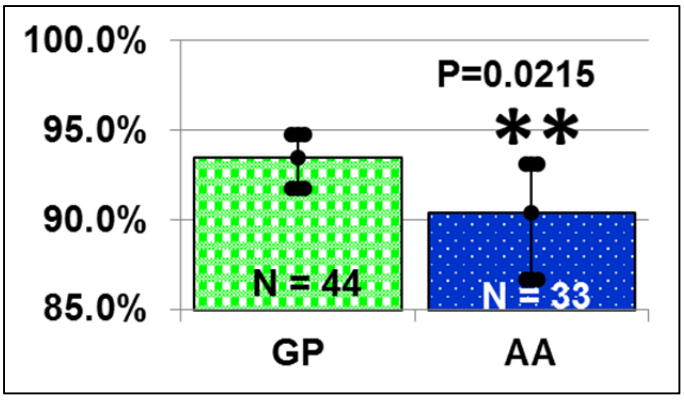


Figure 101 AA less Empathetic to Anxiety than GP

AA members are less empathetic to anxiety (p=0.0215) (90.4%; CI: 86.7%-93.1%) than the GP (93.5%; CI: 91.8%-94.8%)

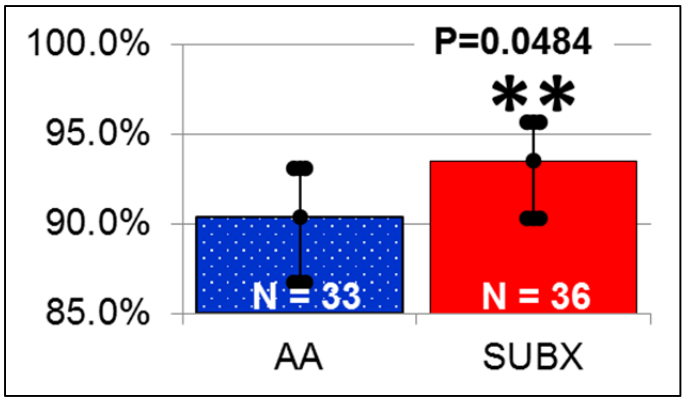


Figure 102 AA less Empathetic to Anxiety than SUBX

AA members are less empathetic to anxiety (p=0.0484) (90.4%; CI: 87.8%-92.5%) than SUBX patients (93.5%; CI: 90.3%-95.7%)

: 92.1%-94.9%) than Males (p=0.0820) (91.8%; CI: 89.2%-93.9%)

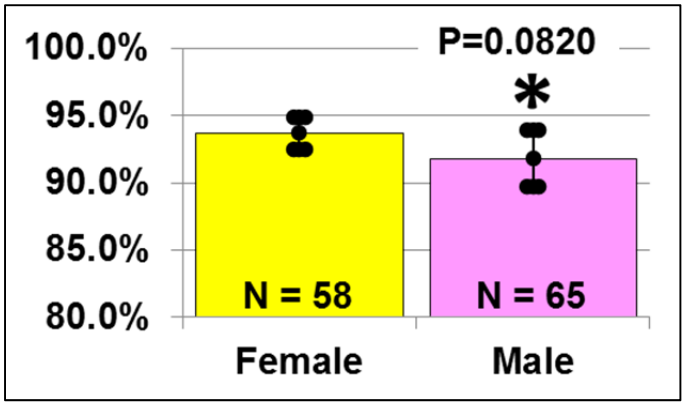


Figure 103 Trend that Males less Empathetic to Anxiety than Females

There is a trend that Females are more empathic to anxiety (93.7%; CI

There are no emotional health differences across language.

5.11 Anger Effects

Table 43 summarizes the effects (p<0.05) and trends (p<0.1) within happiness truth, self-assessment, self-awareness, and empathy across group, gender, language. Each effect is explained in detail in the following pages. Detailed multilevel sadness regression analysis can be found in appendix G.

Table 43 Anger Effects (p<0.05) and Trends (p<0.1)

Anger Health indicator	Fixed effect	Probability	95% confidence interval	p-value Pr(>|z|)
Formula: angry<CROWD|SELF|SELFAWARE|EMPATHY> ~ (1 | p)				
Self-Assessment	2-level null model	3.6%	2.8%-4.6%	<0.0001
Emotional Truth	2-level null model	4.4%	3.4%-5.6%	<0.0001
Self-Awareness	2-level null model	96.4%	95.5%-97.1%	<0.0001
Empathy	2-level null model	94.6%	93.9%-95.2%	<0.0001
Trend	**Formula: angryEMPATHY ~ gender + (1 | p)**			
Empathy	Female	95.1%	94.2%-96.0%	<0.0001
Empathy	Male	94.1%	92.5%-95.6%	0.0991
Effect	**Formula: angryEMPATHY ~ language + (1 | p)**			
Empathy	English	93.0%	91.0%-94.6%	<0.0001
Empathy	French	95.4%	93.4%-96.8%	0.0183

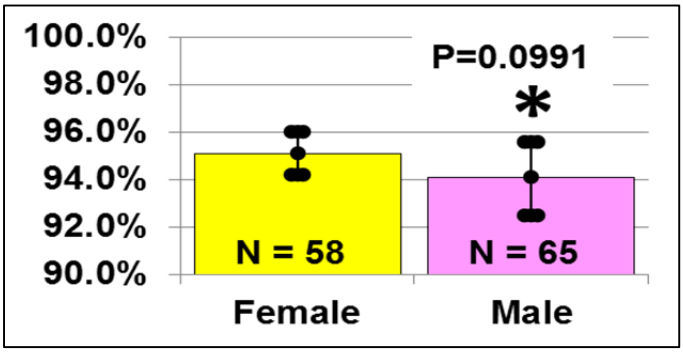


Figure 104 Trend that Males less Empathetic to Anger than Females

There is a trend that Females are more empathic to anger (95.1%; CI: 94.2%-96.0%) than Males (p=0.0991) (94.1%; CI: 92.5%-95.6%)

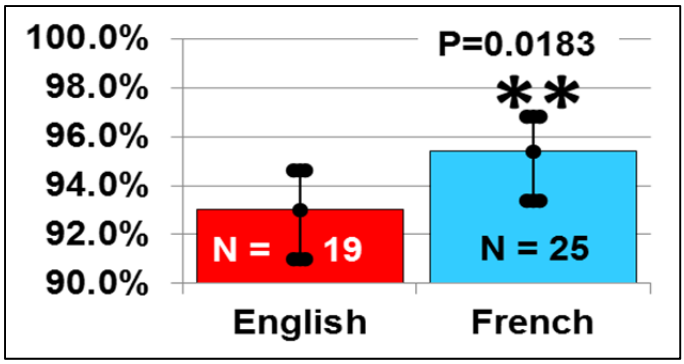


Figure 105 Trend that French more Empathetic to Anger than English

Within the General Population, French speaking people are more empathic to anger (95.4%; CI: 93.4%-96.8%) than English (p=0.0183) (93.0%; CI: 91.0%-94.6%)

There are no differences in emotional health across SUBX patients, AA members, and the GP.

5.12 Neutral (Okay) Effects

Table 44 summarizes the effects ($p<0.05$) and trends ($p<0.1$) within happiness truth, self-assessment, self-awareness, and empathy across group, gender, language. Each effect is explained in detail in the following pages. Detailed multilevel sadness regression analysis can be found in appendix H.

Table 44 Neutral Effects (p<0.05) and Trends (p<0.1)

Neutral Health indicator	Fixed effect	Probability	95% confidence interval	p-value Pr(>|z|)
	Formula: ok<CROWD|SELF|SELFAWARE|EMPATHY> ~ (1 | p)			
Self-assessment	2-level null model	42.1%	38.1 - 46.2	<0.0001
Emotional Truth	2-level null model	46.6%	43.0 - 50.3	0.0663
Self-Aware	2-level null model	67.8%	65.6 - 69.9	<0.0001
Empathy	2-level null model	73.7%	72.3 - 75.0	<0.0001
Trend	**Formula: okCROWD ~ language + (1 | p)**			
Emotional Truth	English	40.5%	33.9 - 47.5	0.0065
Emotional Truth	French	49.3%	40.3 - 58.3	0.0505
Effect	**Formula: okSELFAWARE ~ group3 + (1 | p) …base-level = GP**			
Self-Aware	GP	70.7%	67.3 - 74.0	<0.0001
Self-Aware	SUBX	63.2%	57.0 - 69.0	0.0083
Effect	**Formula: okEMPATHY ~ group3 + (1 | p) …base-level = AA**			
Empathy	AA	71.7%	68.9 - 74.3	<0.0001
Empathy	SUBX	76.5%	72.3 - 80.2	0.0223

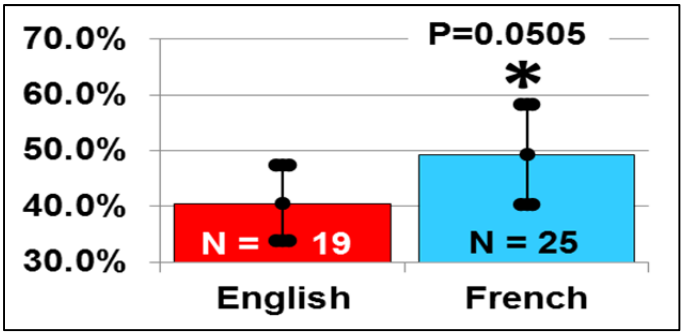


Figure 106 Trend that French People are more Neutral than English People

Within the General Population, there is a trend that French speaking people are more neutral (p=0.505) (49.3%; CI: 33.9% - 47.5%) than English (40.5%; CI: 33.9% - 47.5%)

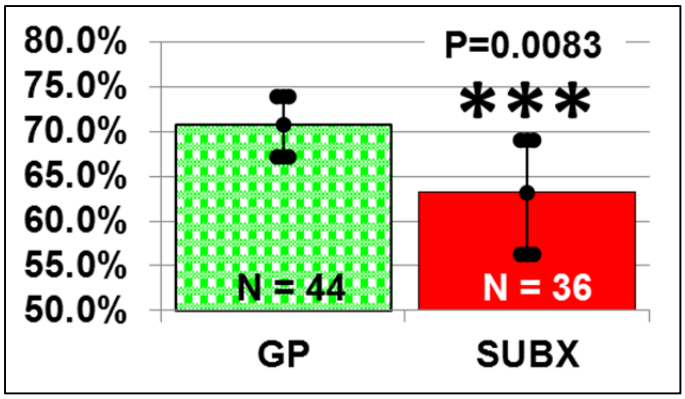


Figure 107 GP more Self-Aware of Neutral State than SUBX Patients

GP are more self-aware of their neutral emotional state (p=0.0083) (70.7%; CI: 67.3% - 74.0%) than SUBX patients (63.2%; CI: 57.0% - 69.0%)

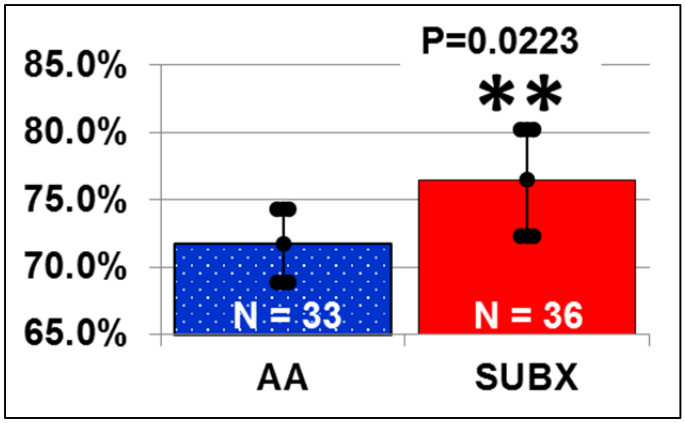


Figure 108 SUBX more Empathic to Neutral State than AA members

SUBX are more empathic to other peoples' neutral emotional state (p=0.0223) (76.5%; CI: 72.3% - 80.2%) than AA members (71.7%; CI: 68.9% - 74.3%)

There are no differences in neutral state emotional health across gender.

5.13 Expressiveness and Affect

Emotional expressiveness or affect is measured by two methods: The length-of-speech, and the confidence score (confusability) of the emotional truth. Length-of-speech is short for people who respond with phrases like "fine", "ok", "not bad"; and longer for people who are more expressive about how they feel (e.g. "having a great day! The sun is shining!"). The confidence score $e_{ij}^{confidence}(X_{ij})$ of $e_{ij}^{truth}(X_{ij})$ is a measure of emotion confusability. The higher the score, the less confusable the emotion is (see section 2.4.1).

5.13.1 Length of Speech Effects

Table 45 summarizes the effects ($p<0.05$) and trends ($p<0.1$) within length-of-speech across group and emotion. Each effect is explained in detail in the following pages. Detailed multilevel sadness regression analysis can be found in appendix I.

Table 45 Length-of-Speech Effects ($p<0.05$) and Trends ($p<0.1$)

Fixed effect	**Seconds**	**95% confidence interval**	**p-value**
Formula: eLENLOG ~(1|p) …eLENLOG is log of eLENGTH			
Null model mean	3.07	2.89 - 3.25	<0.0001
Formula: eLENLOG ~group3+(1|p) …base-level = GP			
GP	3.46	3.15 - 3.80	<0.0001
SUBX	2.39	2.07 - 2.76	<0.0001
Formula: eLENLOG ~group3+(1|p) …base-level = AA			
AA	3.31	2.97 - 3.68	<0.0001
SUBX	2.39	2.05 - 2.78	<0.0001
Formula: eLENLOG ~eCROWD+(1|p)) …base-level = Neutral			
Neutral	2.97	2.83 - 3.12	<0.0001
Happy	3.36	3.27 - 3.45	<0.0001
Sad	3.15	3.05 - 3.25	<0.0001
Anger	3.41	3.27 - 3.56	<0.0001
Anxious	3.60	3.44 - 3.76	<0.0001

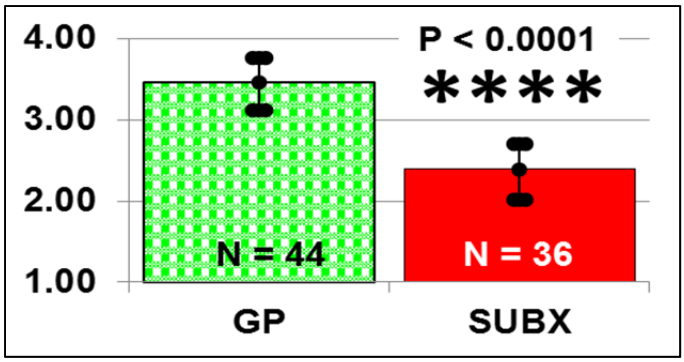


Figure 109 GP more Self-Aware of Neutral State than SUBX Patients

SUBX are less expressive, as measured by the length of their emotional expression ($p<0.0001$) (2.39 seconds; CI: 2.07 -2.76) than the GP (3.46 seconds; CI: 3.15-3.25)

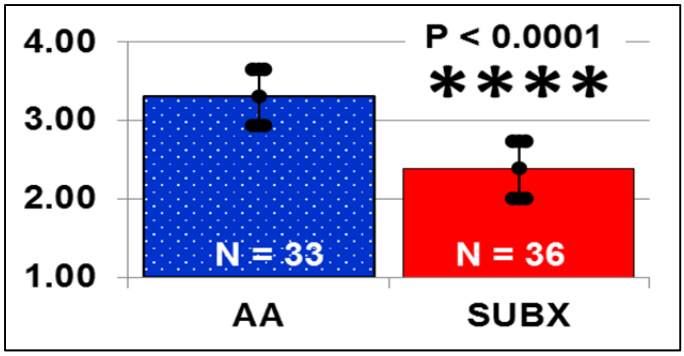


Figure 110 GP more Self-Aware of Neutral State than SUBX Patients

SUBX are less expressive, as measured by the length of their emotional expression ($p<0.0001$) (2.39 seconds; CI: 2.05 -2.78) than AA members (3.31 seconds; CI: 2.97-3.68)

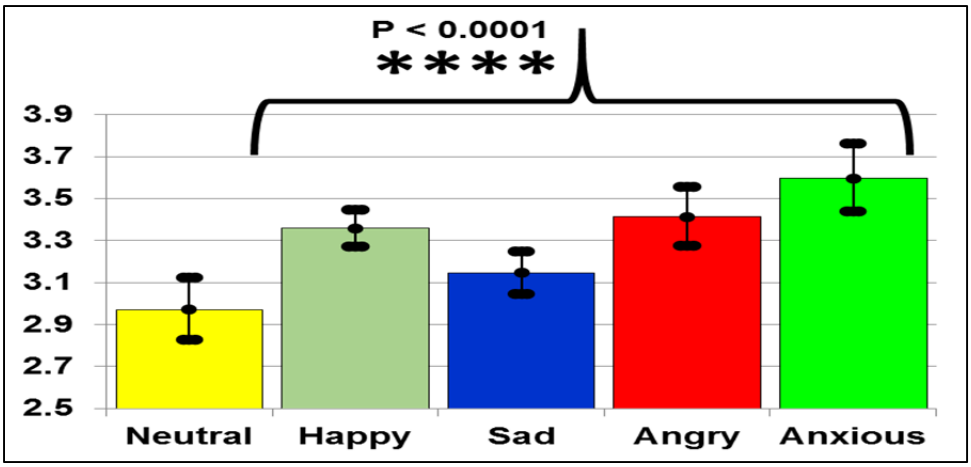


Figure 111 Length-of-speech across emotions

Figure 111 shows that all emotions differ significantly from neutral ($p<0.0001$). Across the entire population, the neutral emotion is the least expressive (2.97 seconds; CI: 2.83 – 3.12), followed by Sadness (3.15; CI: 3.05-3.25), Happiness (3.36; CI: 3.27-3.45), Anger (3.41; CI: 3.27-3.56) and Anxiety (3.60; CI: 3.44-3.76).

5.13.2 Confusability Effects

Table 46 summarizes the effects ($p<0.05$) and trends ($p<0.1$) within confusability across group, emotion, gender, and language. Detailed multilevel sadness regression analysis can be found in appendix J.

Note: The residuals are not normal, which violates the assumptions for HLM described in section 1.7. Attempts made in appendix J including power transformation and log-normalization approached normalization, but fell short.

Table 46 Confidence Score Effects (p<0.05) and Trends (p<0.1)

Fixed effect	Confidence	95% CI	p-value
Formula: ecCONFIDENCE~(1|p)			
Null model mean	75%	0.73 – 0.76	<0.0001
Formula: ecCONFIDENCE ~group3+(1|p) …base-level = GP			
GP	71.4%	0.70 – 0.73	<0.0001
SUBX	65.0%	0.63 – 0.67	<0.0001
Formula: ecCONFIDENCE ~group3+(1|p) …base-level = AA			
AA	70.9%	0.69 – 0.72	<0.0001
SUBX	65.0%	0.63 – 0.67	<0.0001
Formula: ecCONFIDENCE ~eCROWD+(1|p)			
Neutral (Intercept)	69.2%	0.68 – 0.70	<0.0001
Happy	72.7%	0.72 – 0.74	<0.0001
Sad	68.2%	0.67 – 0.70	0.119
Angry	70.3%	0.70 – 0.72	0.217
Anxious	71.1%	0.69 – 0.73	0.038
Formula: ecCONFIDENCE ~gender+(1|p)			
Female	68.9%	0.67 - 0.70	<0.0001
Male	70.4%	0.68 - 0.72	0.0912
Formula: ecCONFIDENCE ~language+(1|p)			
English	69.3%	0.68 – 0.70	<0.0001
French	71.9%	0.69 – 0.74	0.0216

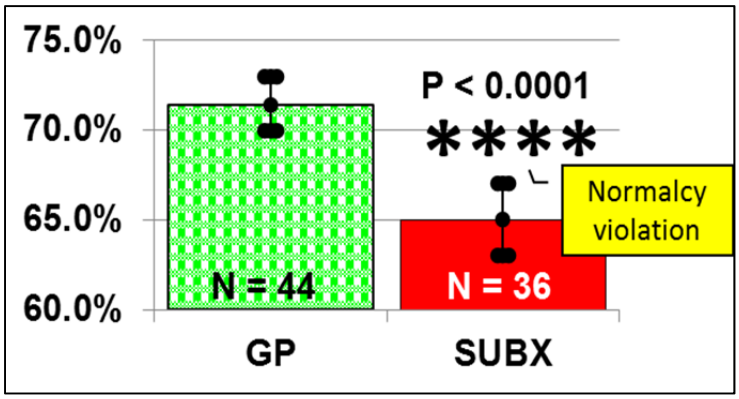


Figure 112 SUBX more Confusable than GP

SUBX are more confusable (p<0.0001) (65.0%; CI: 0.63 – 0.67) than the GP (71.4%; CI: 0.70 – 0.73). Note: the normalcy assumption was violated; the results are not conclusive.

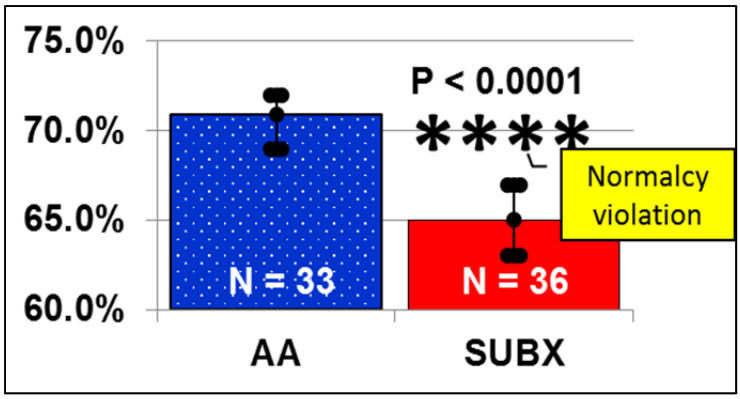


Figure 113 SUBX more Confusable than AA

SUBX are more confusable (p<0.0001) (65.0%; CI: 0.63 – 0.67) than AA members (70.9%; CI: 0.69 – 0.72). Note: the normalcy assumption was violated; the results are not conclusive.

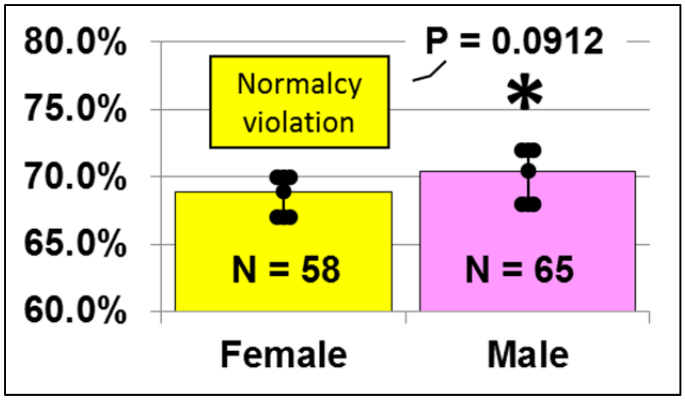


Figure 114 Females more Confusable than Males

There is a trend that Females are more confusable (p=0.0912) (68.9%; CI: 0.67 – 0.70) than Males (70.4%; CI: 0.68 – 0.72). Note: the normalcy assumption was violated; the results are not conclusive.

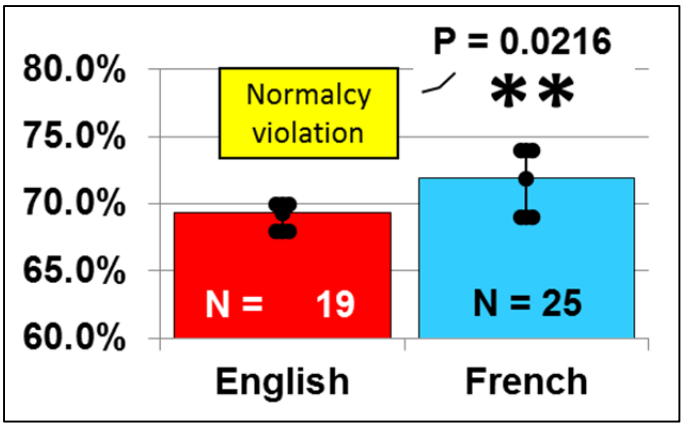


Figure 115 English more Confusable than French

English are more confusable (p=0.0216) (69.3%; CI: 0.68 – 0.70) than French people (71.9%; CI: 0.69 – 0.74). Note: the normalcy assumption was violated; the results are not conclusive.

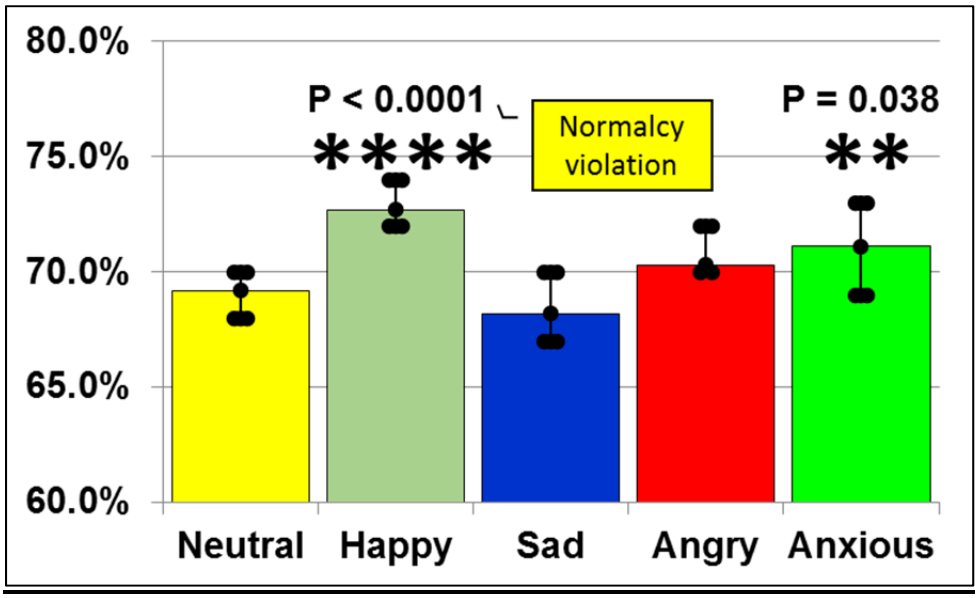


Figure 116 Neutral more Confusable than Happy and Anxious

Figure 116 shows that Neutral (69.2%; CI: 0.68 – 0.70) is more confusable than Happiness (p < 0.0001) (72.7%; CI: 0.72 – 0.74) and Anxiety (p=0.038) (71.1%; CI: 0.69 – 0.73). Note: the normalcy assumption was violated; the results are not conclusive.

5.14 Call rate Analysis

The population average for call completion is 40%. Call completion is a measurement of whether a participant answered the call, and completed the ESM dialogue.

Table 47 Call Completion Effects

Fixed effect	**Probability**	**95% CI**	**p-value Pr(>|z|)**
Formula: complete~ (1|p)			
Null model mean	40.0%	33.6 - 46.7	0.00321
Formula: complete~group3+(1|p) …base-level = GP			
GP	56.7%	45.6 – 67.2	0.228
SUBX	18.9%	10.8 - 30.8	<0.0001
Formula: complete~group3+(1|p) …base-level = AA			
AA	49.3%	45.6 – 67.2	0.903
SUBX	18.9%	10.8 - 30.8	<0.0001

Although the regression intercepts in Table 47 are not significant, there is an indication that SUBX patients completed less calls than the GP or AA members. An observation from Dr. Moehs was that there are SUBX patients who covertly continue to use Opioids while under Suboxone treatment; timing usage to avoid urine detection. Typical opioid detection times [111] are listed in Figure 117.

Buprenorphine	Up to 11 days
Codeine	2 to 4 days
Fentanyl	2 to 3 days
Hydrocodone	2 to 4 days
Hydromorphone	2 to 4 days
Meperidine	2 to 4 days
Methadone	Up to 14 days
Morphine	2 to 4 days
Oxycodone	2 to 4 days
Oxymorphone	2 to 4 days
Propoxyphene	Up to 7 days
Tramadol	2 to 4 days
6-acetylmorphine	Less than 8 hours

Figure 117 Typical Opioid Detection Times in Urine

It is speculated that Opioid usage may be linked to call avoidance. Figure 118 is a SUBX patient's call completion rates. It is evident that there was a lapse. This lapse is observable in many Opioid-Suboxone subjects.

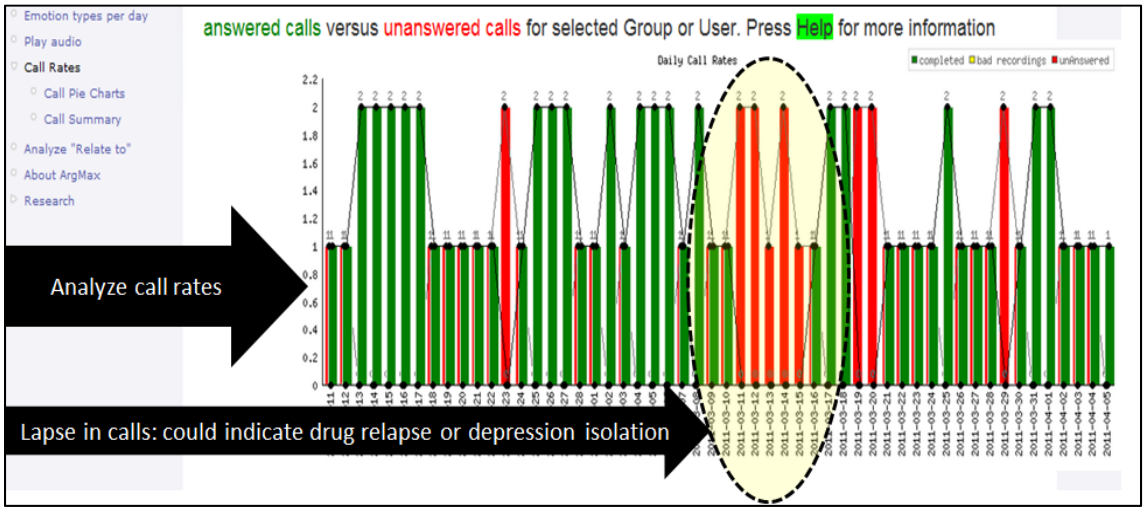


Figure 118 Lapse in Daily Call Completion for SUBX patient

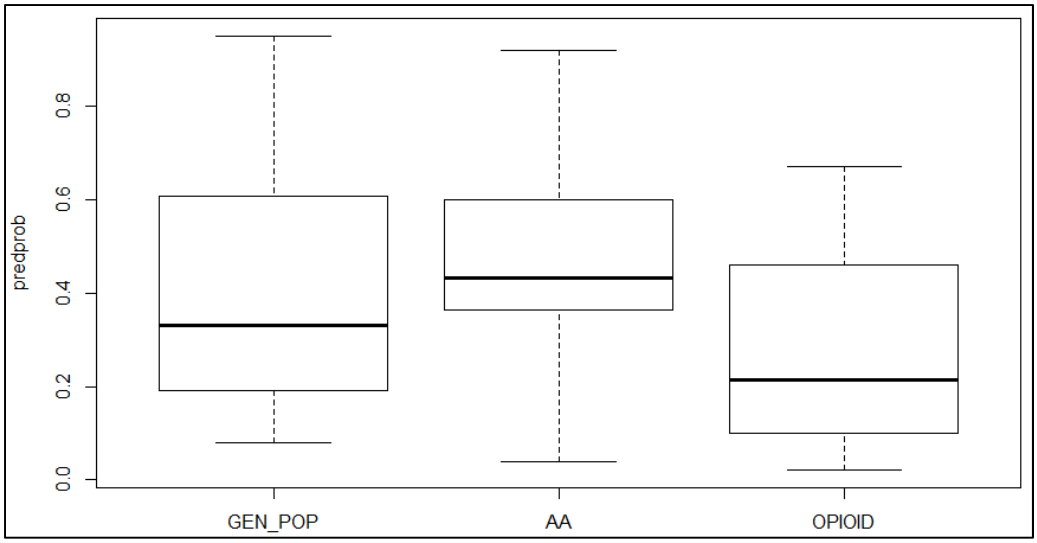


Figure 119 Predicted Probabilities of Call Completion versus Group

Figure 119 provides a boxplot of call completion ~ group3 predicted probabilities. The large IQR for the General Population could be related to the polarity of participants who either liked or disliked taking calls; AA members were more consistent (hence the smaller IQR), especially those that found the calls therapeutic (see the post-trial survey comments in section 1.5.3). It is speculated that SUBX patients' call completion IQR may be related to a participant's commitment to "kicking the habit" and going to any lengths to get better.

5.14.1 Emotional Health Trial Survival Analysis

In the context of trial participation, "survival" is a measurement of whether a participant completed the trial; or quit before the trial end.

5.14.1.1 Emotion collection call data

Nineteen thousand five hundred and thirty–nine (19539) telephone calls were made to the 129 trial participants. 8266 calls were successfully completed (an emotional recording was collected, and a self-assessment was made). There were three general date intervals for data collection, as shown in Table 48.

Table 48 Trial Dates and 60-Day Normalization Factor

Trial	start date	end date	total days	normalize factor (to 60 days)
Trial 1 (AA members & English General Pop)	8/16/2010	9/25/2010	40	1.5
Trial 2 (French General Pop)	11/10/2010	2/1/2011	83	0.722891566
Trial 3 (Opioid addicts on Suboxone)	2/8/2011	4/1/2011	52	1.153846154

5.14.1.2 Trial data normalization

The initial goal of this thesis' trial design was to collect as much data as possible for the purpose of automatic emotion detection model training. A new user could be added and

configured in less than 30 seconds. Thus trials started at different times; trial durations were as long as the majority of participants were willing to participate; participants were added after trial start dates, and some participants continued beyond trial stop dates. The disadvantage of this method is that longitudinal survival analysis is not directly possible. In order to perform a Kaplan-Meier estimate, normalization of trial data is required.

The first step is to normalize the trial durations. Table 48 provides the approximate start and end dates for the three main data collection trials. A factor is calculated to group normalize the trials to 60 days (arbitrarily chosen). The next step is to normalize for participants added after trial start dates, and for participants who continued beyond trial stop dates. There is no survival penalty for either of these cases. We want to measure participants who quit the trial before it ended. Let $t(E_i)$ denote the trial stop date for trial i and D_i the duration of the trial in days.

Participant's p_j actual stop date is denoted as $t(e_j)$. $d_j = [t(S_i) + \lceil t(e_j) \rceil]$ is the number of days the participant partook in the trial and is what we want to normalize and compare. $past\_end = \lceil t(e_j) - t(E_i) \rceil$ is the number of days beyond the trial end that the participant continued in the trial. There is no survival penalty for this case, and therefore $\lceil t(e_j) - t(E_i) \rceil$ is a normalization factor for p_j.

Let $t(S_i)$ denote the trial start date for trial i. Participant's p_j actual start date is denoted as $t(s_j)$. There is no survival penalty. $delay\_start = [t(S_i) - t(s_j)]$ is a normalization factor for p_j.

Table 49 Normalization of Trial Data

idUser	group3	start	end	actual days	delay_ start	past_ end	norm_factor	Di	D60
1	AA Member	8/2/2010	9/25/2010	54	-14	0	0.740740741	40	60
7	AA Member	8/16/2010	9/25/2010	40	0	0	1	40	60
16	AA Member	8/16/2010	9/30/2010	45	0	5	0.888888889	40	60
33	AA Member	8/16/2010	8/23/2010	7	0	0	1	7	10.5
55	AA Member	8/26/2010	9/23/2010	28	0	0	1.333333333	37	55.5
73	General Pop	11/10/2010	1/27/2011	78	0	0	1	78	56.4
80	General Pop	11/10/2010	4/14/2011	155	0	72	0.535483871	83	60
103	Opioid-Suboxon	2/8/2011	3/19/2011	39	0	0	1	39	45
105	Opioid-Suboxon	2/8/2011	4/5/2011	56	0	4	0.928571429	52	60

Table 49 provides some records from the trial data. Example calculation of group normalization is as follows:

- Participant (1) started the trial 14 days early, and finished the trial on time. Participant (1) days are trial normalized from 54 to 40 (column Di) and group normalized from 40 to 60.
- Participant (16) started the trial on time, and finished the trial 5 days past the trial end. Participant (7) days are trial normalized from 45 to 40 (column Di) and group normalized from 40 to 60.
- Participant (33) started the trial on time, and quit before trial end. Participant (33) days are group normalized from 7 to 10.5.

5.14.1.3 Kaplan-Meir Survival analysis

```
csfit <- survfit(Surv(D60) ~ 1, data = CS)
summary(csfit)
plot(csfit,ylab="survival probability",xlab="Trial participation (days)")
```

Code Snippet 26 Kaplan-Meier Survival Analysis of Trial Participation in R

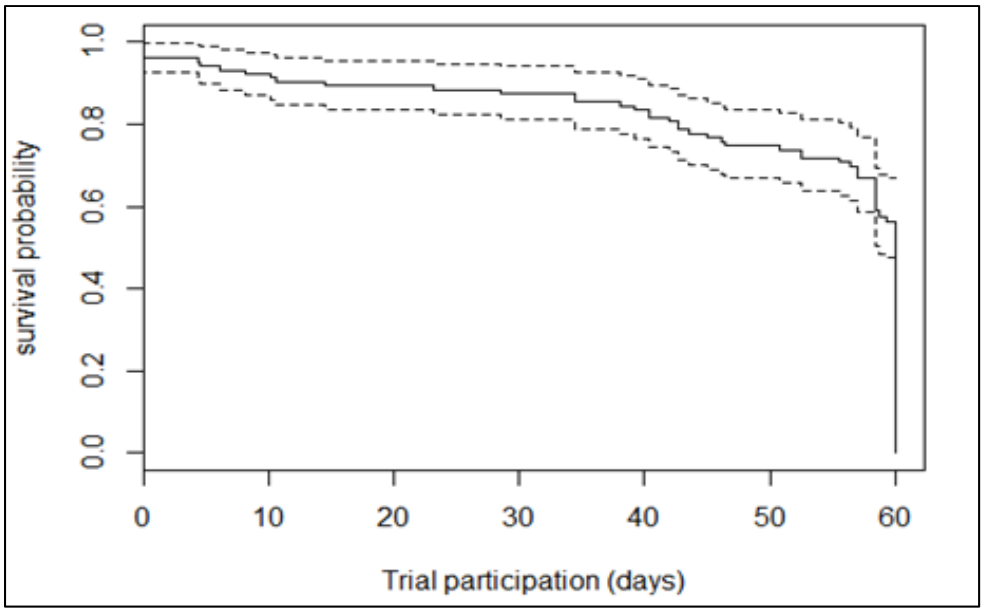


Figure 120 Kaplan-Meier Survival Analysis of Trial Participation

The graph in Figure 120 displays the expected survival probability over the length of the normalized trial. The dotted lines are the 95% confidence intervals. There is a 56.3% probability that a participant will finish the trial.

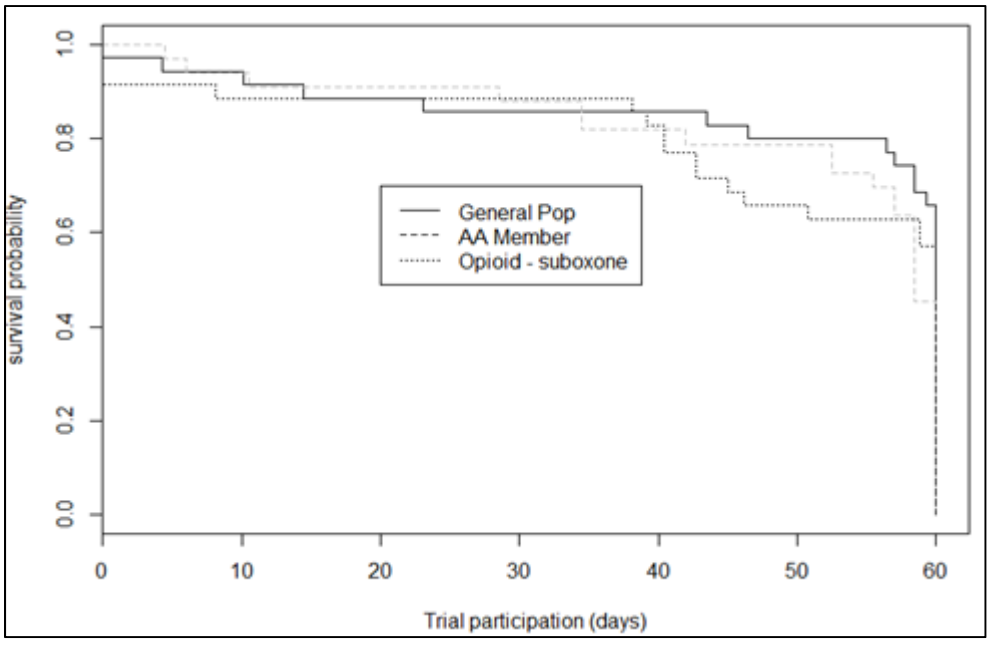


Figure 121 Kaplan-Meier Survival Analysis of Trial Participation by Group

Figure 121 gives the impression that SUBX patients who quit the trial, quit earlier than AA Members and GP. In order to assess the statistical significance, a log-rank test is required.

```
> survdiff(Surv(fdiff) ~ group3, data = CS)

Call:
survdiff(formula = Surv(fdiff) ~ group3, data = CS)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
group3=General Pop     35       35     39.2   0.45392    1.7648
group3=AA Member       33       33     29.2   0.49864    1.5402
group3=Opioid-Suboxone 35       35     34.6   0.00473    0.0171

 Chisq= 2.2  on 2 degrees of freedom, p= 0.334
```

Code Snippet 27 Kaplan-Meier Survival Analysis of Trial versus Group in R

The Chi squared p-value > 0.05, indicating there is no significant difference between groups.

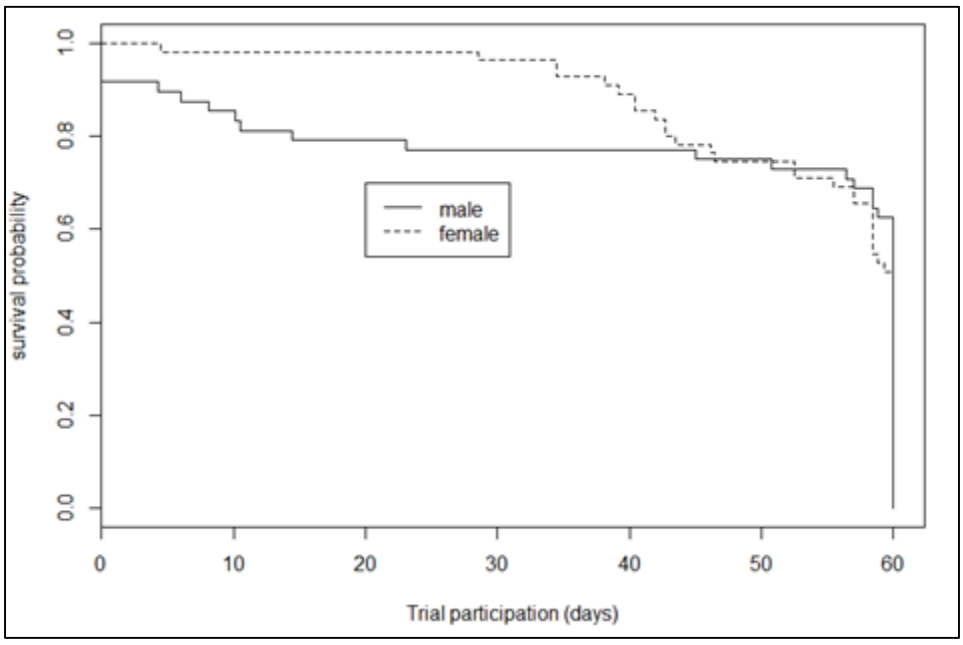


Figure 122 Kaplan-Meier Survival Analysis of Trial Participation by Gender

Figure 122 leads to the impression that most males who quit the trial, decide so quickly as compared to females who persevered longer.

```
> survdiff(Surv(fdiff) ~ gender, data = CS)
Call:
survdiff(formula = Surv(fdiff) ~ gender, data = CS)
```

```
          N Observed Expected (O-E)^2/E (O-E)^2/V
gender=F 48       48     50.5     0.127      0.59
gender=M 55       55     52.5     0.122      0.59

 Chisq= 0.6  on 1 degrees of freedom, p= 0.442
```

Code Snippet 28 Kaplan-Meier Survival Analysis of Trial versus Gender in R

However, there is no significant difference in gender survival rates.

5.15 Discussion

There is an inference from the call analysis in section 5.14 that SUBX patients may covertly continue to misuse licit and illicit drugs during treatment, timing their usage to avoid urine detection. SUBX patients were tested on a scheduled monthly basis. In urine the detection time of chronic opioid users is 5 days after last use [112]. A patient may correctly anticipate that once a urine specimen has been obtained in a certain calendar month, no further specimen will be called for until the next month. Indeed, we have heard from many patients that they understand this only too well—that they have a ‘‘free pass’’ until the next month [112].

Long-term SUBX patients are significantly less happy than the GP ($p=0.0171$) and AA members ($p=0.0310$). There is evidence that SUBX patients perceive others as neutral more often than AA members ($p=0.0223$), and feel themselves as neutral (self-awareness) incorrectly more often than the GP ($p=0.0083$). These observations suggest SUBX patients are living with flat affect.

Long-term SUBX patients are less expressive ($p< 0.01$), and have less self-awareness of being happy, sad, and anxious compared to both the GP and AA groups ($p<0.05$). According to Dr. Blum [19], this motivates a concern that long-term SUBX patients have a diminished ability to perceive “reward” (an anti-reward effect [113]) and may misuse psychoactive drugs, including opioids, during their recovery process. We are cognizant that patients on opioids, including SUBX and methadone, experience a degree of depression and are in some

cases prescribed anti-depressant medication. The resultant flat affect reported herein is in agreement with the known pharmacological profile of SUBX [19].

We did not monitor the AA group participants in terms of length of time in recovery in the AA program and this may have an impact on the results obtained. If the participants in the AA group had been in recovery for a long time the observed anxiousness compared to the SUBX group may have been reduced. However, according to Dr. Blum [19], it is well-known that alcoholics are unable to cope with stress and this effect has been linked to dopaminergic genes [114].

It is well known that individuals in addiction treatment and recovery clinics tend to manipulate and lie not only about the licit and or illicit drugs they are misusing, but also their emotional state [115]. However, it is speculated that these individuals could be flagged through call rate analysis and irregular emotional health patterns. For example a patient consistently reporting that they are OK or happy, violates the normalcy of at least 20% negative emotions concluded by Lyubomirsky et al. [23]. Repeating emotional health measurement experiments in conjunction with proven objective methodologies could provide further validation of the emotional health toolkit for clinical efficacy. Dr. Blum has proposed combining expert assessments of psychological state, advanced urine screening, known as Comprehensive Analysis of Reported Drugs (CARD) [116] determination of affective states as a “true ground emotionality” to compare to the Emotional Health toolkit.

5.16 Conclusion

The null hypothesis that there are no differences in happiness, self-awareness, empathy, or affect between the SUBX group and the General Population is rejected.

Long-term SUBX patients are significantly less happy than the GP ($p=0.0171$) and AA members ($p=0.0310$). There is evidence that SUBX patients perceive others as neutral more often than AA members ($p=0.0223$), and feel themselves as neutral (self-awareness)

incorrectly more often than the GP ($p=0.0083$). These observations suggest SUBX patients are living with flat affect. In addition, SUBX patients are less expressive ($p< 0.01$), and have less self-awareness of being happy, sad, and anxious compared to both the GP and AA groups ($p<0.05$).

- The flat affect and lower happiness is in agreement with the known pharmacological profile of SUBX [19]. Opioid addicts on methadone are less reactive to mood induction. Methadone blunts both elative and depressive emotional reactivity [12]. Patients on opioids, including SUBX and methadone, experience a degree of depression and are in some cases prescribed anti-depressant medication [19].

- The lowered self-awareness of SUBX is in agreement with Scott's conclusion that most chemically-dependent individuals cannot identify their feelings and do not know how to express them effectively [2].

This corroboration of results provides compelling evidence that capturing and measuring Emotional Health in speech can provide a mechanism for Medical Doctors and Psychiatrists to measure the effectiveness of psychotropic medication, and to provide evidence of psychotherapy effectiveness in mental health and substance abuse treatment programs. The acceptance of Emotional Health Toolkit statistical analysis results in the rigorously peer-reviewed science journal PLOS ONE [19] further reinforces the validity of the toolkit to measure clinic efficacy of emotional health.

CHAPTER 6

MULTINOMIAL MULTILEVEL ANALYSIS

Including categorical variables as independent variables in regression models involves constructing dummy variables corresponding to different categories of the independent variable. For example, a 3-category variable D can be split into $D1, D2, D3$ and analyzed by $Y_{ij} = \beta_{0j} + \beta_{1j}D2_{ij} + \beta_{1j}D3_{ij} + \varepsilon_{ij}$ (D1 is the reference level, and thus not included in the equation). Analysis of variance can proceed as described in section 1.7.14.

Including categorical variables as dependent outcome variable is a more difficult problem. Emotional Categorical variables are dependent discrete-choice outcome variables (a.k.a. unordered polytomous variables). The standard statistical model for discrete-choice is logistical regression; where each binomial choice is split out from the multinomial category (e.g. eCROWD split into neutral, happy, sad, angry, and anxious binomials) and independent logistical regressions are performed on each binomial as described in section 1.7.20. The categorical variable eCROWD depends on the mutual exclusivity of choices and as such violates IIA [77] where the odds of preferring one class over another do not depend on the presence or absence of other "irrelevant" alternatives.

The sum of binomial split-out binomial means should equal 100%. However, as shown in Table 37 the sum over all binomials is 88%.

The multilevel multinomial logit model enables the analysis of discrete-choice dependent variables accommodating dependence at unit and cluster levels [78]. There were two software packages experimented: Stata's Generalized Linear Latent And Mixed Models (GLLAMM) [117] and R's Markov Chain Monte Carlo Generalized Linear Mixed Models (MCMCglmm) [79]. GLLAMM was abandoned due to bugs in the tutorials. MCMCglmm multinomial emotion modeling results will be compared to the individual and independent emotion binomial results.

The prior specification is passed to MCMCglmm via the argument **prior** which takes a list of three elements specifying the priors for the fixed effects (B), the G-structure (G) variance-covariance matrix for the random effects, and the residual variance R-structure (R). Default priors were initially attempted but led to both inferential and numerical problems resulting in MCMCglmm failing to converge.

A tutorial from Florian Jaeger [118] set the priors: "The R-structure in this case is set to have a fixed form (fix = 1). Each observation is a single sample from a distribution over k categorical outcomes; we can't actually estimate the residual variance because it depends entirely on the mean (think of the binomial distribution in the simplest case). We follow the prescriptions from the package authors for working with data from this type of distribution, fixing the variance to be 1 for all of the diagonal terms (variances) and .5 for all of the off-diagonal terms (covariance)."

```
> library(MCMCglmm)
> library(spidadev)
> EMO <- THloadEMO() # load the data
> #EMO <- read.csv("2012_FINAL_DATA/EMOgllamm.csv")
> #EMO$eCROWD <- relevel(EMO$eCROWD,ref="OK")
> # make sure there are no null rows (MCMCglmm does not like this)
> EMO <- EMO[!is.na(EMO$group3),]
>
> k <- length(levels(EMO$eCROWD))
> I <- diag(k-1)
> J <- matrix(rep(1, (k-1)^2), c(k-1, k-1))
>
> # prior values
> bp <- list(
+   R = list(fix=1, V=0.5 * (I + J), n = 4),
+   G = list(
+     G1 = list(V = diag(4), n = 4),
+     G2 = list(V = diag(3), n = 3)))
```

Code Snippet 29 MCMCglmm Priors Calculation in R

The MCMCglmm algorithm is then executed with a 2-level grouping under each participant p with dependent outcome variable eCROWD with choices OK, HAPPY, SAD, ANXIOUS, ANGRY.

"The ***burnin*** argument sets the number of samples generated in the burn-in period. The goal is to converge on a set of stable estimates for the model parameters. We know that this won't

happen instantaneously, and so we allow the model some time to try to settle on a good set of parameters before we start collecting samples. These initial throwaway samples comprise the burn-in period, and the number of such samples to be generated is set by the ***burnin*** argument. After the burn-in period, we sample each parameter from the model a certain number of times, controlled by the *nitt* argument. There's no guarantee that the model will have settled on stable estimates by the end of burn-in, though, so it will be important to test whether the model has converged in the samples we end up using. It should also be noted that it is possible to discard samples as they're collected using the thin parameter. This is usually done to try to minimize dependence between samples" [118]

```
> EMO$eCROWD <- relevel(EMO$eCROWD,ref="OK")
> # run MCMCglmm
> m <- MCMCglmm(eCROWD ~ -1 + trait-1,
+               random = ~ us(trait):p + us(group3):p,
+               rcov = ~ us(trait):units,
+               prior=bp,
+               burnin = 15000,
+               nitt = 80000,
+               family = "categorical",
+               data = EMO)
```

Code Snippet 30 MCMCglmm Emotional Truth versus Group in R

```
                      MCMC iteration = 0
  Acceptance ratio for latent scores = 0.000278
. . .
                      MCMC iteration = 79000
  Acceptance ratio for latent scores = 0.309923
                      MCMC iteration = 80000
  Acceptance ratio for latent scores = 0.309415

burnin = 15000,   nitt = 80000,
    traiteCROWD.SAD   traiteCROWD.ANGRY traiteCROWD.ANXIOUS   traiteCROWD.HAPPY
         0.15966442          0.06496719          0.05634249          0.28867925

Iterations = 15001:79991 Thinning interval = 10 Number of chains = 1
Sample size per chain = 6500
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
                       Mean     SD Naive SE Time-series SE
traiteCROWD.SAD     -1.6697 0.1406 0.001744       0.006050
traiteCROWD.ANGRY   -2.6608 0.1798 0.002230       0.009796
traiteCROWD.ANXIOUS -2.8708 0.1713 0.002124       0.011138
traiteCROWD.HAPPY   -0.9279 0.1193 0.001480       0.003737

2. Quantiles for each variable:
                      2.5%    25%    50%     75%   97.5%
traiteCROWD.SAD     -1.946 -1.766 -1.667 -1.5700 -1.4014
traiteCROWD.ANGRY   -3.021 -2.776 -2.659 -2.5385 -2.3186
traiteCROWD.ANXIOUS -3.209 -2.985 -2.866 -2.7540 -2.5399
traiteCROWD.HAPPY   -1.157 -1.009 -0.927 -0.8484 -0.6946
```

Code Snippet 31 MCMCglmm Emotional Truth versus Group Results in R

Results in Code Snippet 31 indicate convergence after 80000 iterations. The trait OK is the base trait, so it is not shown in the results. All trait probabilities add up to 100%. The results of the generalized multilevel linear regression for each null-level emotion binomial are listed in Table 37 are compared in Table 50.

Table 50 MCMCglmm Categorical Means versus Glmer Binomial Means

Emotional Truth	MCMCglmm approximation	glmer Binomial approximation	Ratio (glmer/MCMCglmm)
SAD	15.97%	11.6%	72.6%
ANGRY	6.50%	4.4%	67.7%
ANXIOUS	5.63%	3.9%	69.3%
HAPPY	28.87%	21.5%	74.5%
NEUTRAL	43.03%	46.6%	108.3%
total	100.00%	88.0%	

The glmer binomial approximations exhibit a fairly consistent ratio of approximately 70% of the MCMCglmm results across SAD, ANGRY, ANXIOUS, and HAPPY. The Neutral approximation is closer.

The differences could be attributed to the small emotional corpus. For small data sets the random effects variances are difficult to estimate and require a large number of iterations to converge (80000 iterations for the emotional data set from 126 participants compared to 3000 iterations for the 9205 patients in the IMPACT (International Mission on Prognosis and Clinical Trial design in TBI) set [119]. To improve the MCMCglmm approximation, we would need better priors and more data.

6.1 Conclusion

Emotional Categorical variables are dependent discrete-choice outcome variables. Splitting out binomials from a discrete-choice outcome variable violates IIA [77] where the odds of preferring one class over another do not depend on the presence or absence of other "irrelevant" alternatives.

It is statistically valid to conclude differences in means based on analysis of variance of split-out binomials, but invalid to report actual means. This is evident from Table 37 that shows the sum over all split-out eCROWD binomials is 88%, rather than 100%.

MCMCglmm produces valid means for each discrete-choice outcome. However, a larger corpus is required to improve the approximation results.

GENERAL CONCLUSION

This thesis described the development and validation of an evidence-based toolkit that captures a patient's emotional state during a fifteen second telephone call, and then accurately measures and analyzes indicators of Emotional Health based on emotion detection in speech, majority voting, and multilevel regression analysis. The results, in terms of the goals outlined in the Introduction on page 1, are as follows:

1. **Build an Interactive Voice Response (IVR) cloud platform to monitor and analyze the emotional health of a patient in their natural environment.**

RESULTS: ACHIEVED. This IVR ESM system is the only method known that can measure emotional truth, self-awareness, expressiveness, affect, and empathy from speech. The system is based on ESM; the best method to collect momentary emotional states in a person's natural environment [45]. A patient is registered for daily ESM telephone calls in under 30 seconds. Subject burden is low by limiting call duration to as little as fifteen seconds, and providing an intuitive user interface design that eliminate the need for training. Calling subjects at times of their convenience further maximizes compliance rates. Trial data analysis indicates a 40% overall call completion mean and a 56.3% probability that a participant will record ESMs for at least 60 days.

IVR momentary emotional state capture is universally accessible and avoids deployment costs associated with self-report systems on smartphones. There are five Billion mobile and phone users worldwide; only 1.5 Billion have access to a smartphone [48]. To deploy on all smartphones natively, you must build Apple iOS, Android, Blackberry, Symbian, etc. applications; which is expensive. Providing patients with a smartphone is also expensive; typical unit cost is $500 with reoccurring monthly telephony carrier charges of $30 or more. Furthermore, a severely afflicted addict may sell their smartphone for drugs.

2. **Sample, capture, and collect an emotional speech corpus of sufficient size to enable measurement and statistical analysis.**

RESULTS: ACHIEVED. Eight thousand three hundred and seventy-six (8,376) momentary emotional states were collected from 2010 to 2011, from one hundred and thirteen (113) participants including three groups: Opioid Addicts undergoing Suboxone® treatment at Dr. Charles Moehs MD MPH clinic (Occupational Medicine Associates of Northern New York) N = 36 [13 men; Expressions = 1054] with an average SUBX continued maintenance period of 1.66 years (Standard Deviation (SD) = 0.48); General Population (GP), N = 44 [15 men; Expressions = 2440]; and Alcohol Anonymous (AA), N = 33 [29 men; Expressions = 3848]. There are statistical significant differences in Emotional Truth, Expressiveness, Affect, Self-Awareness, and Empathy across group, gender, and language. These results move the toolkit towards clinical efficacy and acceptance as a tool for Physicians and Psychotherapists.

3. **Devise an unsupervised crowd-sourced emotion corpus labeling technique.**

RESULTS: ACHIEVED. Fused classifiers based on crowd-sourced majority voting from anonymous, professional transcribers, and self-assessment with optimized weighting were developed to approximate emotional truth of an audio recording for emotion detection algorithm training and statistical analysis.

Unsupervised emotional truth corpus labeling requires automatic chunking of audio into an utterance with a single emotion, and unsupervised automatic emotional truth labeling. Automatic chunking is implemented and verified. The anonymous crowd-sourced MV classifier has 70% accuracy as compared to the ground truth of the fused anonymous-transcriber-self MV classifier. This accuracy is promising, considering the accuracy measurements are based on only 2132 recordings in the corpus with 3 or more anonymous votes, and votes from SUBX patients and AA members were included.

4. **Accurately measure the emotional health of a patient over time.**

RESULTS: PARTIALLY ACHIEVED. Is the $e_{ij}^{detect}(X_{ij})$ performance at 41.92% good enough for reliable detection of emotional truth? We discovered in Chapter 3 the following facts:

- The 5-class winner from INTERSPEECH 2009 had an accuracy was 41.65% [97].
- Performance of human classification of utterances into six classes was 65.7% [42].
- Emotion labeler agreement in most cases is 3 out of 5. This equates to 60% [43].
- The commercialization threshold for automatic classification systems is 80% [41].
- The concordance of $e_{ij}^{self}(X_{ij}) == e_{ij}^{\widehat{truth}}(X_{ij})$ was 61.66% (section 3.8.1.1) .

The conclusion is that at least 60% accuracy, approaching humans, is required. 41.92% emotion truth accuracy is not sufficiently reliable for clinical patient monitoring or to establish clinical efficacy through statistical analysis. By fusing Majority Vote classifiers to $e_{ij}^{detect}(X_{ij})$ we were able to achieve reliability.

5. **Devise a real-time auditable approach to emotional health measurement for monitoring. This method will improve the accuracy of measurements as reinforcement data becomes available; and provide a score to indicate the certainty of the measurement.**

RESULTS: ACHIEVED. Professional intervention can be triggered on patient trends such as missed call rates and multiple ESMs containing negative emotions. Trend detection windows would logically be over a period of at least two consecutive days.

Emotional trends are dependent on emotional truth accuracy. Accuracy can be incrementally improved over time, as new data becomes available. The pseudo real-time classifier can provide a preliminary accuracy of 42% (UA). Accuracy can be maximized within a few days, as demonstrated during the emotion trial to measure the validity of the automatic emotion detector in detecting mood predictive of performance on the Iowa gambling task conducted

by Ogura et al. [110], which should provide ample time to trigger professional intervention on negative emotions.

Emotional truth accuracy is not a black and white measurement and in some cases nondeterministic; people have flat affect which confuses emotional truth (a statement that will be proven in section 5.14). Certainty scores and confusability scores provide a good indication of emotional truth accuracy.

6. **Evidence-based practices are interventions for which there is consistent scientific evidence showing that they improve client outcomes [8]. In general the highest standard is several randomized clinical trials comparing the practice to alternative practices or to no intervention [8]. A key outcome of this thesis is to provide statistical evidence that capturing and measuring Emotional Health in speech can provide a mechanism:**
    a. **To assist Cognitive Behavioural Therapy (CBT) for psychiatrists and therapists and patients to become aware of symptoms and make it easier to change thoughts and behaviors;**

    b. **For evidence of psychotherapy effectiveness in mental health and substance abuse treatment programs;**

    c. **For Medical Doctors and Psychiatrists to measure the effectiveness of psychotropic medication.**

RESULTS: ACHIEVED. The null hypothesis that there are no differences in happiness, self-awareness, empathy, or affect between the SUBX group and the General Population is rejected.

Long-term SUBX patients are significantly less happy than the GP ($p=0.0171$) and AA members ($p=0.0310$). There is evidence that SUBX patients perceive others as neutral more

often than AA members ($p=0.0223$), and feel themselves as neutral (self-awareness) incorrectly more often than the GP ($p=0.0083$). These observations suggest SUBX patients are living with flat affect. In addition, SUBX patients are less expressive ($p< 0.01$), and have less self-awareness of being happy, sad, and anxious compared to both the GP and AA groups ($p<0.05$).

- The flat affect and lower happiness is in agreement with the known pharmacological profile of SUBX [19]. Opioid addicts on methadone are less reactive to mood induction. Methadone blunts both elative and depressive emotional reactivity [12]. Patients on opioids, including SUBX and methadone, experience a degree of depression and are in some cases prescribed anti-depressant medication [19].

- The lowered self-awareness of SUBX is in agreement with Scott's conclusion that "most chemically-dependent individuals cannot identify their feelings and do not know how to express them effectively" [2].

This corroboration of results provides compelling evidence that capturing and measuring Emotional Health in speech can provide a mechanism for Medical Doctors and Psychiatrists to measure the effectiveness of psychotropic medication, and to provide evidence of psychotherapy effectiveness in mental health and substance abuse treatment programs. The acceptance of Emotional Health Toolkit statistical analysis results in the rigorously peer-reviewed science journal PLOS ONE [19] further reinforces the validity of the toolkit to measure clinic efficacy of emotional health.

RESULTS: ACHIEVED.

7. **Devise patient monitoring and trend analysis tools to provide empirical insight into a patient's emotional health and accelerate the interview process during monthly assessments by overburdened physicians and psychotherapists. Crisis intervention**

can be triggered on conditions including the detection of isolation from unanswered calls, or consecutive days of negative emotions.

RESULTS: ACHIEVED. As described in Appendix B and C, it takes less than 30 seconds to enroll and configure a patient in the emotional health toolkit. Once the toolkit has been configured with all patients, the supervisor can monitor patients and analyze results. Trend analysis charts and views provides professionals with insight in a patient's emotional health between appointments. Audio dialogue prompts are recorded by the supervisor or delegate to promote familiarity for the patients in the system thus increasing their willingness and openness to use the system (verified with Dr. Moehs' patients). This system was tested during data collection by the author, and further validated by Dr. Moehs during the SUBX data collection, and during an emotion trial to measure the validity of the automatic emotion detector in detecting mood predictive of performance on the Iowa gambling task conducted by Ogura et al. [110] from January through March of 2013.

RECOMMENDATIONS

Accurate emotional truth detection is required to accurately determine the emotional health of patients. An international collaboration between ETS, MIT, and MIT Lincoln Labs has been formed to improve acoustic emotion detection accuracy. The emotional health toolkit will assist in collecting data towards emotion detection algorithm and accuracy improvements.

Social and medical sciences recognize statistically significant results as a measure of clinical efficacy. The results of this thesis require further validation through collaboration with experts in the fields of psychiatry, psychology, mental health medicine, and addiction medicine.

A project is planned for 2014 with Dr. Kenneth Blum to determine if the dopamine D2 agonist (KB220Z) increases the level of happiness of addicts in treatment. The results of the Emotional Health Toolkit will be compared to a baseline measurement system that combines expert assessment with advanced urine screening. This system is known as the Comprehensive Analysis of Reported Drugs (CARD) [52] with accurate determination of affective states (“true ground emotionality”).

A possible collaboration with the centre for Studies on Human Stress would measure stress and compare ESM measured stress against the clinically proven stress measurement method of cortisone levels. Trials are in the planning stages for validity with a "normal" group differentiated by personality characteristics and problem gambling tendencies before and after performing the Iowa Gambling Task (IGT); predicting moods shifts predictive of violent outbursts in forensic mental disordered inpatients; Canadian Correctional Service (CSC) study with a focus on detecting neurophysiological differences in inmates with a) different degrees of substance abuse/addictions and b) propensities for aggression. There are possibilities of trials in with the VivaVoz addiction service telephone program with the

Brazilian Government, and as an addiction therapy monitoring tool with the Government of Ireland; and more trials with Dr. Charlie Moehs MD MPH in Watertown New York.

The toolkit should be upgraded to incorporate the statistically significant findings as a baseline indicator for individual measurements comparison. Anxiety, anger, sadness and call completion thresholds could trigger alerts to therapists for intervention.

Internationalization is required for some of the global opportunities (e.g. Portuguese for Brazil, Chinese, French and Spanish).

Longitudinal statistical analysis techniques should be investigated in order to analyze effects. For example, a patient group could be measured before and after the administration of KB220Z.

Electronic Medical Record compliance (e.g. Canada Health Infoway) is required for commercialization.

PUBLICATIONS

Journal Publications

Hill E., Han D., Dumouchel P., Dehak N., Quatieri T., Moehs C., Oscar-Berman M., Giordano J., Simpatico T., Blum K. (2013). *Long Term Suboxone™ Emotional Reactivity As Measured by Automatic Detection in Speech.* PLoS ONE 8(7): e69043. doi:10.1371/journal.pone.0069043.

Hill E., Dumouchel P., Moehs C. (2011). *An evidence-based toolset to capture, measure and assess emotional health.* Studies in Health Technology and Informatics. Retrieved Annual Review of Cybertherapy and Telemedicine 2011 - Advanced Technologies in Behavioral, Social and Neurosciences, 167.

Conference Proceedings

Hill E., Dumouchel P. (2010). *An Evidence-Based Toolset to Capture, Measure, Analyze & Assess Emotional Health.* Canadian Society Addiction Medicine, Charlottetown, PEI.

Hill E., Dumouchel. P. (2012). *An Evidence-Based Toolset to Capture, Measure, Analyze & Assess Emotional Health*. International Conference and Exhibition on Addiction Research and Therapy, Los Vegas, Nevada.

Hill E., Moehs C., Dumouchel P. (2013). *An Evidence-Based Toolset to Capture, Measure, Analyze & Assess Emotional Health.* American Society of Addiction Medicine, 44th Annual Medical-Scientific Conference, Chicago, Illinois.

Invited Speaker

Hill E., Dumouchel P. (2013). *An Evidence-Based Toolset to Capture, Measure, Analyze & Assess Emotional Health.* Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Press Coverage

Louden K. (2013), *Robocalls Flag Recovering Addicts' Relapse Risk*, Medscape.

Hill E., Dumouchel P. (2013). *Un ordinateur qui décode les émotions dans la voix*. Le Code Chastenay, telequebec.tv. http://video.telequebec.tv/video/13373/un-ordinateur-qui-decode-les-emotions-dans-la-voix.

APPENDIX A

MORE ON STEP 3: EMOTIONAL HEALTH ESM

This appendix analyzes and compares pen and pencil, mobile device form entry, IVR questionnaire, and acoustic IVR experience sampling methods.

The conclusion will show that Acoustic IVR has superior ESM capabilities, and is the best method to capture and measure emotional health.

Pen-and-Pencil Journaling

A common ESM in Cognitive Behaviour Therapy is for a patient to maintain a daily journal of the day's events and associated feeling, emotions, and actions. This journal contributes to the therapist's assessment of the patient's cognitive and behavioural health [120].

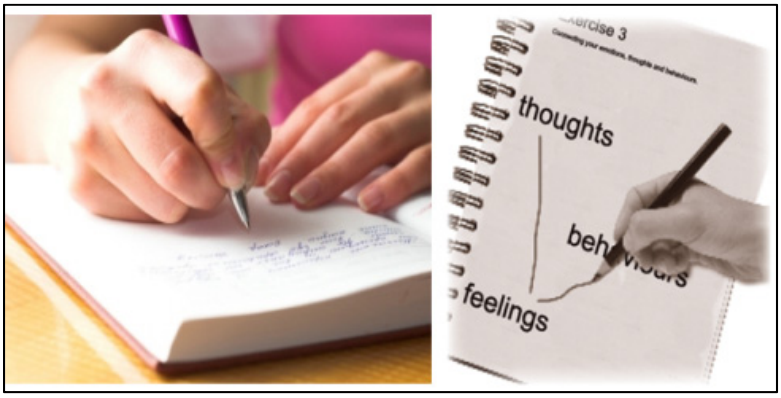


Figure 123 Pen and Pencil Journaling

Table 51 Pen-and-Pencil ESM Capabilities

id	ESM Requirement	Pen-and-Pencil
1	Time sampling ESM	NO. There are no empirical methods to ensure time compliance, or to avoid entry "hoarding"
2	Minimal time per ESM	POOR. A diary entry may typically require 5 minutes of thought and expression from pen to paper
3	Subject burden (PRO length, PRO user interface, and PRO complexity)	POOR. Putting thoughts to paper requires time. People have a tendency to procrastinate or avoid a task if it takes too long. A participant's willingness is negatively impacted (as little as 11% compliance)
4	Anytime, anywhere	POOR. The participant must carry a diary
5	No literacy required	NO. If a person cannot write, they cannot record.
6	Language independence	NO. The person must express themselves in a language the therapist can understand
7	Low cost per unit	BEST. 0$. No cost for this method.
8	Recall situation, behaviour, & feeling associated with ESM	POOR. Difficult to trigger situational memory (situation, behaviour, emotion, mood) from a journal entry.
9	Honesty of ESM entry	POOR. Professional must rely on honesty of participant.
10	Ease-of-deployment	GOOD. Pen or pencil and a notebook. Instructions on when and how to perform ESM.

Table 52 Pen-and-Pencil ESM for Emotional Health

id	Emotional Health Requirement	Pen-and-Pencil
11	Emotional expression	POSSIBLE. Must express feelings in writing. If a person has limited literacy, their journal entries will not correlate with their emotional state.
12	Measure emotional expression	NO. There is no way to accurately measure the actual emotion(s) expressed in the journal entry
13	Self-assessment of emotional state	YES. A journal entry can contain an indication of emotional state
14	Self-assessment verification	NO. There is no method of automatically verifying a patient's ability to identify their emotional self-assessment.
15	Empathy measurement	NO. There is no possibility of capturing or measuring a patient's ability to relate: A key component of Emotional Health.
16	Empathy verification	
17	Complete emotional health capture and measurement	NO. Emotional health capture is incomplete without measuring a patient's ability to identify their own emotion and ability relate to others.
18	Emotional Health analysis	POSSIBLE. However, An average community mental health services' psychiatric follow-up session for a stable patient is estimated at 20 minutes [36]. This affords little time for in-depth analysis of journal data, verification of data validity, or long-term trend analysis.

ESM technology

ESM Technology can capture and measure some emotional and behavioural aspects of the daily journal that may help professionals interpret and assess the emotional health of their patients. Computer-assisted delivery of cognitive-behavioural therapy is effective and an empirically-supported therapy [56].

Research has recently commenced in evidence-based methods to capture and measure momentary emotional state using windows-form mobile devices [121] and IVR systems [122]. These systems can time-sample self-assessment of emotional state, but do not provide

empirical methods to measure a person's ability to express emotions, identify their own emotions, or relate to other people's emotions. In addition, these methods suffer from busy bias, resulting in participation apathy and neglect.

Mobile device ESM

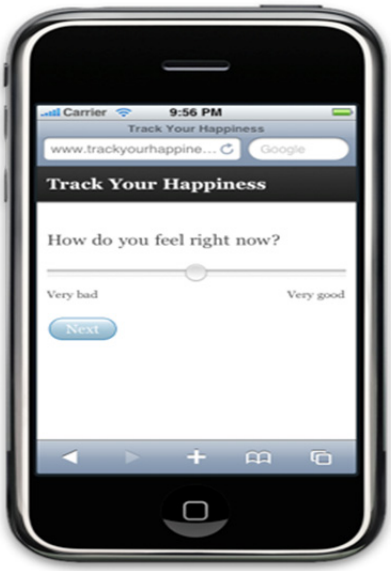


Figure 124 Example of trackyourhappiness.org

Data entry on a smart phone or mobile device can provide momentary emotional state time sampling functionality. However, mobile device data capture suffers from most Pen & Paper deficiencies, with possible clinical efficacy, and a possible reduction in busy bias.

In addition, it is costly to deploy mobile devices on a large scale to low-income people. In chronic mental disorders s well as addiction maintenance and recovery, the majority of patients are predominantly low-income.

However, speech recordings can also be collected on smart phones and smart phones can interact with a server. This device can then support all emotional health suitability aspects acoustic ESM. Physiological approaches (Heart-rate, skin conductance) and multi-modal

(self-assessment combined with Global Positioning System (GPS), movement, etc.) are surfacing in 2012 but are beyond the scope of this thesis.

Table 53 Mobile Device ESM Capabilities

id	ESM Requirement	Mobile device
1	Time sampling ESM	YES. Periodic Beep prompts participant
2	Minimal time per ESM	OK. Time depends on User Interface design, number of questions asked, and input modalities. To capture all elements of emotional health, estimate is minimum 30 seconds
3	Subject burden (PRO length, PRO user interface, and PRO complexity)	POOR. Form-based entry requires hands and eyes. User Interface requires computer skills. Combine with 30+ seconds per ESM. May have an effect on compliance.
4	Anytime, anywhere	POOR. The participant must carry a mobile
5	No literacy required	NO. If a person cannot read and use a computer they cannot record.
6	Language independence	NO. must be able to read the User Interface to comply
7	Low cost per unit	POOR. $500++ per patient
8	recall situation, behaviour, & feeling associated with ESM	POOR. Difficult to trigger situational memory from checkbox. Contextual information would need t be collected as well.
9	Honesty of ESM entry	POSSIBLE. iPhone apps like Auto Lie Detector HD are emerging.
10	Ease-of-deployment	POOR. For those with no mobile device, a device must be supplied & the user trained. For everyone else, must support operating system (i.e. IOS, Android, blackberry, Symbian, etc.) and language

Table 54 Mobile ESM for Emotional Health

id	Emotional Health Requirement	Mobile device
11	Emotional expression	POSSIBLE. No known devices which capture emotional expression (only self-assessment).
12	Measure emotional expression	
13	Self-assessment of emotional state	YES. Emotional state can be entered or chosen from a mobile form
14	Self-assessment verification	NO. There is no method of automatically verifying a patient's ability to identify their emotional self-assessment.
15	Empathy measurement	POSSIBLE. but no known devices which capture assessment of another person's emotional expression
16	Empathy verification	
17	Complete emotional health capture and measurement	POSSIBLE. But without acoustic measurements, Missing emotional expression, self-assessment verification, ability to relate to others' emotion
18	Emotional Health analysis	

IVR Experience Sampling

Time-sampled and self-initiated ESM are performed by scheduled outbound dialling and inbound dialling respectively over the PSTN.

Emotional state is momentary and may not coincide with the time sampled. If the time-based sampling is performed while the patient is in a particular mood, the emotional state can still be captured. Participants can also self-initiate emotional speech registration by calling in to the IVR system. Details on system and software architecture are described in appendix B.

IVR Questionnaire

IVR questionnaire is an ESM approach widely used. There were at least 54 ESM studies based on IVR between 1989 and 2000 alone [122]. This methodology denoted "IVR

checkbox" calls the participant on their telephone, prompts the participant with a series of questions and possible choices, and registers the keypad responses.

Table 55 IVR Questionnaire ESM Capabilities

id	ESM Requirement	IVR checkbox ESM
1	Time sampling ESM	YES. Call Patient on their telephone.
2	Minimal time per ESM	OK. Time depends on Voice User Interface design, number of questions asked, and input modalities. To capture all elements of emotional health, estimate is minimum 30 seconds
3	Subject burden (PRO length, PRO user interface, and PRO complexity)	POOR. Voice and keypad-based entry requires hands and eyes. UI is easy - most know how to operate a phone. Combine with 30+ seconds per ESM. May have an effect on compliance.
4	Anytime, anywhere	BEST. Call-in on any telephone. Mobile or Call forward for incoming calls.
5	No literacy required	YES. Numeric keypad entry.
6	Language independence	YES. Response is Language independent
7	Low cost per unit	< $0.05/minute
8	Recall situation, behaviour, & feeling associated with ESM	POSSIBLE. Record emotional expression. Situational memory can be triggered from ESM audio recording playback.
9	Honesty of ESM entry	POOR. Professional must rely on honesty of participant.
10	Ease-of-deployment	GOOD. Need user's phone number, , call times, language, and any other factors pertinent for the clinical trial (e.g. age, gender, etc.).

Table 56 IVR Questionnaire ESM for Emotional Health

id	Emotional Health Requirement	IVR checkbox ESM
11	Emotional expression	NO. keypad cannot capture emotional expression
12	Measure emotional expression	
13	Self-assessment of emotional state	Yes. The set of Emotional state can be in a voice prompt, and the choice can be entered from a keypad
14	Self-assessment verification	NO. There is no method of automatically verifying a patient's ability to identify their emotional self-assessment.
15	Empathy measurement	POSSIBLE. But no known IVR checkbox applications capture assessment of another person's emotional expression
16	Empathy verification	
17	Complete emotional health capture and measurement	NO. Missing emotional expression, self-assessment verification, ability to relate to others' emotion
18	Emotional Health analysis	

IVR Acoustic ESM

IVR acoustic ESM (A-ESM) can capture an experience sample ESM_{ij} for $participant_j$ during an automated IVR telephone call c_{ij}. The ground truth e_{ij}^{truth} of the emotionally charged speech utterance X_{ij} can then be subsequently calculated. Telephones are universally accessible.

Table 57 IVR Acoustic ESM Capabilities

id	ESM Requirement	IVR A-ESM
1	Time sampling ESM	YES. Call Patient on their telephone.
2	Minimal time per ESM	BEST. An average A-ESM call duration is 12 secs
3	Subject burden (PRO length, PRO user interface, and PRO complexity)	BEST. Short call duration and full hands-free overcomes busy bias (note: hands-free version not used in trial)
4	Anytime, anywhere	BEST. Call-in on any telephone. Mobile or Call forward for incoming calls.
5	No literacy required	YES. voice + numeric keypad
6	Language independence	YES. Response is Language independent. Emotion recognition in speech is language independent. There is no need to express feelings in the therapist's language.
7	Low cost per unit	< $0.05/minute
8	Recall situation, behaviour, & feeling associated with ESM	YES. Record emotional expression. Situational memory can be triggered from ESM audio recording playback.
9	Honesty of ESM entry	POSSIBLE. Lie detection in speech is commercially available
10	Ease-of-deployment	GOOD. Need user's phone number, call times, language, and any other factors pertinent for the clinical trial (e.g. age, gender, etc.).

Table 58 IVR Acoustic ESM for Emotional Health

id	Emotional Health Requirement	IVR acoustic ESM
11	Emotional expression	YES. Record emotionally charged audio elicited form the prompt "how are you feeling?"
12	Measure emotional expression	YES. Measure emotion through crowd-source and automatic detection
13	Self-assessment of emotional state	YES. The set of Emotional state can be in a voice prompt, and the choice can be entered from a keypad or (speech recognition in hands-free mode)
14	Self-assessment verification	YES. The ground truth of the Emotional expression can be calculated and compared to the self-assessment
15	Empathy measurement	YES another person's Emotional expression can be played, and the participant's assessment can be entered by keypad or voice
16	Empathy verification	YES. The ground truth of the other person's Emotional expression can compared to the assessment
17	Complete emotional health capture and measurement	YES. All aspects of emotional health are automatically captured or measured.
18	Emotional Health analysis	YES. A toolkit can analyze groups for clinical efficacy and compare individual scores to group norms.

Comparison of Experience Sampling methods

Pen-and-Pencil, Mobile device, IVR checkbox, and IVR A-ESMs are assessed for their ESM capabilities (summarized in **Table 59**) and their ability to collect all aspects of emotional health (summarized in Table 60). Blackened cells indicated the ESM method is not compliant with the requirement. Greyed cells indicate the ESM method can partially meet requirement. White cells indicate full compliance.

Acoustic IVR has superior ESM capabilities, and is the best method to capture and measure emotional health

Table 59 Comparison of ESM Capabilities across ESM

id	ESM Requirement	Pen-and-Pencil	Mobile device	IVR checkbox ESM	IVR acoustic ESM
1	Time sampling ESM	NO	YES	YES	YES
2	Minimal time per ESM	POOR	OK	OK	BEST
3	Subject burden	POOR	POOR	POOR	BEST
4	Anytime, anywhere	POOR	POOR	BEST	BEST
5	No literacy required	NO	NO	YES	YES
6	Language independence	NO	NO	YES	YES
7	Low cost per unit	$0	$500++	< $.05/min	< $.05/min
8	Recall situation, behaviour, & feeling associated with ESM	POOR	POOR	POSSIBLE	YES
9	Honesty of ESM entry	POOR	POSSIBLE	POOR	POSSIBLE
10	Ease-of-deployment	EASY	POOR	POSSIBLE	EASY

Table 60 Comparison of Emotional Health Measurement Suitability

id	Emotional Health Requirement	Pen-and-Pencil	Mobile device	IVR ESM	IVR acoustic ESM
11	Capture emotional expression	POSSIBLE	POSSIBLE	NO	YES
12	Measure emotional expression	NO			YES
13	Self-assessment of emotional state	YES	YES	YES	YES
14	Self-assessment verification	NO	POSSIBLE	NO	YES
15	Empathy assessment	NO	POSSIBLE	POSSIBLE	YES
16	Empathy verification				YES
17	Complete emotional health capture and measurement	NO	POSSIBLE	NO	YES
18	Emotional Health analysis	POSSIBLE			YES

APPENDIX B

SYSTEM DESIGN

Participant and data management

Participants are signed up, along with the data collection period and call times through a Drupal 6.0 web 2.0 site [123]. Drupal is an open source Content Management Framework (CMF). Unlike a typical Content Management System (CMS), it is geared more towards configurability and customization.

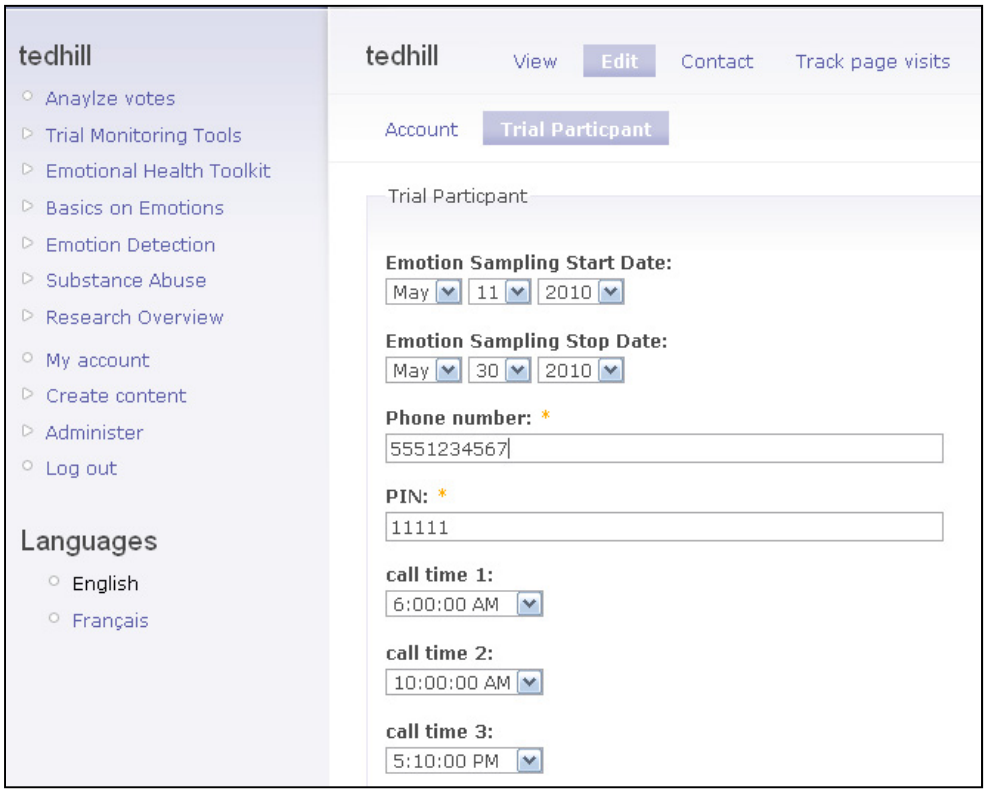


Figure 125 Participant Profile in Drupal

Static content has been created with the Drupal add Page (Node) & Menu Item configuration tools. Dynamic content for the site was created with a combination of HyperText Markup Language (HTML), PHP, Flash and JavaScript. Asynchronous client-side JavaScript (AJAX)

and server-side PHP provide a dynamic user experience during the Flash playback and emotional labeling of recorded audio.

Emotional speech capture and collection

IVR was selected as the best method to automate momentary emotional speech capture.

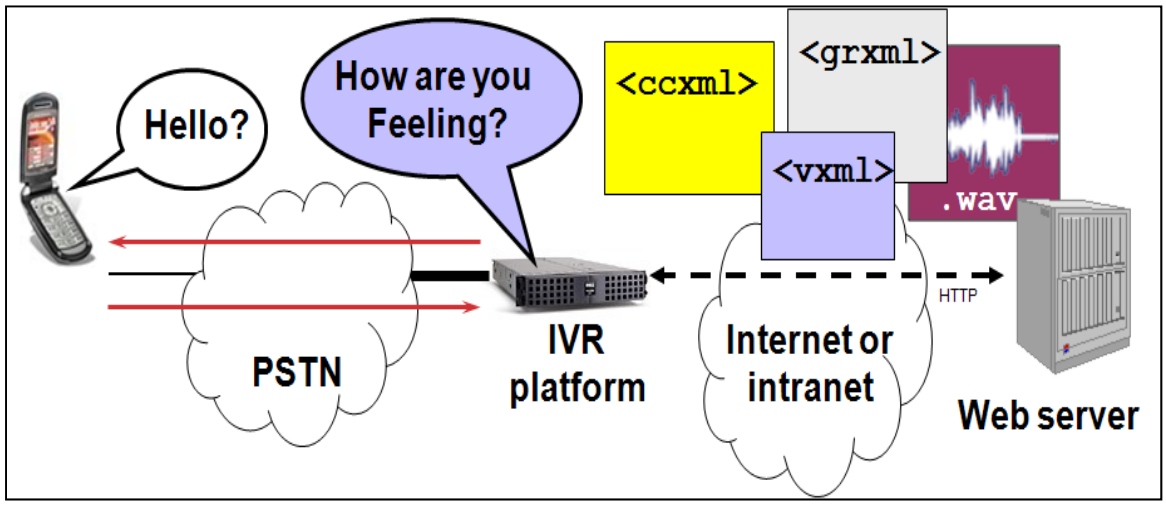


Figure 126 IVR Network Architecture

Both pre-arranged and random time sampling are performed by scheduled outbound dialing over the PSTN through the power of CCXML and CRON. The CRON daemon invokes a PHP script that checks the database for “ripened” call times. Once the call is successfully answered, the VoiceXML application, with speech recognition and DTMF recognition grammars coded in GRXML, is invoked. Call status and user responses are captured to a database indexed by user and timestamp.

The World Wide Web Consortium (W3C) standard CCXML [124] is an event driven markup language that is designed to provide telephony call control support for VoiceXML and has features such as answer machine detection, busy detect, and connection time out (no answer).

A simplified CCXML outbound dialing script is shown on the next page. The <eventprocessor> element is a container for the <transition> elements that drive the CCXML asynchronous execution. When the script is loaded, the *ccxml.loaded* event is triggered, executing the <createcall> outbound dialing element with the passed session parameter *numbertodial,* and a timeout of 30 seconds to abort on no answer. If the call is answered, the *connection.connected* event is triggered, and the dynamic VoiceXML dialogue "vxmlDialog.php" is invoked with the parameter *numbertodial*, which provides reference to the user's database records. Transition events *connection.failed, connection.disconnected, error.*,* allow for more call processing such as call state logging. The user-defined timed event *DIE_ZOMBIE_DIE* created in the <send> element ensures the call does not spin forever due to buggy VoiceXML code or a platform error.

The CCXML answer detection feature was experimented with; however, the algorithm implementation was based on a human's trait to answer with a short interrogative like "Hello?" versus a long preamble from an answer machine. The algorithm gets confused when there is excessive background noise that can occur in public places such as Restaurants. Thus the feature was removed from the CCXML state machine. Instead, the call state is tracked over VoiceXML dialogue legs – logging dialogue progress. If the User does not respond to the first question in the dialogue, it can be assumed the call was unsuccessful, and logged and processed as such. Call completion statistics are then mined from the call state logs.

```
<?xml version="1.0" encoding="UTF-8"?>
<ccxml version="1.0">
 <var name="app_state" expr="'init'"/>
 <var name="numbertodial" expr="session.values.numbertodial"/>
 <eventprocessor statevariable="app_state">
    <transition event="ccxml.loaded">
      <createcall dest="'tel:' + numbertodial"
           callerid="'tel:55555555555'" timeout="'30s'" />
      <send name="'DIE_ZOMBIE_DIE'" target="session.id" delay="'150s'"/>
    </transition>
    <transition event="connection.connected">
    <dialogstart src="'vxmlDialog.php'" type="'application/xml+vxml'"
                  namelist="numbertodial"/>
    </transition>
    <transition event="connection.failed">
      <log expr="'*** CALL FAILED ***'"/>
    </transition>
    <transition event="connection.disconnected"> </transition>
    <transition event="error.*"> </transition>
    <transition event="dialog.exit"> </transition>
    <transition event="DIE_ZOMBIE_DIE"> </transition>
  </eventprocessor>
</ccxml>
```

Code Snippet 32 CCXML Script to Call Participants in CCXML

The W3C standard VoiceXML [125] allows voice applications to be developed and deployed in an analogous way to HTML for visual applications. Just as HTML documents are interpreted by a visual web browser, VoiceXML documents are interpreted by a Voice browser.

VoiceXML can be statically created, or dynamically generated by server-side scripts written in Java, PHP, C# etc. that personalize the interactive voice dialog based on the user's profile extracted from a database. The key VoiceXML elements used in this application are: <prompt>, <audio>, <record>, and <grammar>. The <prompt> and <audio> elements allow for playback of text-to-speech or recorded audio respectively.

```
<?xml version="1.0" encoding="utf-8"?>
<vxml version = "2.1">
<form id="F_1">
  <block>
    <prompt> Hello world </prompt>
     <audio src="MySoundFile.wav"/>
  </block>
</form>
</vxml>
```

Code Snippet 33 Play an Audio Recording over the Phone in VoiceXML

The <record> element allows the participant's response to "How are you feeling?" to be captured into variable, and subsequently saved to disk.

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version = "2.1">
<form id="F1">
  <record name="R_1" beep="true" maxtime="10s" finalsilence="1s">
    <prompt> How are you feeling? </prompt>
    <filled>
      <prompt> your emotion was <value expr="R_1"/> </prompt>
    </filled>
  </record>
</form>
</vxml>
```

Code Snippet 34 Record an Audio Recording over the Phone in VoiceXML

The grammar element allows dialogue interaction with the participant. This dialogue captures the participant's emotional choice:

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version="2.1">
<var name="emotion" expr="'none'"/>
 <form id="slabel">
    <field name="labelEmotion">
        <prompt bargein="true">
         Are you happy, angry, sad, nervous or okay?
        </prompt>
        <grammar src="emotions.grxml" type="application/srgs+xml"></grammar>
        <noinput>
            <prompt> Sorry I cannot hear you.  </prompt>
            <reprompt/>
        </noinput>
        <nomatch>
        <prompt>
            I did not recognize your choice. Please try again.
        </prompt>
        <reprompt/>
        </nomatch>
</field>
    <filled>
      <assign name="emotion" expr="labelEmotion"/>
      <submit next="SaveFeeling.php" method="get" namelist="userid emotion"/>
    </filled>
  </form>
</vxml>
```

Code Snippet 35 Capture the Participant's Emotional Self-Report in VoiceXML

The grammar *emotions.grxml* that captures the participant's emotional choice looks like this:

```
<?xml version="1.0" encoding="utf-8"?>
<grammar xmlns="http://www.w3.org/2001/06/grammar" version="1.0"
  xml:base="file:/C:/workspace2/TedBerry/grm/emotions.abnf" mode="voice"
  tag-format="" xml:lang="en-US" root="data">
    <rule id="data" scope="public">
    <item>
    <item repeat="0-1">
    <ruleref uri="#filler" /></item>
    <ruleref uri="#emotion" />
    <item repeat="0-1">
    <ruleref uri="#filler2" /></item>
    <tag>out= rules.emotion;</tag></item>
    </rule>
    <rule id="filler" scope="private"> <one-of>
    <item> I feel</item>
    .....et cetera...
    </one-of> </rule>
    <rule id="emotion" scope="private"> <one-of>
    <item> okay <tag>out="1";</tag></item>
    <item> happy <tag>out="2";</tag></item>
    <item> great <tag>out="2";</tag></item>
    <item> sad <tag>out="3";</tag></item>
    .....et cetera...
    </one-of> </rule>
    <rule id="filler2" scope="private">
    <one-of> <item> today</item>
    .....et cetera...
    </one-of> </rule>
  </grammar>
```

Code Snippet 36 Emotional Self-Report Grammar in GrXML

Alternatively, a DTMF grammar or a combination of both DTMF and speech recognition could be used. Server-side scripts generate VoiceXML script personalization. For example the PHP script:

```php
<?php
$dom = new DOMDocument("1.0");
$user = "Ted Hill";
$block= $dom->createElement('block');
$dom->appendChild($block);
$prompt = $dom->createElement('prompt', "Hello $user");
$block->appendChild($prompt);
echo $dom->saveXML();
?>
```

Code Snippet 37 Dynamic VoiceXML in PHP

code snippet 94 generates:

```xml
<?xml version="1.0" ?>
<block>
  <prompt>Hello Ted Hill</prompt>
</block>
```

Code Snippet 38 Dynamic VoiceXML Output from PHP

Use-Case Views

Patient View

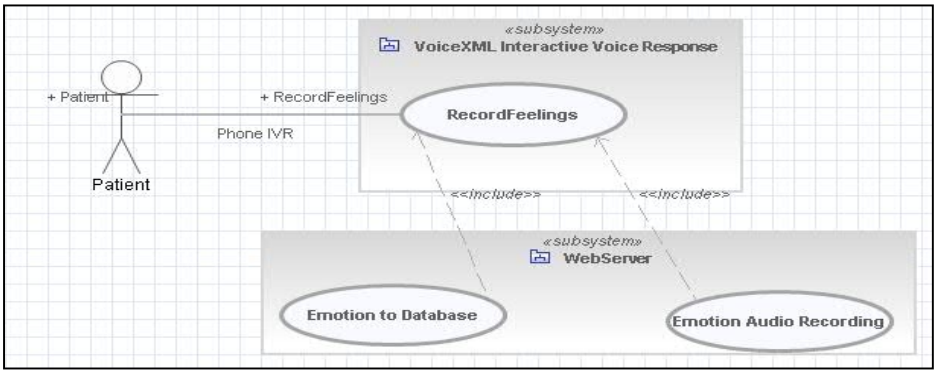


Figure 127 Record Emotional Momentary Experience using IVR

Figure 127 depicts the EMA, or ESM, key to this entire project. A patient dials a 1-800 number, enters their assigned PIN, and records how they are feeling. Once the call is complete, a record of the call is stored in the database, and the recording, indexed by the database record, is stored in the file system.

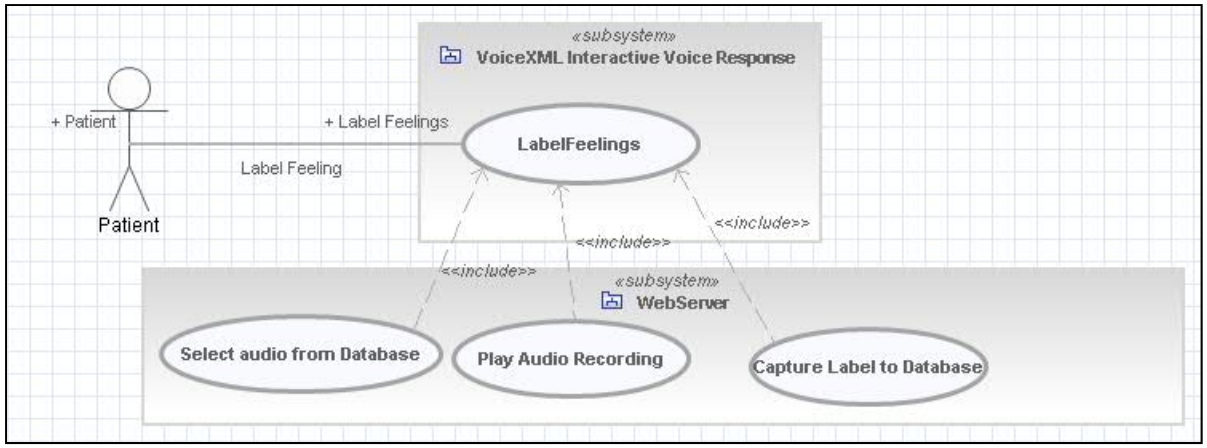


Figure 128 Anonymous Labeling of Emotional Recordings using IVR

Figure 128 depicts the patient "relating" to other people's recorded emotions. A patient dials a 1-800 number, enters their assigned PIN. An anonymous recording is played back to the user and the patient is asked to label the emotion of the recording using word-list speech recognition. Once the call is complete, a record of the call and the labeled emotion is stored in the database.

Professional View

Figure 129 depicts Professional Caregivers (Addiction treatment specialists, Psychologists, Mental Health Clinicians, Doctors, etc.) access to patient monitoring and trend analysis tools. Caregivers can login to the web portal and monitor their assigned patients' progress (listen to their emotional recordings, view history of "check-ins", etc.), and analyze trends (positive/negative emotional trends, ability to relate to others' emotions, etc.).

An Alarm Subsystem can be configured to trigger an email or an outbound telephone call should a patient not check-in for a certain period of time, be too emotionally negative, etc.

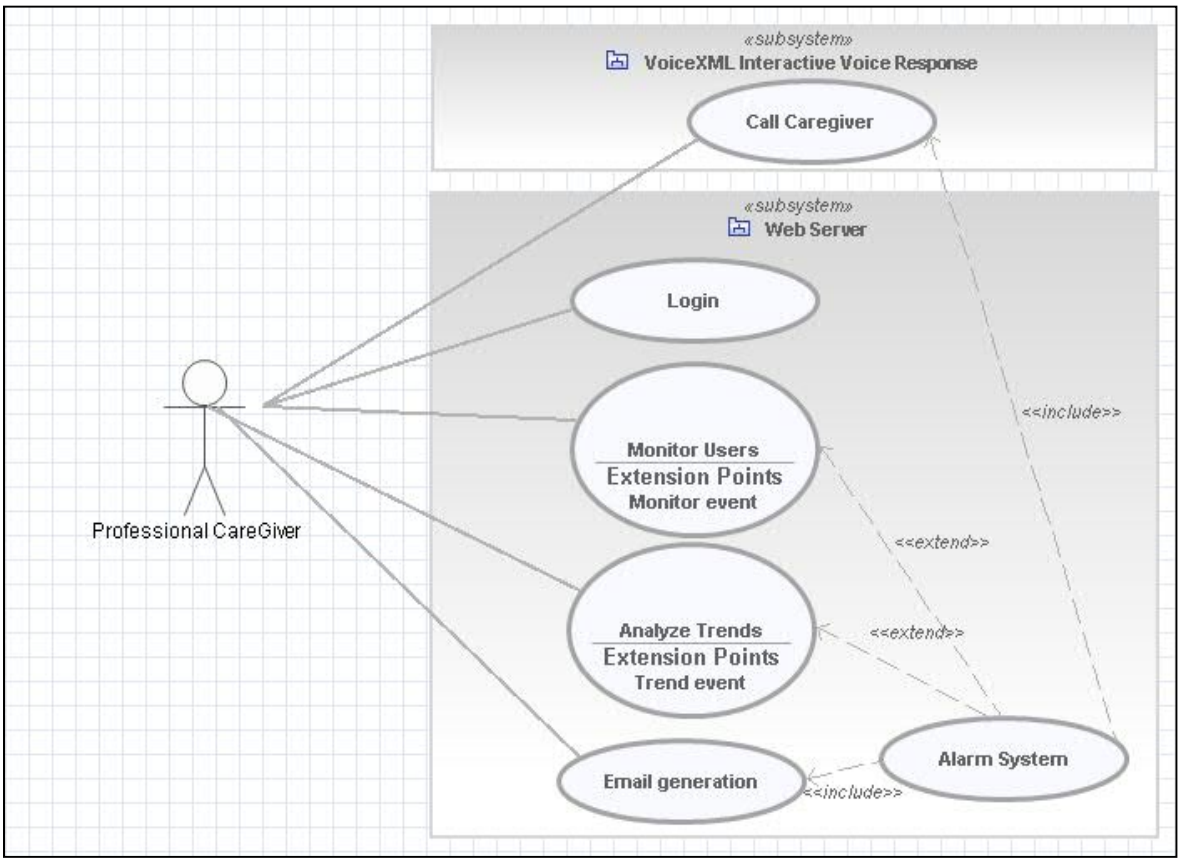


Figure 129 Patient Monitor and Trend Analysis

Speech Scientist View

Figure 130 depicts the training of the emotion detector. The Emotion Detector Server is separated for the Web Server for cost reduction and CPU bandwidth conservation purposes. Recorded audio is transferred to the Emotion Detection Server using the Secure File Transfer Protocol (SFTP) protocol. The Audio is then transformed into a format compatible with the front-end processor. The Front-end processor converts the audio into silence-removed MFCCs. The audio is sorted into directories corresponding to the emotion contained in the audio. The Emotion Detector can then be trained on the classified data.

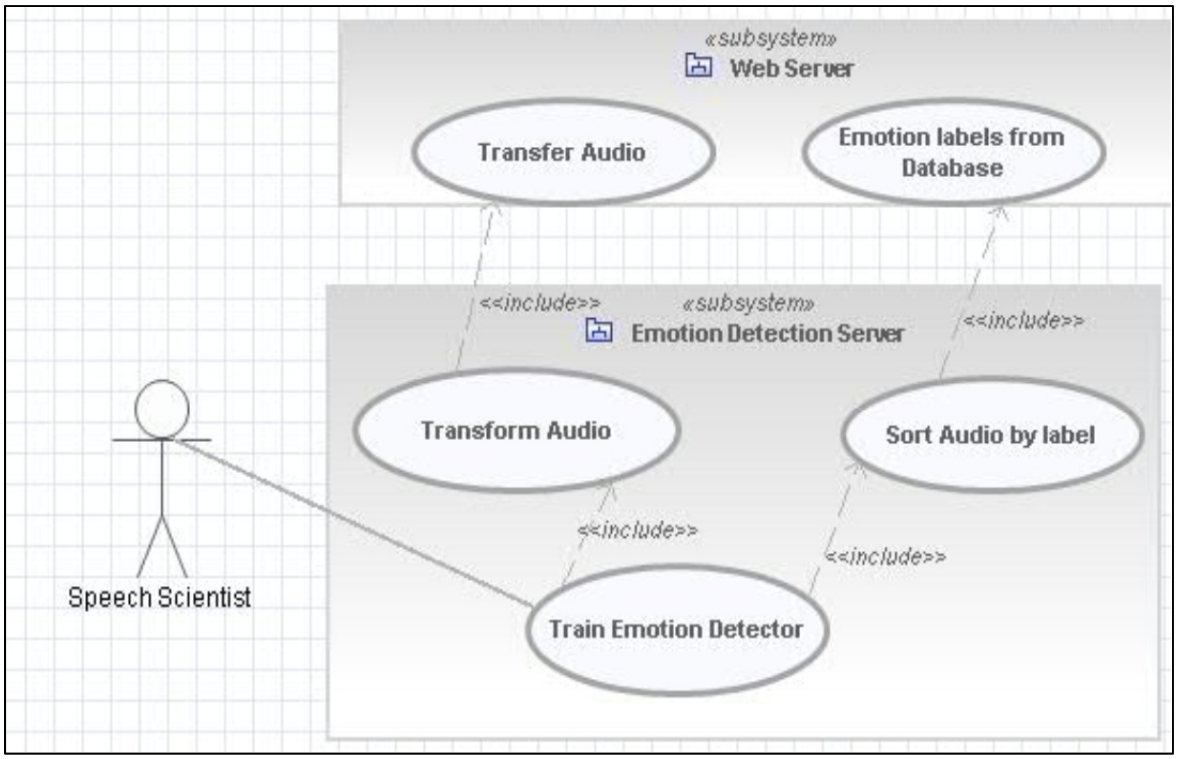


Figure 130 Speech Scientist Training the Emotion Detector

Key Scenarios

IVR and Dynamic VoiceXML

In the scenario of Figure 174, a patient calls into the IVR system. A personalized VoiceXML script is dynamically generated. The Patient interacts with the IVR system by expressing their feelings resulting in an audio recording, a self-label of their emotional state, and a record of the call session in the database.

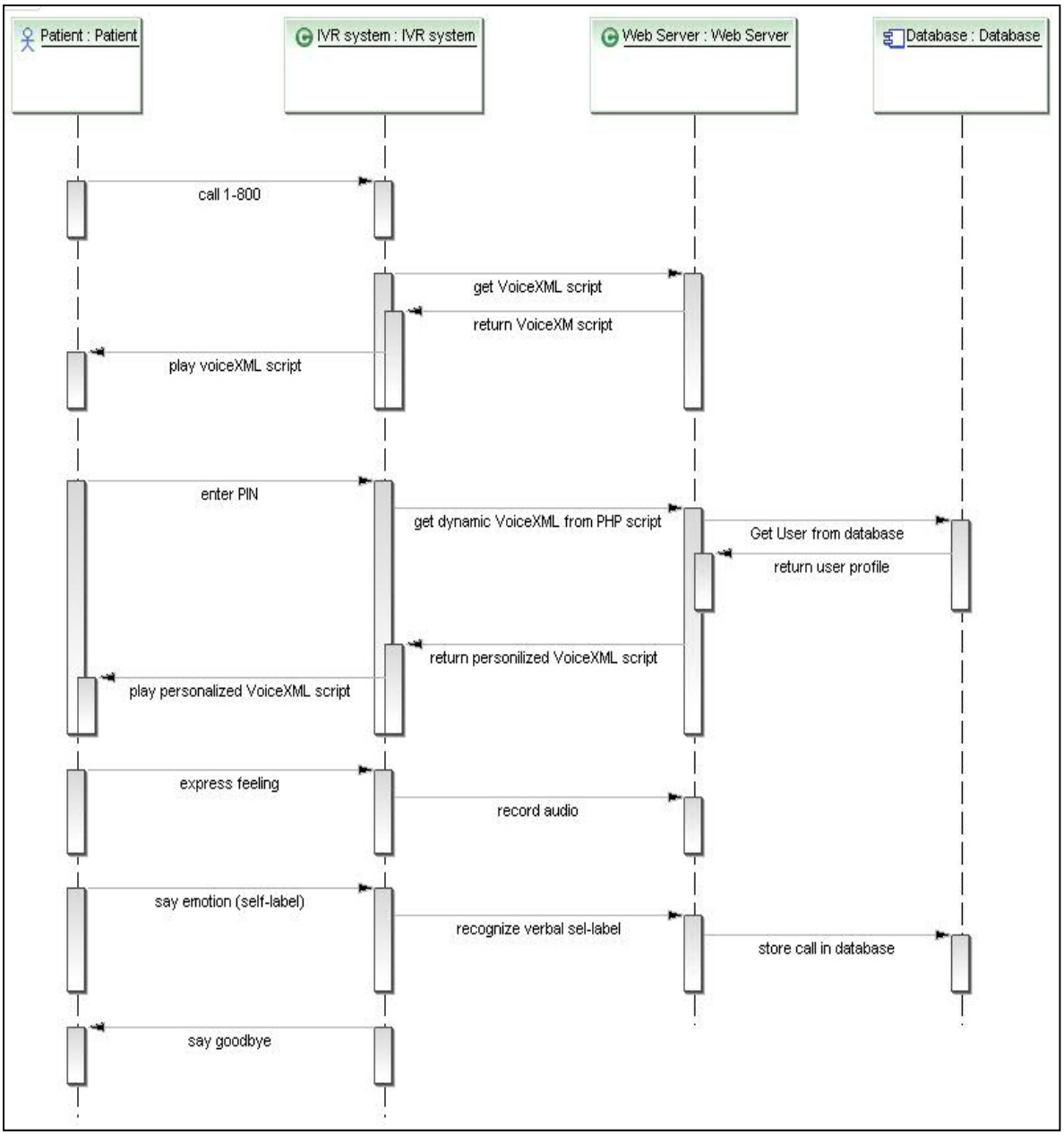


Figure 131 IVR Sequence Diagram

Emotion Detection Model Training

The scenario of Figure 175 depicts the sequence of events needed to train emotional models. A CRON daemon periodically looks for new emotional feeling audio recordings on the web server. The corresponding labels are extracted from the database, and a voting system is used to determine the highest probably emotional label. The emotion labels are used to sort the corresponding audio that is then used to train the emotion models.

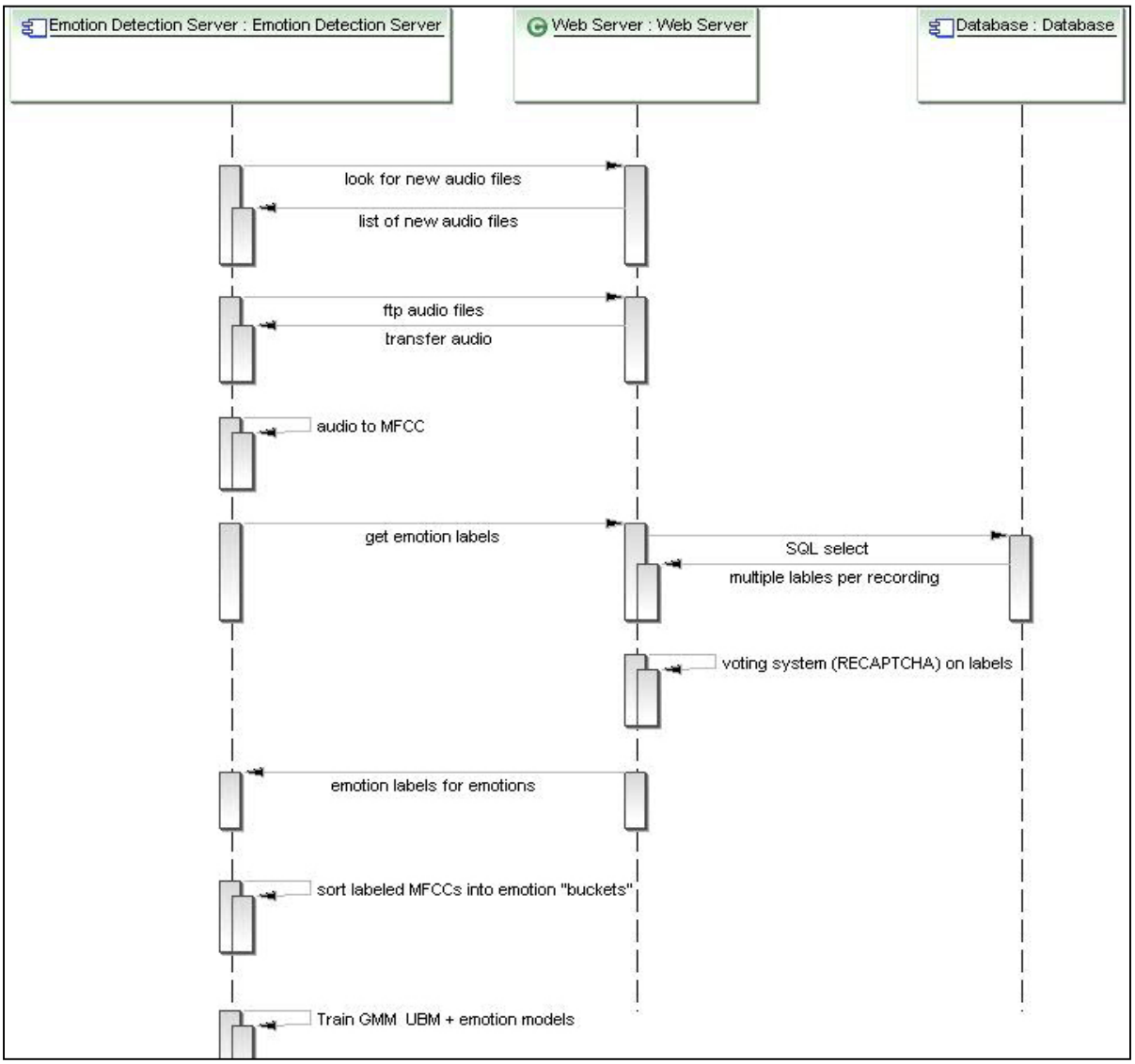


Figure 132 Emotion Detector Training Sequence Diagram

Tooling and Infrastructure

www.emotiondetect.com

The www.EmotionDetect.com Software Architecture has been designed to be flexible, configurable, and easily customizable. There are multiple User Interface modalities supported: (1) Voice Interaction over a telephone; and (2) personalized dynamic Web 2.0 content and user access control with multi-modal interaction including Flash audio playback, and HTML/JavaScript/AJAX web pages

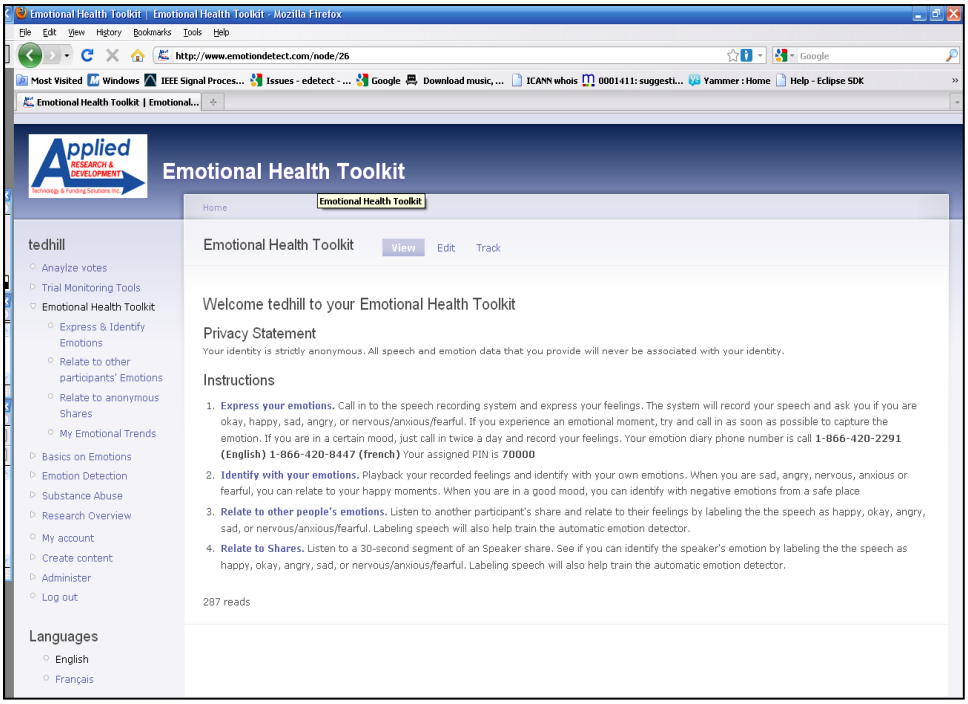


Figure 133 www.emotiondetect.com

Versatility, multi-modality, and rapid evolution are vital to support an iterative approach to the User Interface. The Conceptual models of both the Patient/Addict and the Professional Caregiver will rapidly evolve as more and more insight on their Mental Models becomes available through consultation and usability feedback.

The Drupal Open-Source Content Management Framework supports both (1) a Patient view; allowing them to review their expressed feelings, and relate to others' feelings; and (2) a Professional Caregiver view; to monitor patients and analyze emotional and behavioral trends.

The Emotion detection architecture supports the speech science of audio formatting, conversion, and feature extraction; and the emotion detection algorithm model training and detection.

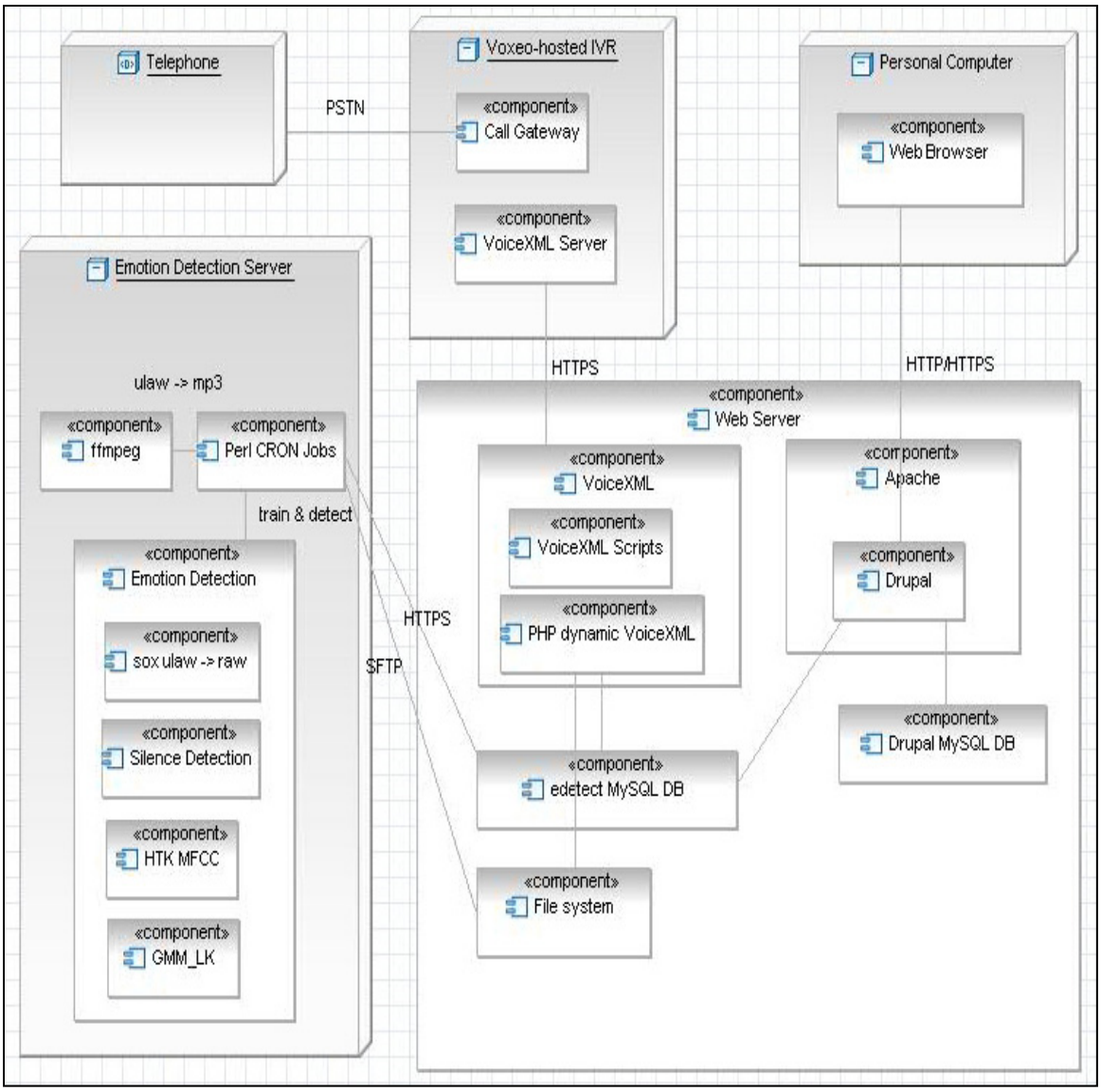


Figure 134 www.emotiondetect.com Deployment Architecture

Emotion Detection Software Architecture

The Emotion Detection Software Architecture consists of myriad of open-source and proprietary audio and speech processing algorithms that will be advanced during the course of this project. This architecture has been improved upon the initial setup.[18]

[18] courtesy of Dr. Najim Dehak, a former student under Dr. Dumouchel

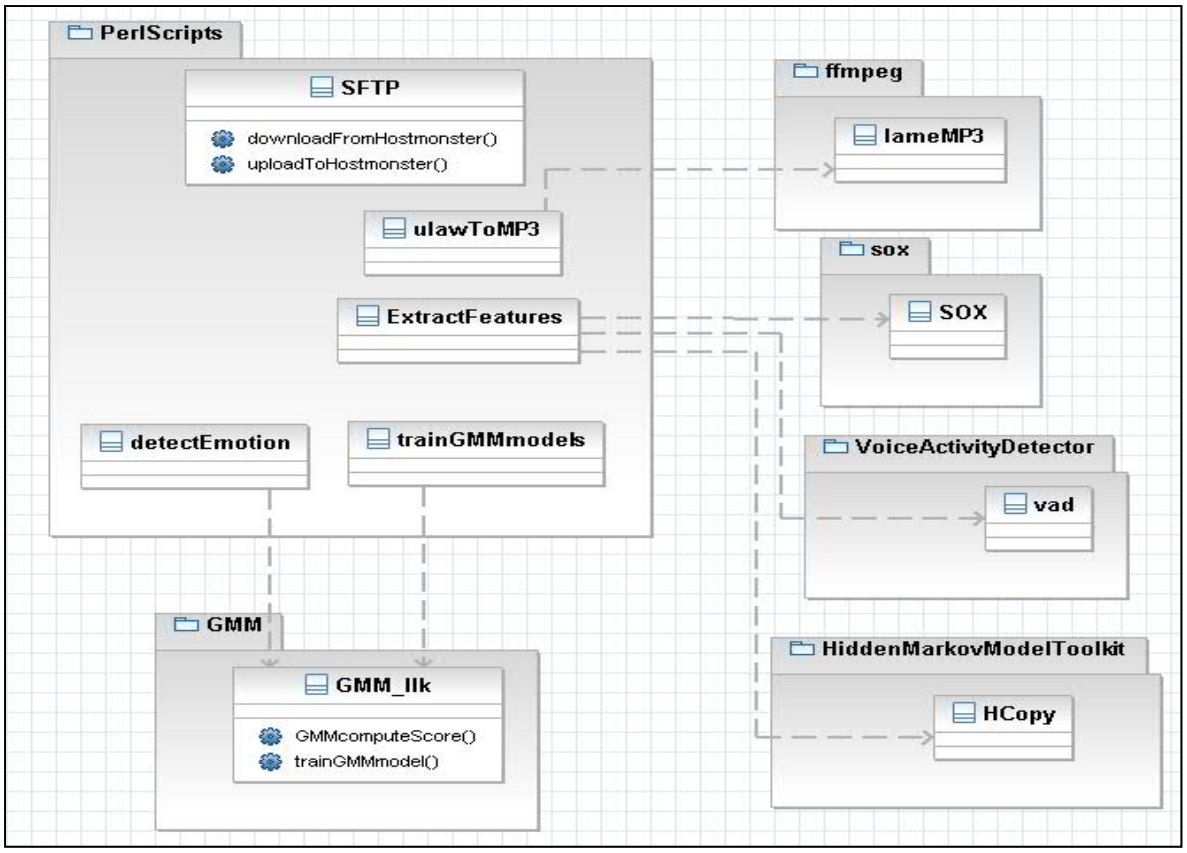


Figure 135 Emotion Detection Deployment Architecture

The SFTP Perl script is used as the glue between the Web server and the Emotion Detection Server for uploading and downloading audio data and emotion detection information.

- **Ffmpeg:** http://ffmpeg.org licensed under GNU (GNU's Not Unix) Lesser General Public License (LPGL): cross-platform solution to record, convert and stream audio and video. It includes libavcodec - the leading audio/video codec library. Includes the lame Moving Picture Experts Group (MPEG) Audio Layer 3 (MP3) encoder required to generate MP3 for the embedded Flash audio player. MP3 encoding is a licensing issue which must be addressed prior to commercialization.

- **Sox:** http://sox.sourceforge.net/ command line utility that converts various formats of computer audio files in to other formats. Required to convert from ulaw to HTK raw audio format.

- **Voice Activity Detector (VAD):** Originally from Institute for Signal & Information Processing, Mississippi State University, licensed under LPGL. VAD has since been upgraded by Centre de recherche informatique de Montréal (CRIM), also under LPGL. VAD is used to detect and remove silence and unvoiced audio.

- **HTK Hidden Markov Model Toolkit:** http://htk.eng.cam.ac.uk/ The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models. The tools provide sophisticated facilities for speech analysis, HMM training, testing and results analysis. The software supports HMMs using both continuous density mixture Gaussians and discrete distributions and can be used to build complex HMM systems. The tool HCopy is used to generate Mel Frequency Cepstral Coefficients (MFCCs) from the raw formatted audio.

- **GMM_llk:** http://www.tsi.enst.fr/~chollet/becars/index.php Speaker Verification Library and Tools for Speaker Verification licensed under LPGL. Open-source software for speaker recognition (BECARS) was developed by the University of Balamand (Lebanon) and the École Nationale Supérieure des Télécommunications (GET-ENST Paris, France), that has successfully taken part in successive NIST evaluations. It provides a C library and several tools that permit to set up of the modeling and scoring phases of a GMM-based Automatic Speaker Verification system. It firstly provides an implementation of the EM algorithm with different kinds of criteria, e.g. ML, MAP and MLLR. It permits then to estimate the likelihood of a set of acoustic vectors given a model. BECARS has been extensively cross-tested by many academic institutions.

Data Schema for Collection and Emotion Annotation

The data schema in Figure 136 is a subset of the www.emotiondetect.com database to collect user speech recordings into the Software Query Language (SQL) Table “Feelings”, user annotations of their own emotions, and user annotations of other people’s emotions (SQL Table “userFeelingLabels”).

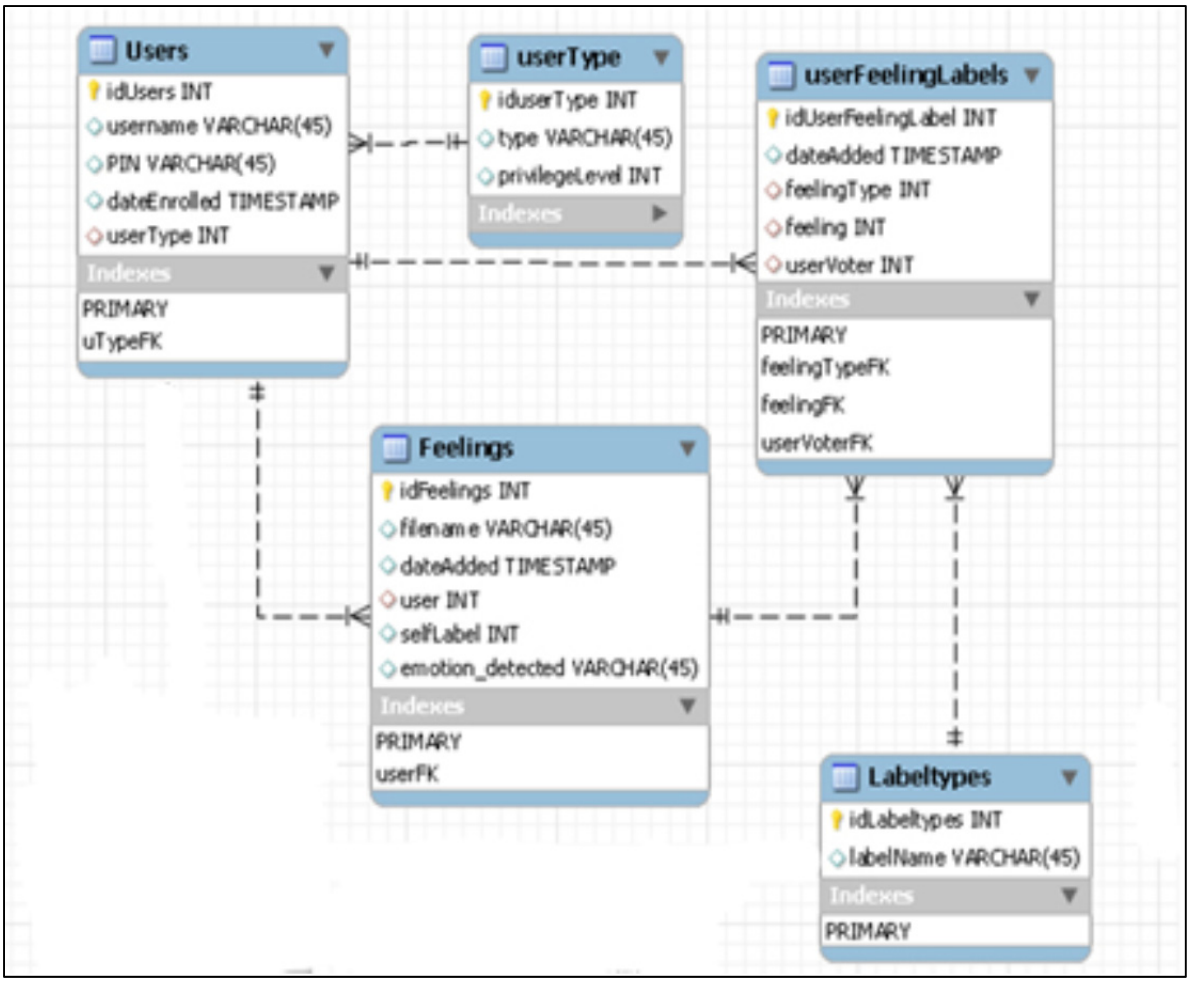


Figure 136 Simplified Data Schema

APPENDIX C

USER INTERFACE DESIGN

Figure 137 Emotional Health Toolkit Login Web Page

The emotional health toolkit is customized to a therapist, researcher, Doctor, etc. (hereafter "supervisor") requirements. In Figure 179 the site has been customized for Dr. Moehs.

Audio dialogue prompts are recorded by the supervisor or delegate to promote familiarity for the patients in the system thus increasing their willingness and openness to use the system.

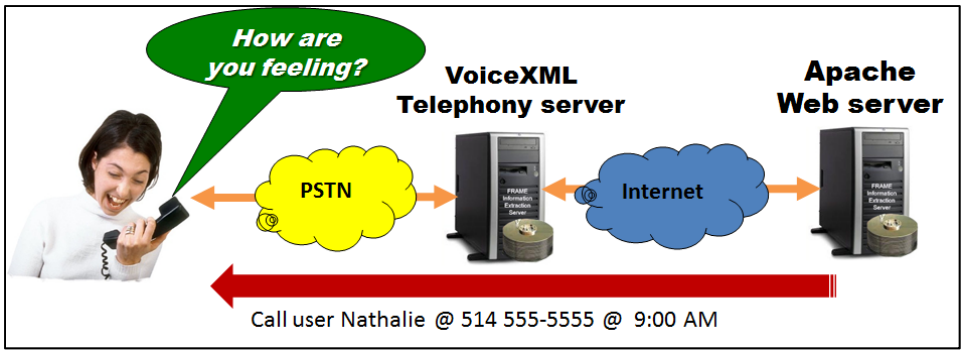


Figure 138 Audio Prompt Customization

There are three different user types (roles). Each role has a set of functionalities and views:

1. Supervisor view: manage patients' profiles and call times. Analyze patients' data
2. Participant view: view personal emotion health indicators.
3. Transcriber view: Transcribe audio data (crowd-sourced emotional truth)

Supervisor view

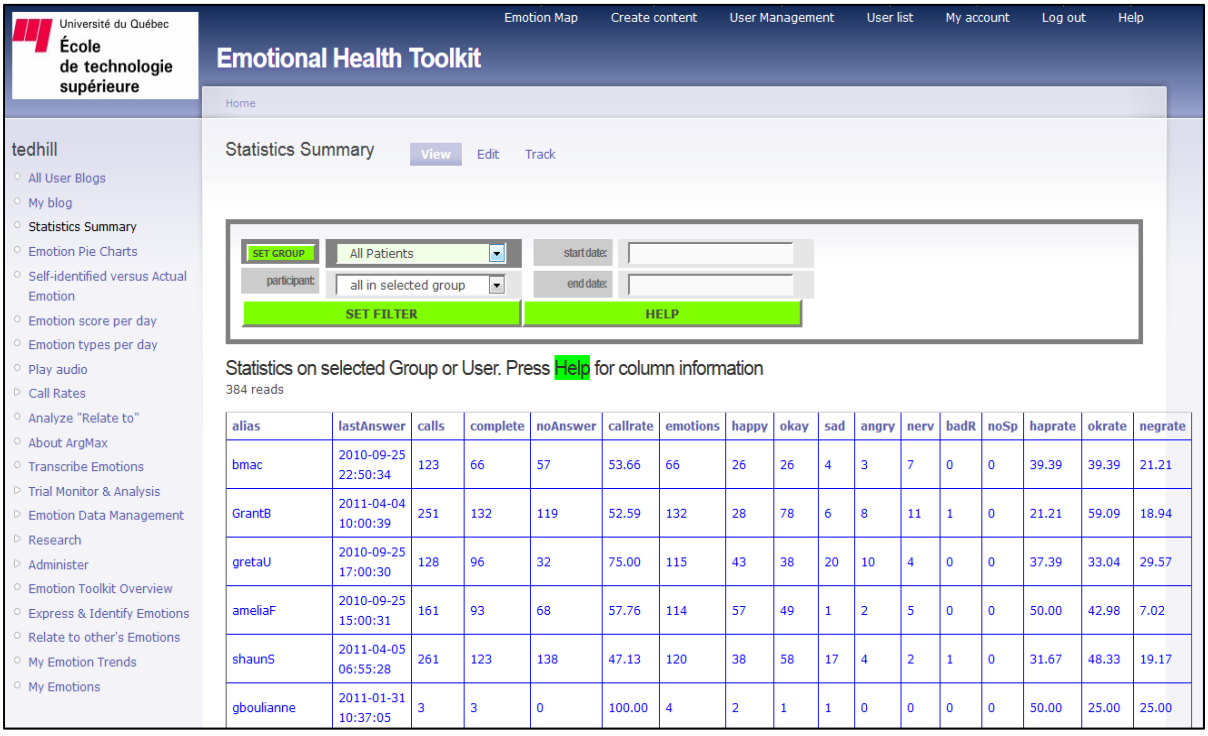


Figure 139 Supervisor Home Page

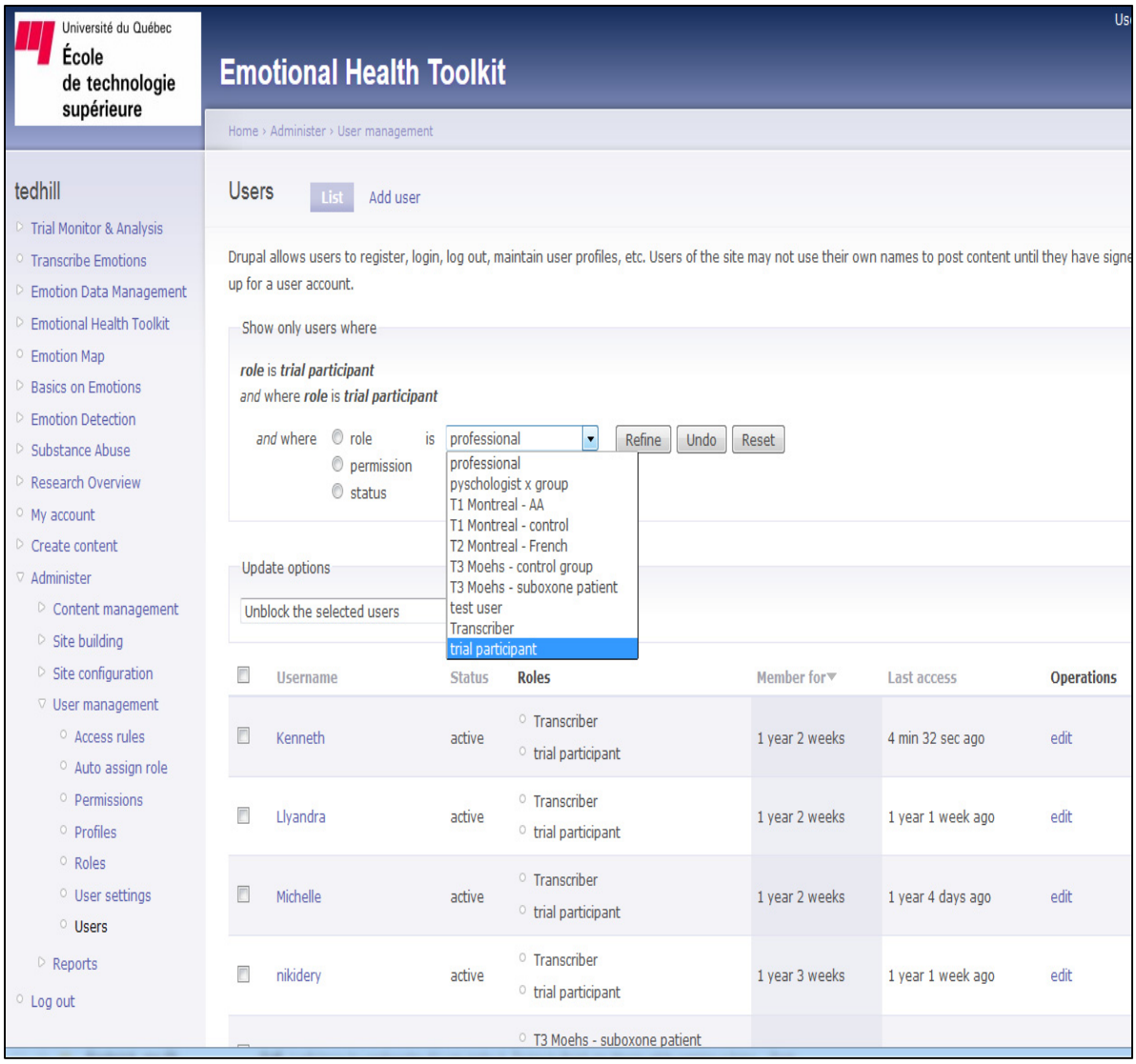


Figure 140 List of Users

The User view allows users to be added, edited or deleted.

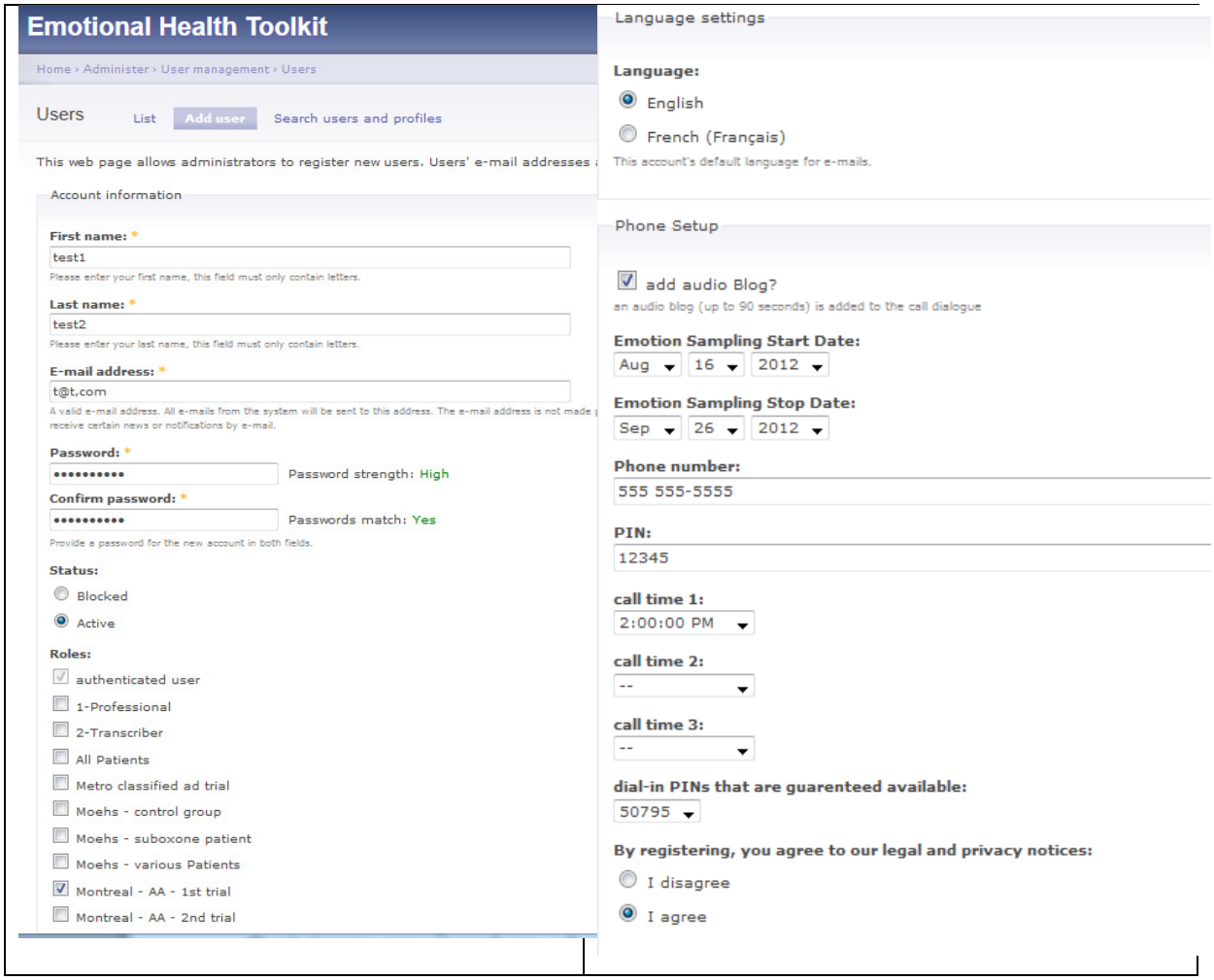


Figure 183 Configure a New User

To configuring a new user the name (or alias) is entered along with email, password, role, and language. The Phone setup activates the automatic call. The phone number is dialed at up to 3 call times. A PIN can be manually entered or assigned to allow the participant to call in and enter the PIN on their telephone keypad when prompted.

Once the toolkit has been configured with all patients, the supervisor can start monitoring and analyzing results.

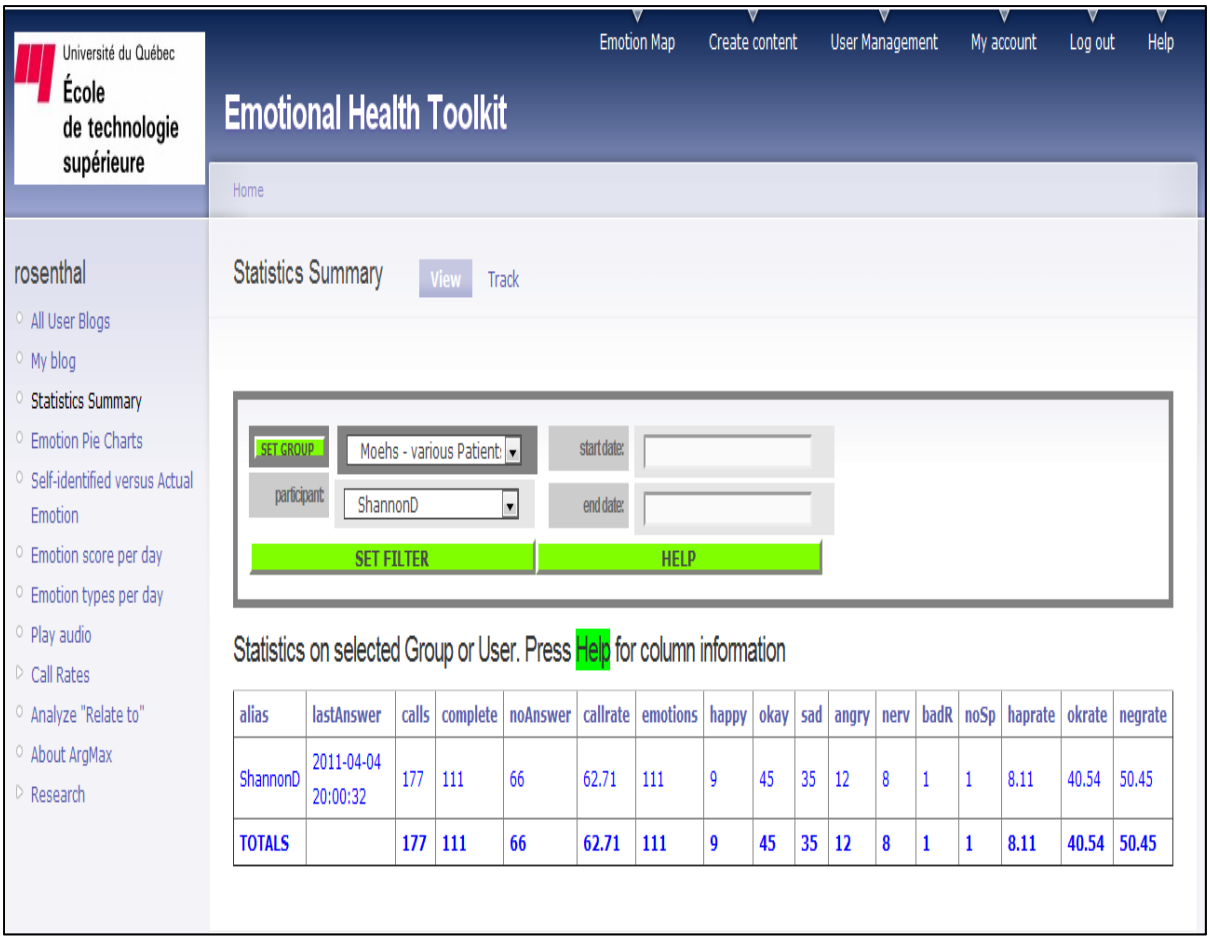


Figure 141 Filter by Group, Patient, Start Date and End Date

Filters can be applied to any supervisor view. A supervisor only sees groups and users assigned to him thus allowing multiple supervisors to use the same emotional health toolkit server but maintain privacy of their own data.

Figure 141 is a filter applied to the home page view of Figure 139.

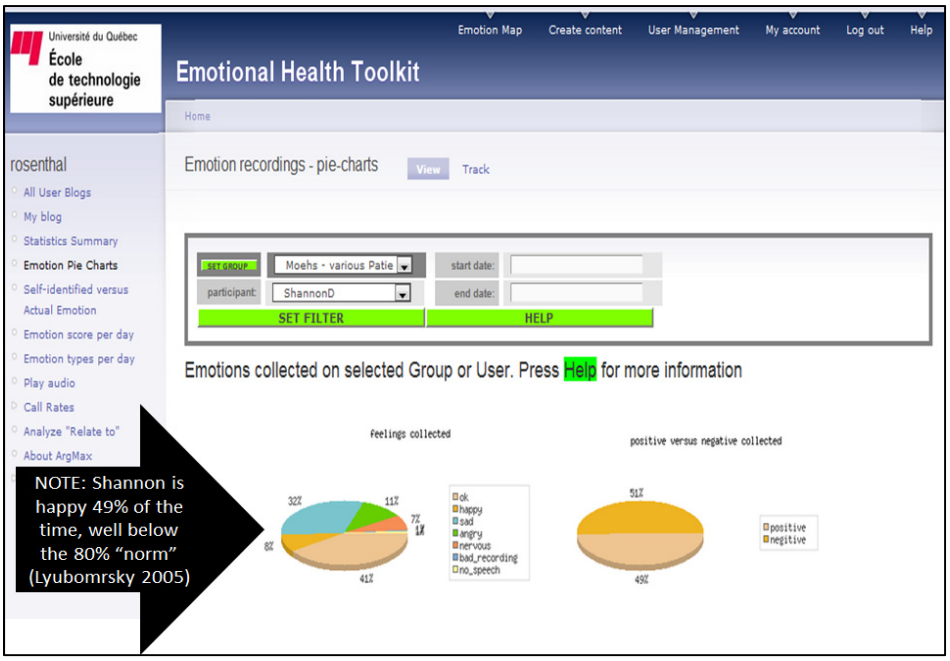


Figure 142 Emotional Truth Pie Charts

Pie charts of emotional truth provide a quick overview of the patient's emotional health. Histograms compare the individual or group's self-assessment to their emotional truth.

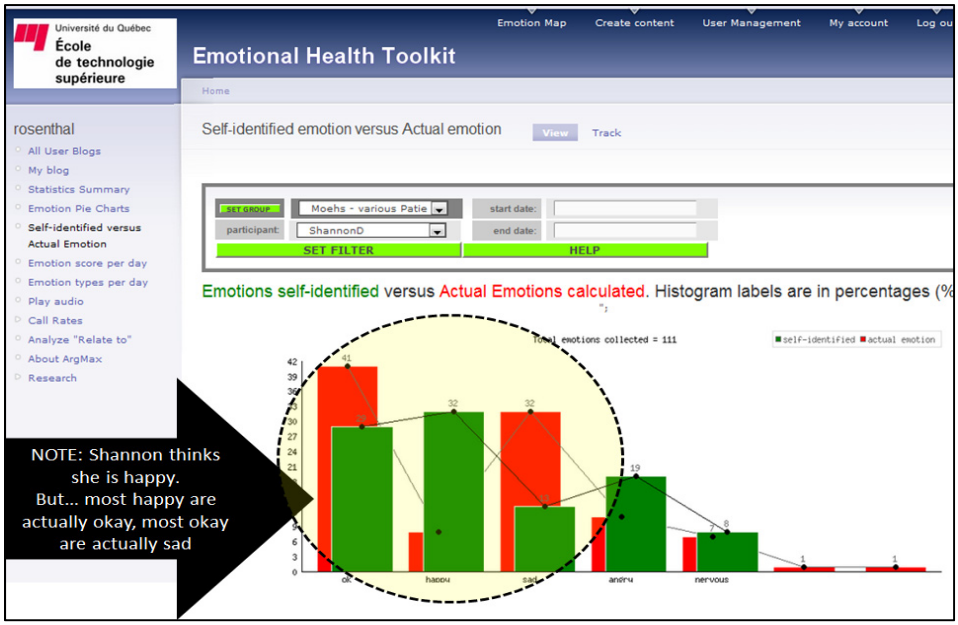


Figure 143 Self-Assessment versus Emotional Truth

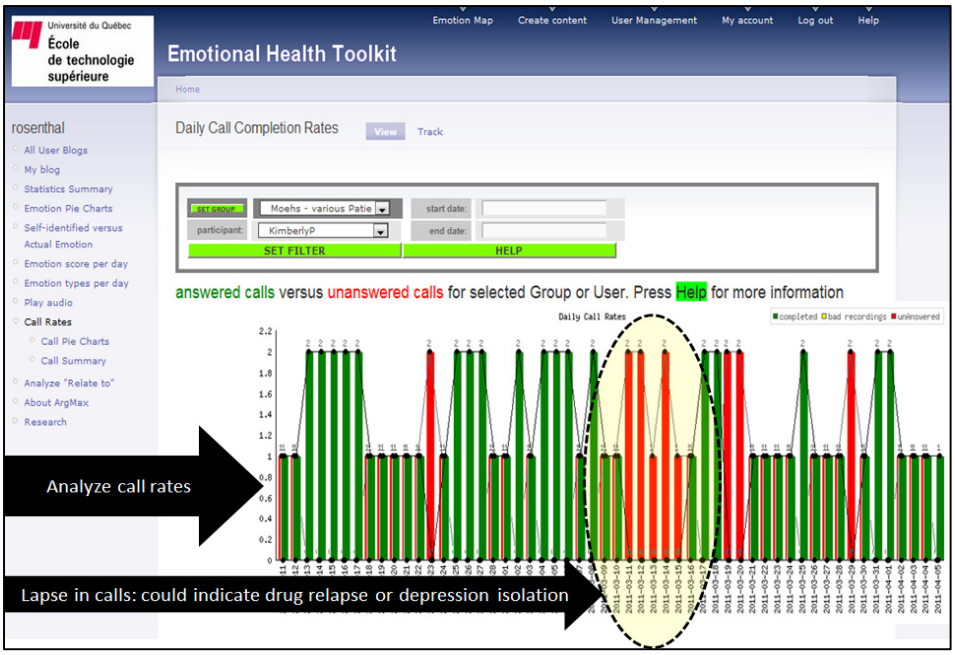


Figure 144 Analyze Call Rates

An individual or group's Call rates can be analyzed. Green indicates a successful call. Red indicates no answer or hang-up. Pie charts are also available.

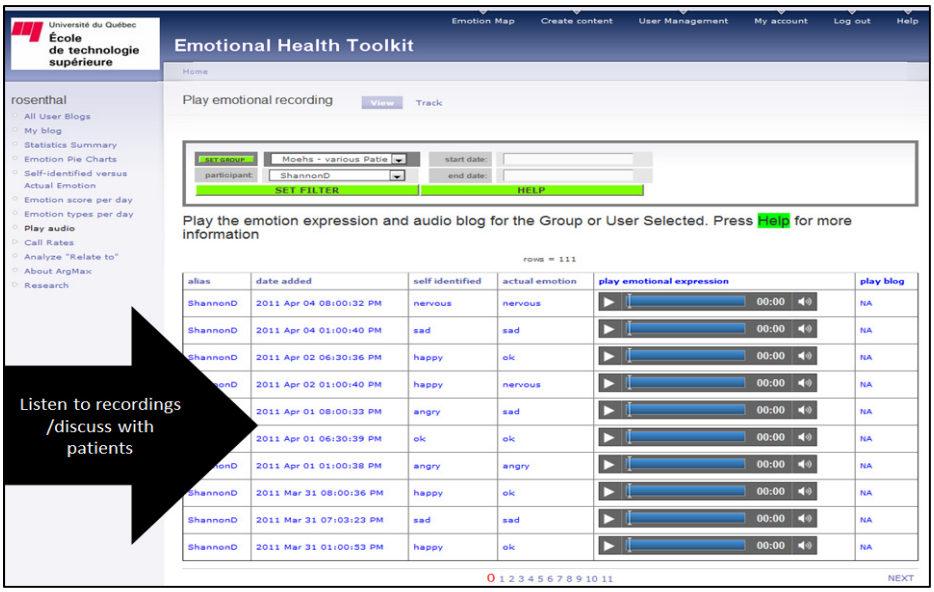


Figure 145 Listen to Audio Data Collected

Collected data can be listened to by the supervisor and discussed with the patient (i.e. in a cognitive behaviour therapy session).

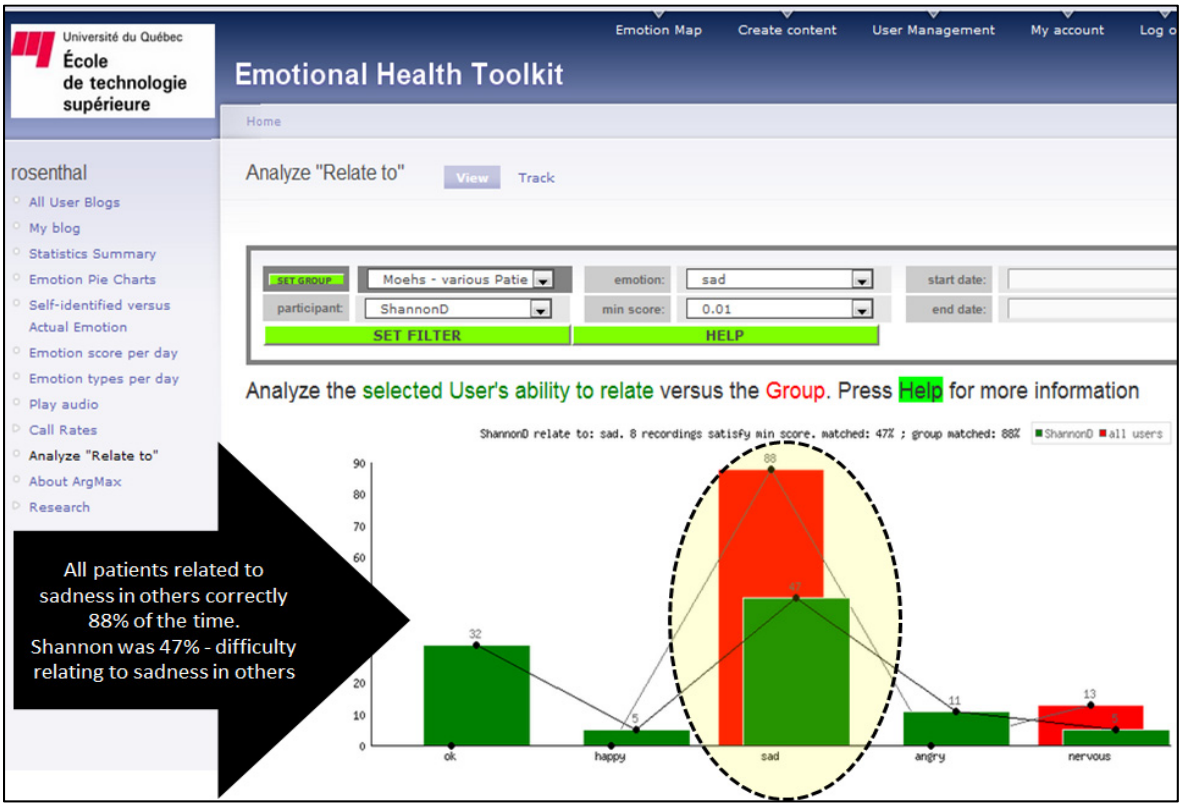


Figure 189 Patient's Empathy (ability to relate)

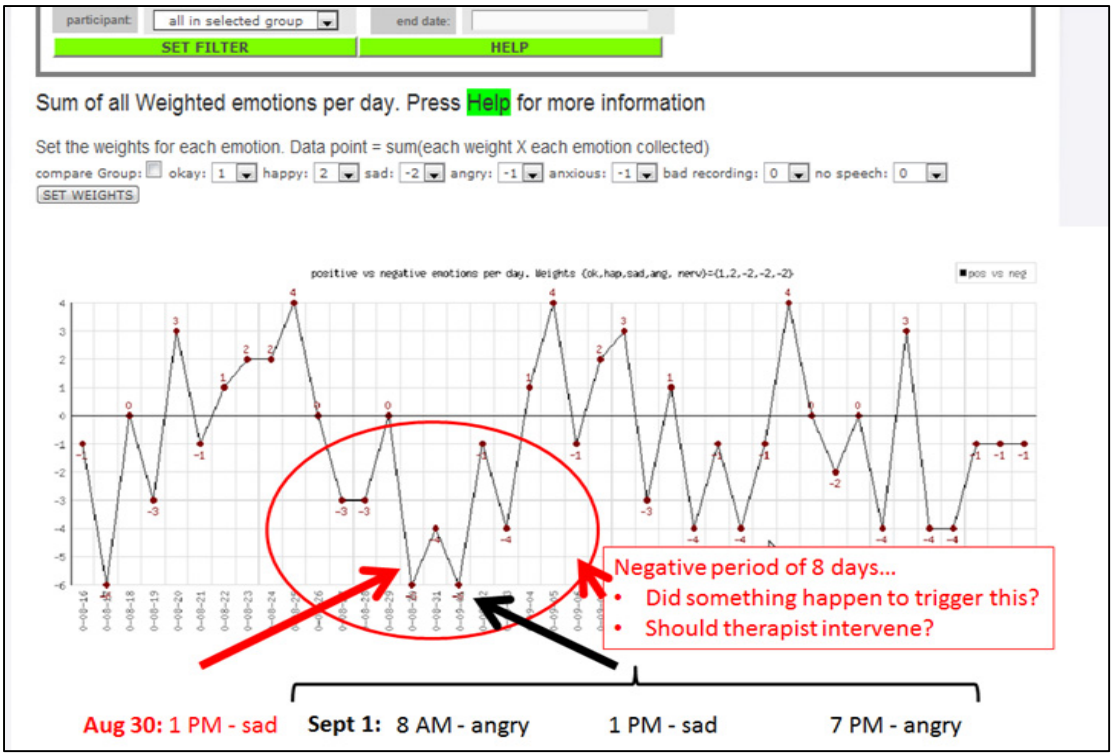


Figure 146 Graph of Emotions over Time

The patient's emotions can be graphed over time to visualize problems. The supervisor can then play back the audio and discuss behaviours, thoughts, and emotions within this period.

Participant view

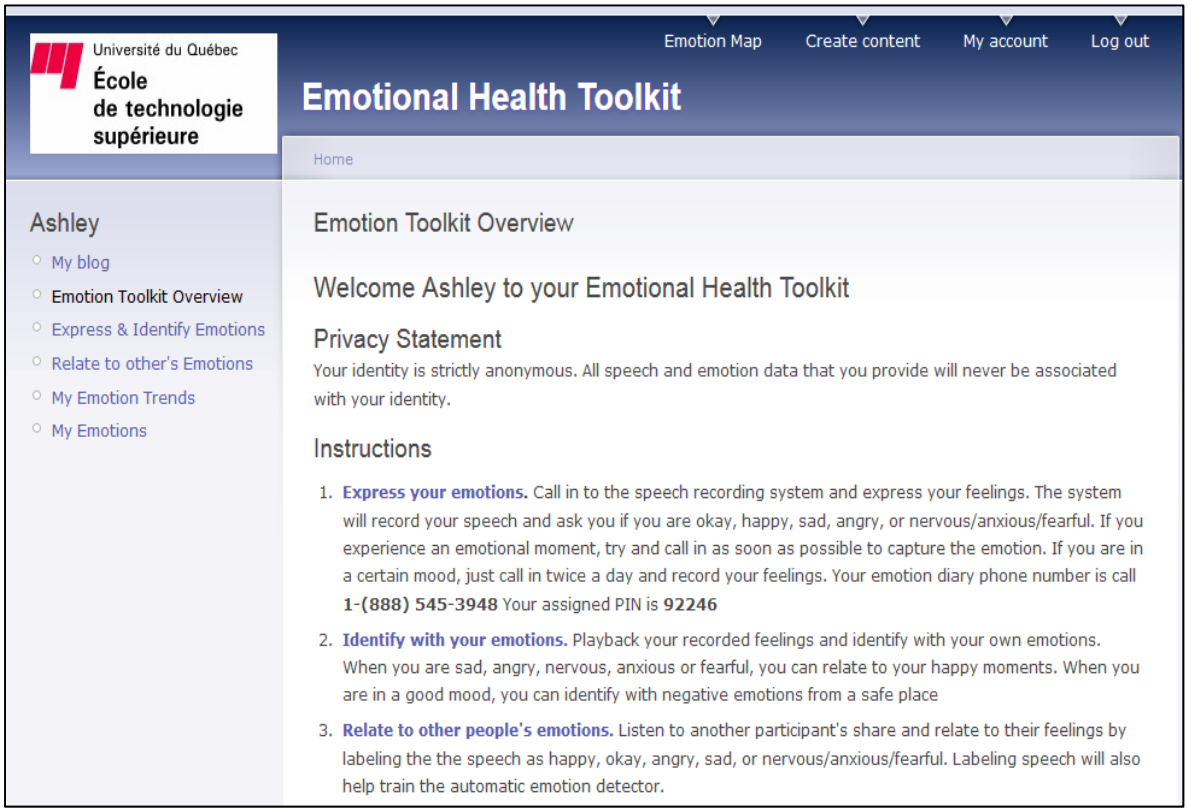


Figure 147 Participant Home Page

A participant logs in, and views their emotional health. Views are similar to the supervisor kit, but restricted to their emotions only.

Transcriber view

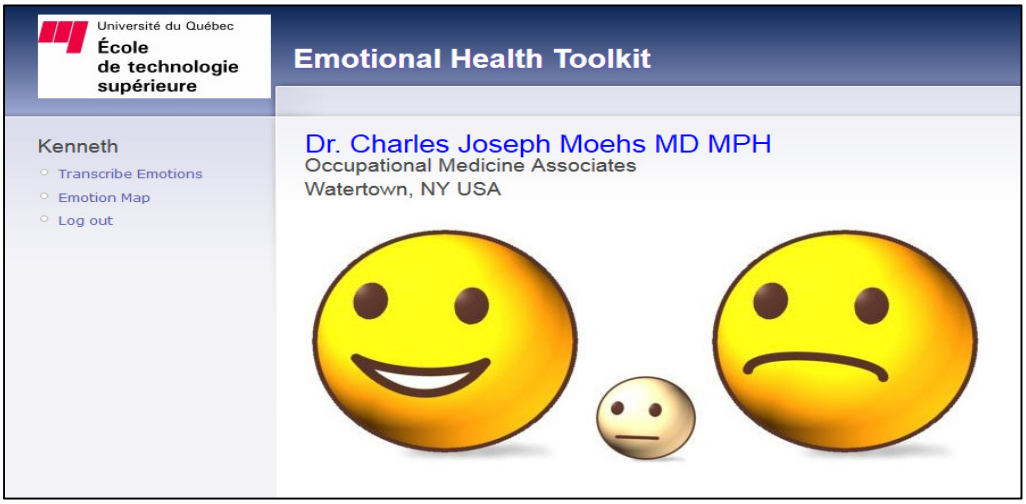


Figure 148 Transcriber Home Page

A transcriber logs in to transcribe emotions.

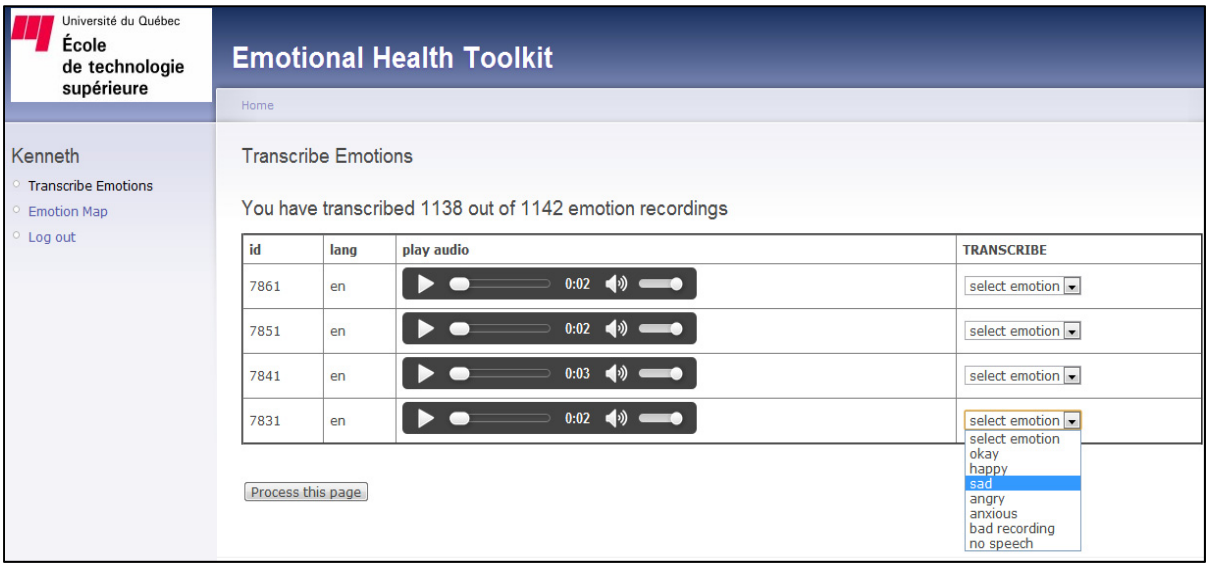


Figure 149 Transcriber Interface

Transcribers are only presented audio to transcribe. The transcriber listens the emotion, and transcribes the emotion.

APPENDIX D

HAPPINESS REGRESSION ANALYSIS

Happiness (happyCROWD)

Table 34 provides frequency counts of happiness occurrence across the collected data. There is a possibility that Opioid addicts are proportionally less happy than the GP and AA members.

Table 61 Happiness Frequencies

	FALSE		TRUE	
General Pop	1728	72.0%	671	28.0%
AA Member	3469	77.8%	990	22.2%
Opioid-Suboxone	866	86.4%	136	13.6%

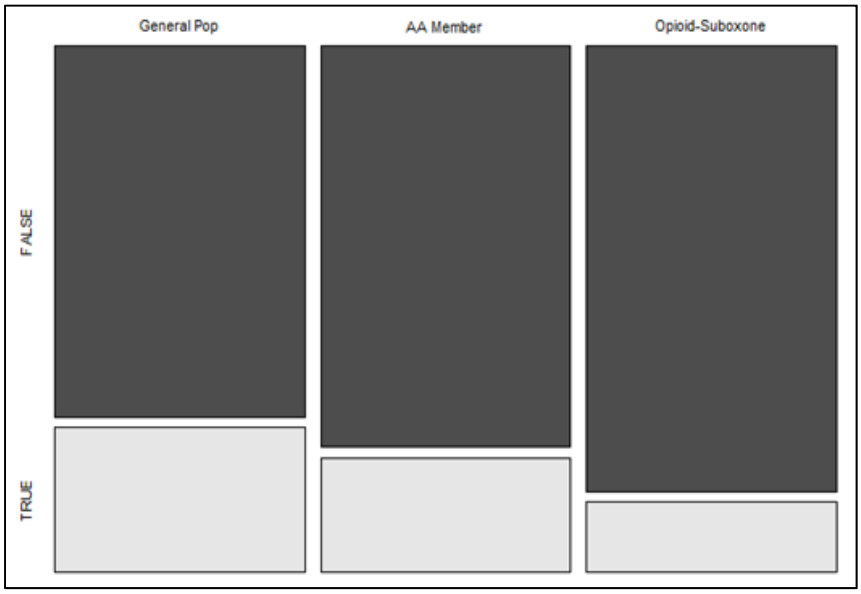


Figure 150 Frequency of Happiness (happyCROWD) across Groups

The procedure for calculating the statistical significance for a discrete-choice outcome variable was described in in section 1.7.20 on page 75 and used Happiness emotional truth as an example. As such, only the results will be repeated in this section.

The probability of a SUBX patient being happy is 9.5% less than the GP ($p < 0.05$). Code Snippet 19 reveals a significant difference of 8.8% between AA members and SUBX ($p < 0.05$).

- GP pr(happyCROWD) = 24.7% (95% CI, 19.2%–31.0%)
- AA Member pr(happyCROWD) = 24.0% (95% CI, 16.4%–33.7%)
- SUBX pr(happyCROWD) = 15.2% (95% CI, 9.7%–22.9%)

```
   Formula: emotion ~ gender + (1 | p)
Data: EMO
  AIC  BIC logLik deviance
 8020 8047  -4006     8012
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.86907  0.93224
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5271     0.4599  -3.321 0.000898 ***
genderF       0.1987     0.4829   0.411 0.680736
genderM       0.2810     0.4778   0.588 0.556510
```

Code Snippet 39 Happiness Emotional Truth versus Gender Two-Level Model in R

There is no significant difference for gender or language.

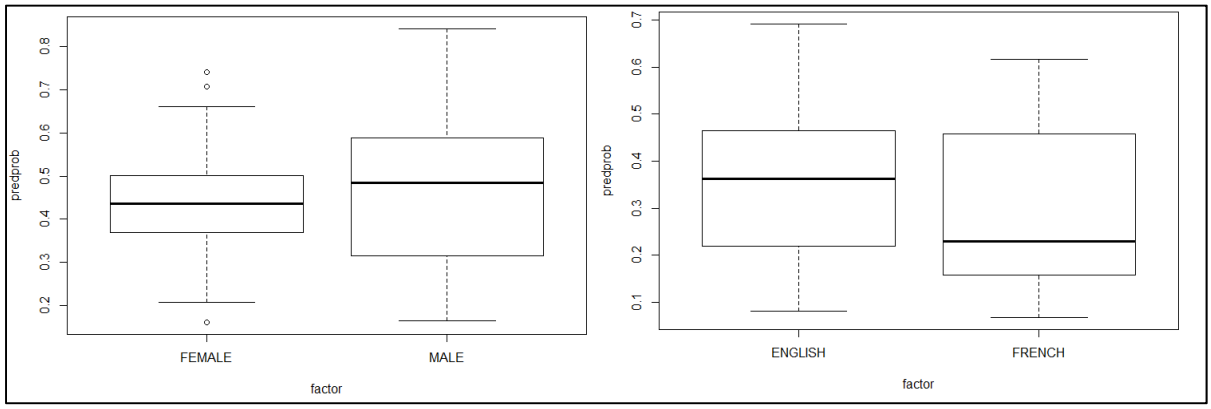


Figure 151 Predicted Probabilities of Happy versus Gender and Language

Happiness Self-report (happySELF)

There are no significant statistical differences of ***happySELF*** across group3, gender, or language. Frequencies of happySELF across groups are very similar.

Table 62 Happiness Self-report Frequencies across Groups

	FALSE	TRUE
GP	68%	34%
AA	72%	30%
SUBX	67%	35%

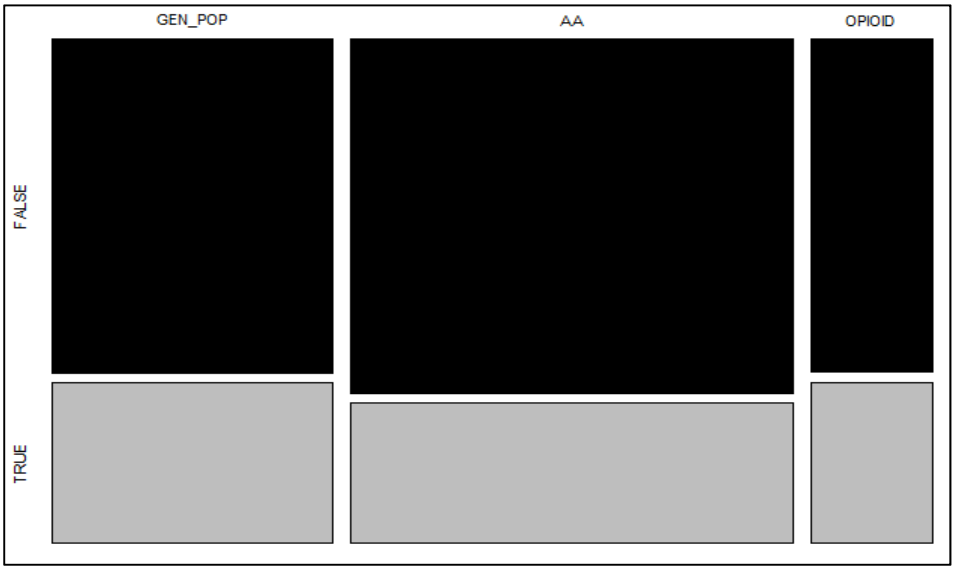


Figure 152 Frequency of Happiness Self-Report across Groups

```
Formula: happySELF ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 9717 9731  -4856     9713
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.81592  0.90328
Number of obs: 8376, groups: p, 130

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.80014    0.09008  -8.882   <2e-16 ***
```

Code Snippet 40 Happiness Self-Report Null Model in R

The log-odds mean for self-assessment for an average participant ($\mu_{0j} = 0$) is significant is estimated at -0.80014 which is a probability of 31.0%. There are no significant differences across group3, gender, or language.

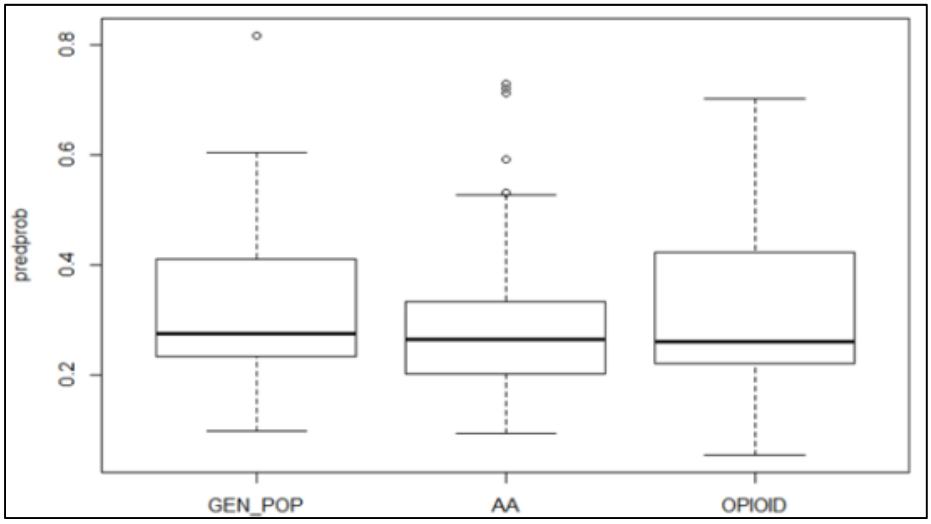


Figure 153 Predicted Probabilities of happySELF versus Group

Happiness Self-Awareness (happySELFAWARE)

From equation (6.3, there is an 80.7% probability (95% confidence interval (CI), 77%–84%) that a General Population participant is self-aware of their happiness. There is a trend ($p < 0.1$) that the probability of an Opioid-Suboxone patient being happy self-aware is 75.3% (95% CI, 72%–83%); 5.4% less than the General Population. There is no significant difference between AA members and the General Population.

HappySELFAWARE is a derived outcome variable from ***happySELF == happyCROWD*** concordance. Opioid addicts seem less self-aware than others

Table 63 Happiness Self-Awareness Frequency across Groups

	FALSE	TRUE
GP	21%	79%
AA	21%	79%
SUBX	26%	74%

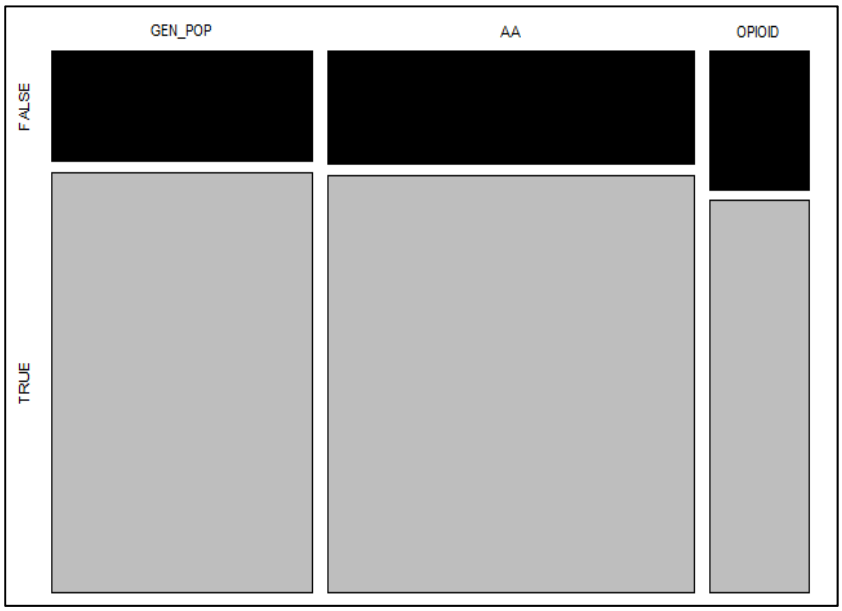


Figure 154 Happiness Self-Awareness Frequency across Groups

```
Formula: happySELFAWARE ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 7746 7760  -3871     7742
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.30728  0.55433
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.31254    0.06367   20.61   <2e-16 ***
```

Code Snippet 41 Happiness Self-Awareness Two-Level Null Model Calculation in R

The log-odds mean for happiness for an average participant ($\mu_{0j} = 0$) is significant is estimated at 1.31254 which is a probability 78.8%. The Log Likelihood test statistic between the one-level and two-level null model is 171 indicating evidence that between-participant variance is non-zero.

```
Formula: happySELFAWARE ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 6823 6850  -3407     6815
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.3106   0.55732
Number of obs: 6650, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.4317     0.1095  13.074   <2e-16 ***
group3AA      -0.1720     0.1595  -1.078   0.2810
group3OPIOID  -0.3170     0.1722  -1.841   0.0656 .
---
```

Code Snippet 42 Happiness Self-Awareness versus Group Model Calculation in R

From the coefficient estimates in Code Snippet 42:

$$logit(\pi_{ij}) = 1.4317 - 0.171958\text{group3AA} - 0.316996\text{group3OPIOID} + \mu_{0j} \quad (6.3)$$

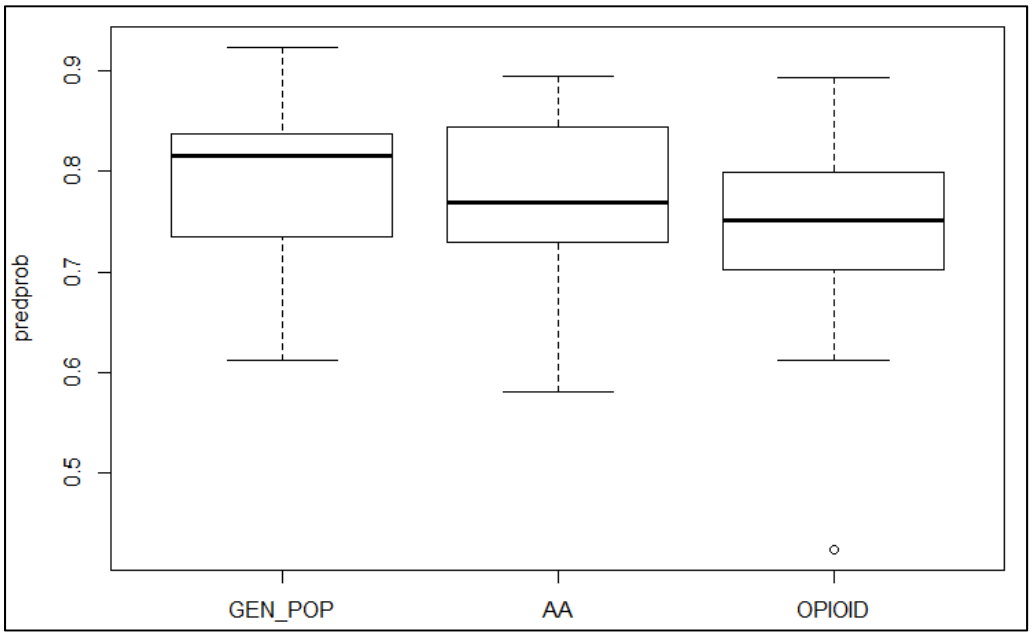


Figure 155 Predicted Probabilities of happySELFAWARE versus Group

The $R^2_{binomial} = 0.003249$ indicating that group3 describes a small amount of variance. The ICC is 0.086267 indicating some degree of correlation within groups (typical good range is 0.1 – 0.25). The deviance from the null 2-level model is highly significant with $X^2 =$ 927 on four degress of freedom.

There is a trend that Opioid addicts are 5.4% less self-aware of their happiness than the General Population ($p<0.1$). Figure 155 indicates Opioid addicts have a much wider IQR for sadness which indicates higher variance.

- GP pr(happySELFAWARE) = 80.7% (95% CI, 77.1%–83.9%)
- AA Member pr(happySELFAWARE) = 77.9% (95% CI, 71.9%–82.9%)
- Opioid Addict pr(happySELFAWARE) = 75.3% (95% CI, 68.4%–81.1%)

There is no significance on gender or language.

Happiness Empathy (happyEMPATHY)

There are no significant differences in happiness empathy across group3, gender, or language. Experiment conditions were not consistent across groups. General Population and AA members related to randomly chosen emotional recordings from General Population and AA members. Opioid addicts related to randomly chosen emotional recordings from Opioid addicts. However there is no significant difference between General Population and AA members even if OPIOID addicts are deleted from the data set.

HappyEMPATHY is a derived outcome variable from ***happyRELATE == happyCROWD*** concordance.

Table 64 Frequency of Happiness Empathy (happyEMPATHY) across Groups

	FALSE	**TRUE**
GP	18%	82%
AA	22%	78%
SUBX	16%	84%

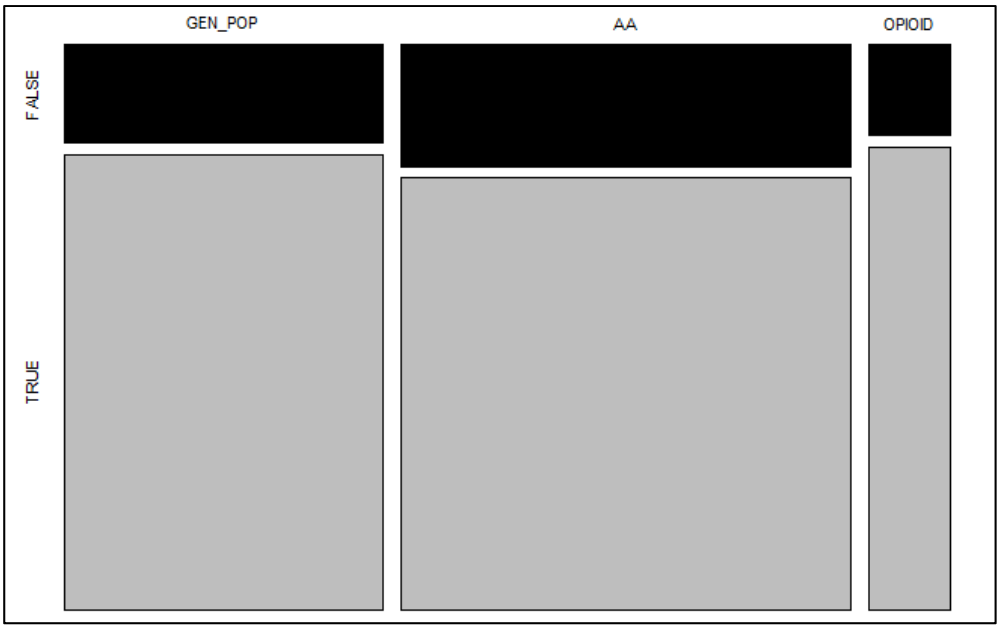


Figure 156 Frequency of Happiness Empathy (happyEMPATHY) across Groups

```
Formula: happyEMPATHY ~ (1 | p)
   Data: EMO
   AIC   BIC logLik deviance
 15616 15631  -7806    15612
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.095171 0.3085
Number of obs: 16001, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.51926    0.04086   37.19   <2e-16 ***
```

Code Snippet 43 Empathy to Happiness Two-Level Null Model in R

The log-odds mean for happiness empathy for an average participant ($\mu_{0j} = 0$) is significant is estimated at 1.51926 (probability 82.0%). The Log Likelihood test statistic between the one-level and two-level null model is 189 indicating evidence that between-participant variance is non-zero; however the residuals are not normally distributed (skew and Kurtosis indicate lognormal or gamma distribution) as evident in the Cullen and Frey graph in Figure 157.

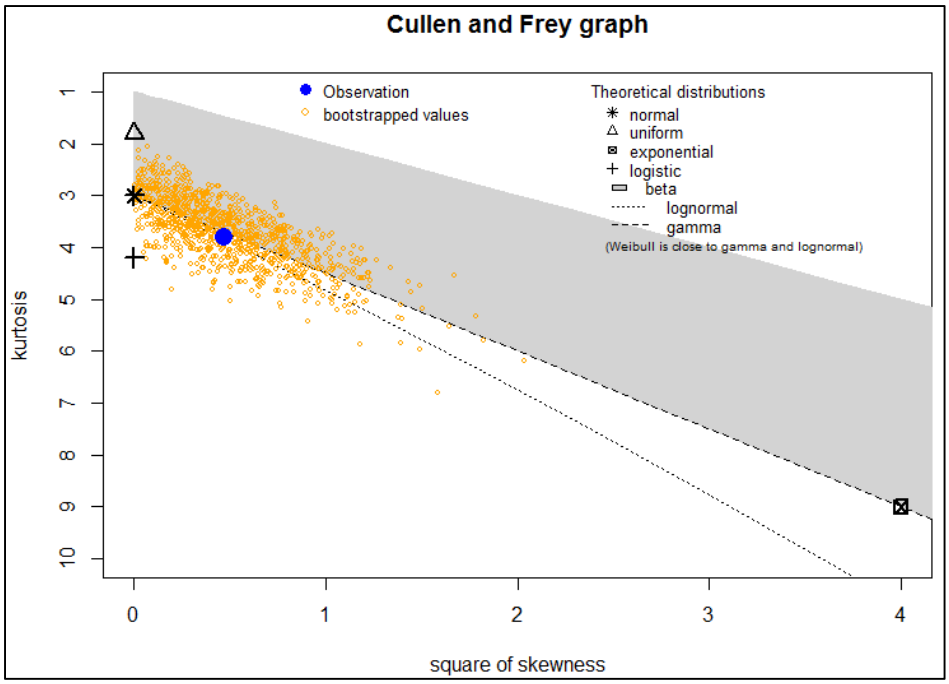


Figure 157 Cullen and Frey Graph of happyEMPATHY Residuals

```
Formula: happyEMPATHY ~ group3 + (1 | p)
   Data: EMO
   AIC   BIC logLik deviance
 13404 13434  -6698    13396
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.10265  0.32038
Number of obs: 13612, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.52793    0.06715  22.753   <2e-16 ***
group3AA     -0.11251    0.10111  -1.113    0.266
group3OPIOID  0.09278    0.12038   0.771    0.441
```

Code Snippet 44 Happiness Empathy versus Group Model in R

There is no significance on group, gender or language.

Figure 80 shows the AA member interquartile range is much wider than the General Population indicating more divergence of empathic ability of AA members that is speculated to be possibly correlated to length of sobriety and/or mood disorders. Males have a much wider IQR than females.

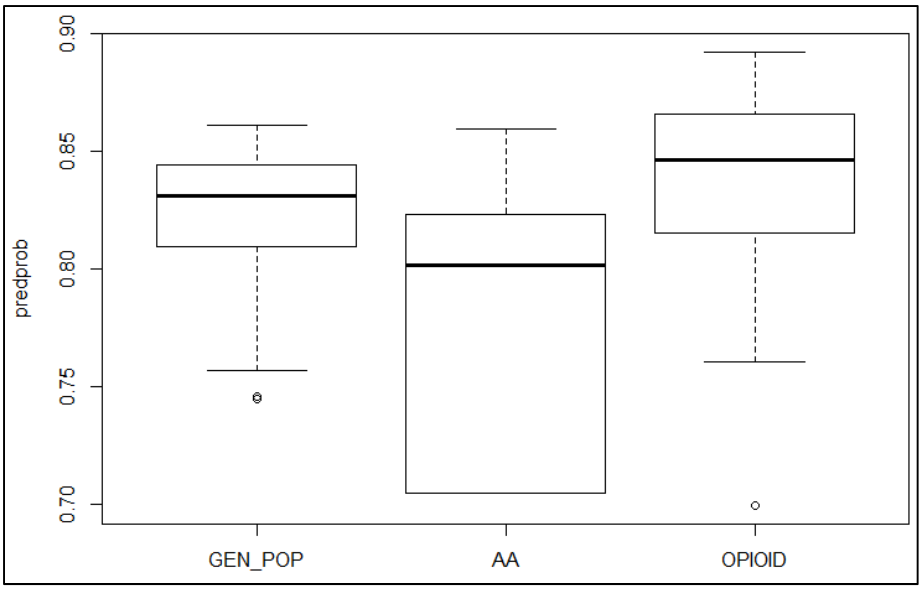


Figure 158 Predicted Probabilities of happyEMPATHY versus Group

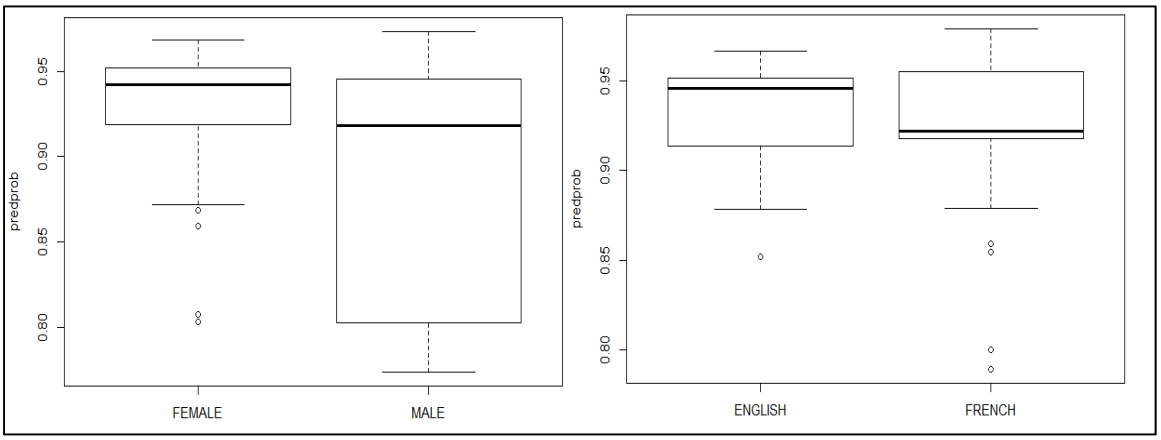


Figure 159 Predicted Prob of happyEMPATHY versus Gender and Language

There is no significance on group, gender or language.

APPENDIX E

SADNESS REGRESSION ANALYSIS

Sadness (sadCROWD)

AA members seem to be the least sad, and Opioid addicts the saddest.

Table 65 Frequency of Sadness (sadCROWD) across Groups

	FALSE	TRUE
GP	85%	15%
AA	88%	12%
SUBX	82%	18%

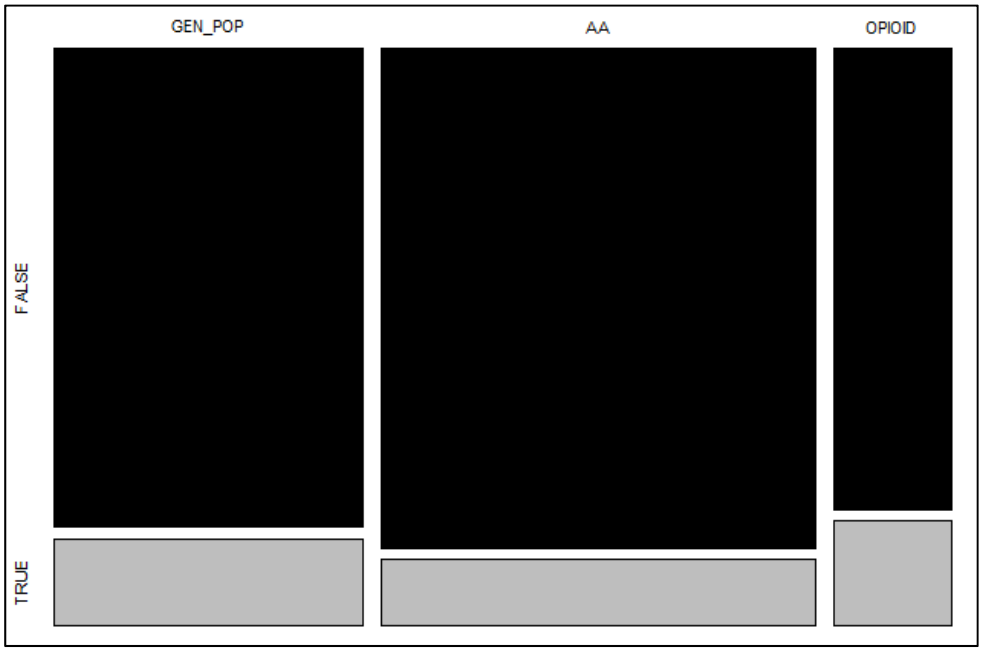


Figure 160 Frequency of Sadness (sadCROWD) across Groups

```
Formula: sadCROWD ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 5673 5687  -2834     5669
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 1.007    1.0035
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.032      0.106  -19.17   <2e-16 ***
```

Code Snippet 45 Sadness Two-Level Null Model in R

The log-odds mean for sadness for an average participant ($\mu_{0j} = 0$) is significant is estimated at -2.031672 which is a probability 11.6% (95% CI, 9.6%–13.9%). The Log Likelihood test statistic between the one-level and two-level null model is 557 indicating STRONG evidence that between-participant variance is non-zero.

```
Formula: sadCROWD ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 4869 4897  -2431     4861
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.9451   0.97216
Number of obs: 6650, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.940211   0.173541 -11.180   <2e-16 ***
group3AA     -0.465996   0.263204  -1.770   0.0766 .
group3OPIOID -0.002465   0.273283  -0.009   0.9928
```

Code Snippet 46 Sadness versus Group Model in R

From the coefficient estimates in Code Snippet 46:

$$logit(\pi_{ij}) = -1.940 - 0.465996\text{group3AA} - 0.002465\text{group3OPIOID} + \mu_{0j} \qquad (6.4)$$

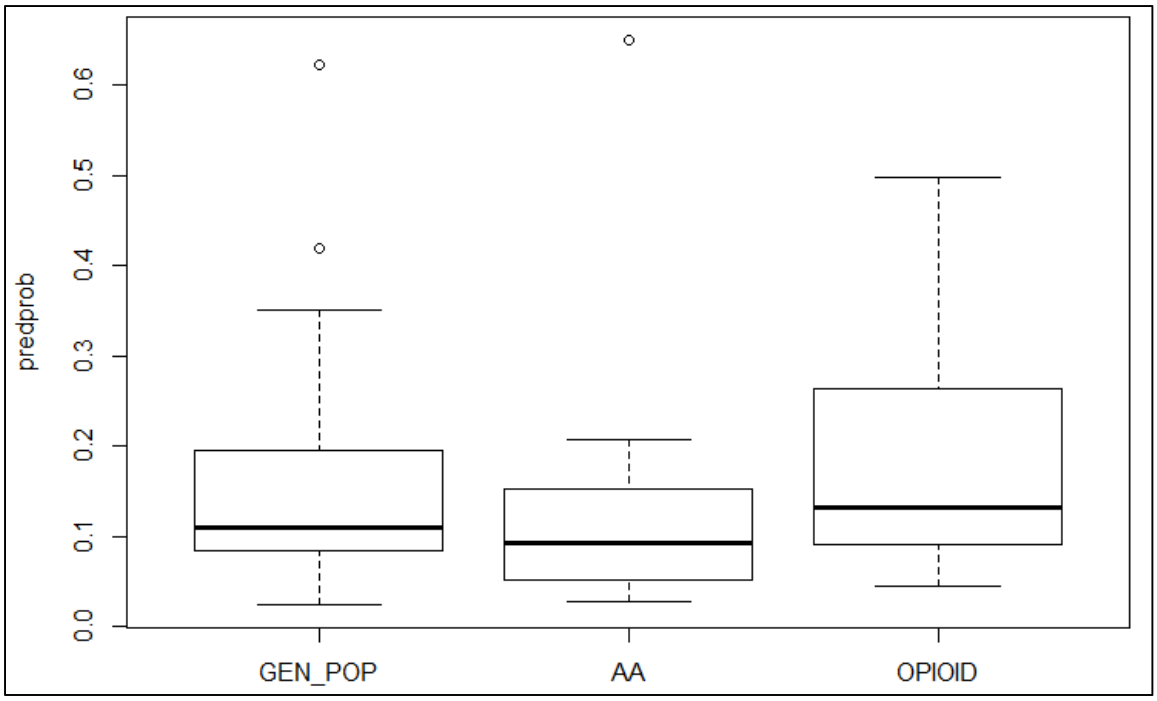


Figure 161 Predicted Probabilities of sadCROWD versus Group

The $R^2_{binomial}$ = 0.012621 indicating that group3 describes some of the ***sadCROWD*** variance. The ICC is 0.223167 indicating a high degree of correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 807$ on four degress of freedom.

Using equation (6.4 and Code Snippet 46, there is a trend that AA members are 4.3% less sad than the General Population ($p<0.1$). Figure 84 indicates Opioid addicts have a much wider IQR for sadness which indicates higher variance.

- General Population pr(sadCROWD) = 12.6% (95% CI, 9.2%–16.9%)
- AA Member pr(sadCROWD) = 8.3% (95% CI, 5.7%–13.2%)
- Opioid Addict pr(sadCROWD) = 12.5% (95% CI, 7.7%–19.8%)

```
Formula: sadCROWD ~ gender + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 5361 5382  -2677     5355
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.92826  0.96346
Number of obs: 7290, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7622     0.1553 -11.350  < 2e-16 ***
genderMALE   -0.5799     0.2115  -2.742  0.00611 **
```

Code Snippet 47 Sadness versus Gender Model in R

From the coefficient estimates for ***sadCROWD ~ gender***:

$$logit(\pi_{ij}) = -1.7622 \quad -0.5799\text{genderMALE} + \mu_{0j} \tag{6.5}$$

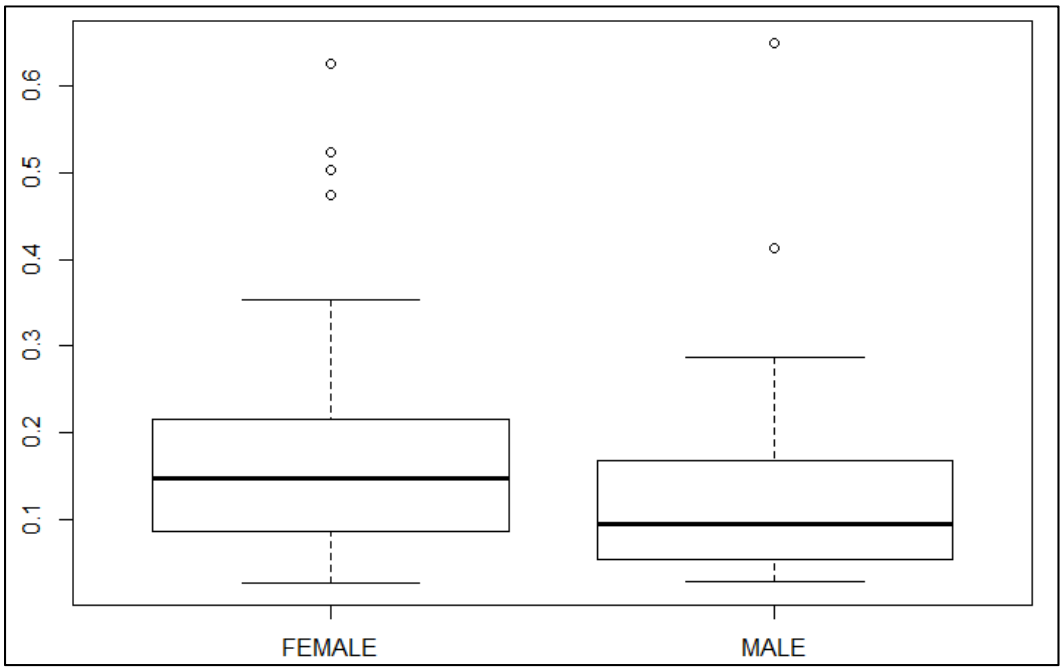


Figure 162 Predicted Probabilities of sadCROWD versus Gender

Using equation 6.5 and Code Snippet 47, Females have 5.9% more probability of being sad than Males ($p<0.05$). Males have an 8.7% probability of being sad (95% CI, 5.9%–12.8%) while the Female probability is 14.6% (95% CI, 11.2%–19.0%);

There is no difference in language.

Sadness Self-Awareness (sadSELFAWARE)

Opioid addicts seem the least self-aware of their sadness.

Table 66 Frequency of Sadness Self-Awareness across Groups

	FALSE	TRUE
GP	11%	89%
AA	11%	89%
SUBX	18%	82%

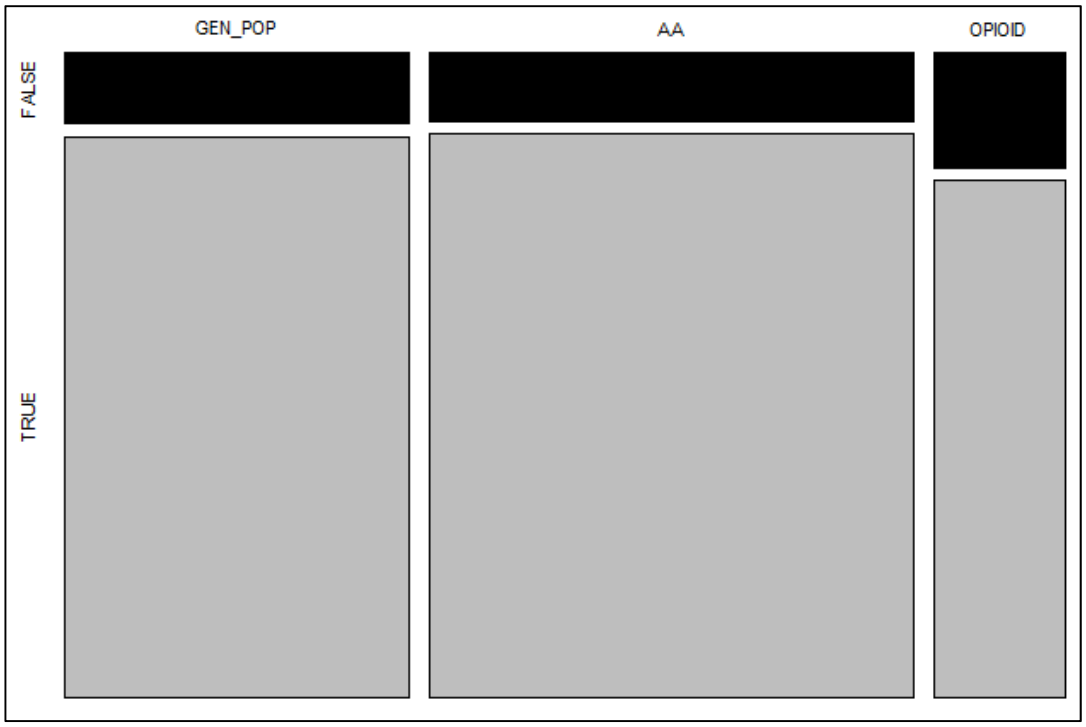


Figure 163 Frequency of Sadness Self-Awareness across Groups

```
Formula: sadSELFAWARE ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 5171 5185  -2584     5167
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.70621  0.84036
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.15740    0.09316   23.16   <2e-16 ***
```

Code Snippet 48 Sadness Self-Awareness Two-Level Null Model Calculation in R

The log-odds mean for self-awareness of sadness for an average participant ($\mu_{0j} = 0$) is significant is estimated at 2.15740 which is a probability 89.6% (95% CI, 87.8%–91.2%). The Log Likelihood test statistic between the one-level and two-level null model is 367; evidence that between-participant variance is non-zero.

```
Formula: sadSELFAWARE ~ group3 + (1 | p)
      Data: EMO
  AIC  BIC logLik deviance
 4571 4598  -2281     4563
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.58354  0.7639
Number of obs: 6650, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.1554     0.1467  14.690   <2e-16 ***
group3AA       0.1981     0.2194   0.903   0.3664
group3OPIOID  -0.3964     0.2277  -1.741   0.0817 .

>EMO$group3 <- relevel(EMO$group3,ref="AA")
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.3535     0.1631  14.433   <2e-16 ***
group3GEN_POP -0.1982     0.2194  -0.903   0.3663
group3OPIOID  -0.5946     0.2385  -2.493   0.0127 *
```

Code Snippet 49 Sadness Self-Awareness versus Group Model Calculation in R

From the coefficient estimates with GEN_POP as reference:

$$logit(\pi_{ij}) = 2.1554 + 0.1981D_{i1} - 0.3964D_{i2} + \mu_{0j} \quad (6.6)$$

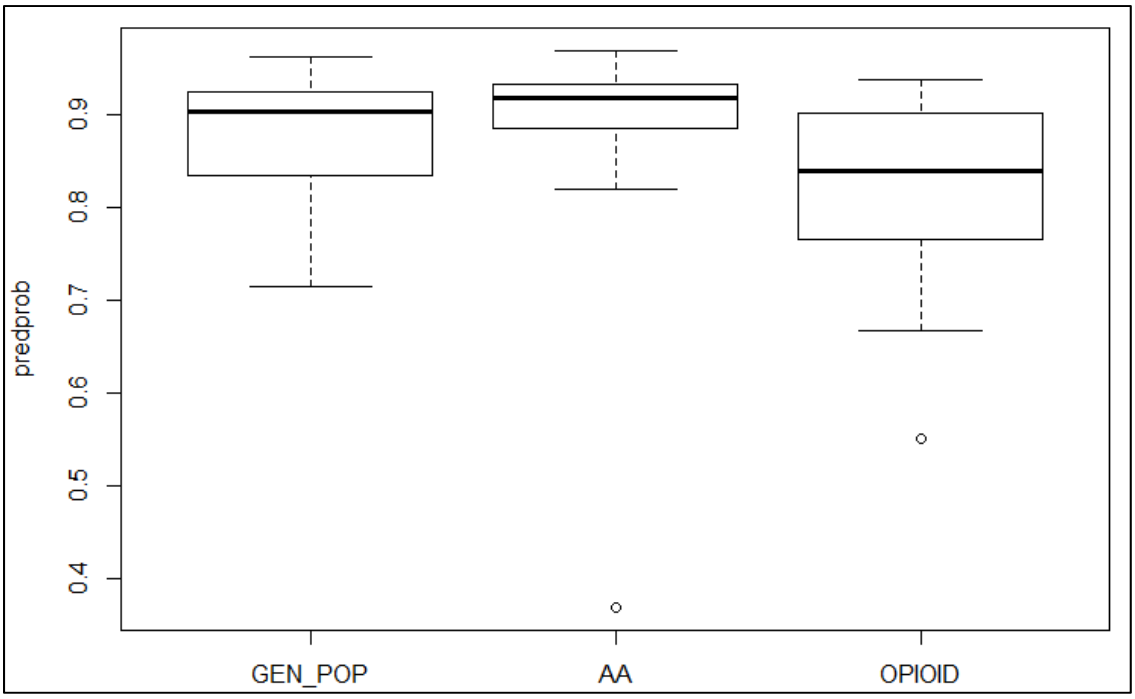


Figure 164 Predicted Probabilities of sadSELFAWARE versus Group

The $R^2_{binomial}$ = 0.010020 indicating that group3 describes some of the ***sadSELFAWARE*** variance. The ICC is 0.150653 indicating correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 604$ on four degress of freedom.

AA members are 6% more self-aware of their sadness than Opioid addicts ($p<0.05$). There is a trend that the General Population is 4.3% more self-aware of their sadness than Opioid addicts ($p<0.1$).

- General Population pr(sadSELFAWARE) = 89.6% (95% CI, 84.8%–93.0%);
- AA pr(sadSELFAWARE) = 91.3% (95% CI, 88.3%–93.6%);
- Opioid addict pr(sadSELFAWARE) = 85.3% (95% CI, 78.3%–90.3%);

```
Formula: sadSELFAWARE ~ gender + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 4985 5006  -2489     4979
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.6813   0.82541
Number of obs: 7290, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9475     0.1396  13.952   <2e-16 ***
genderMALE    0.3645     0.1888   1.931   0.0535 .
```

Code Snippet 50 Sadness Self-Awareness versus Gender Model Calculation in R

From the coefficient estimates for ***sadSELFAWARE ~ gender***:

$$logit(\pi_{ij}) = 1.9475 + 0.3645X_{i1} + \mu_{0j} \quad (6.7)$$

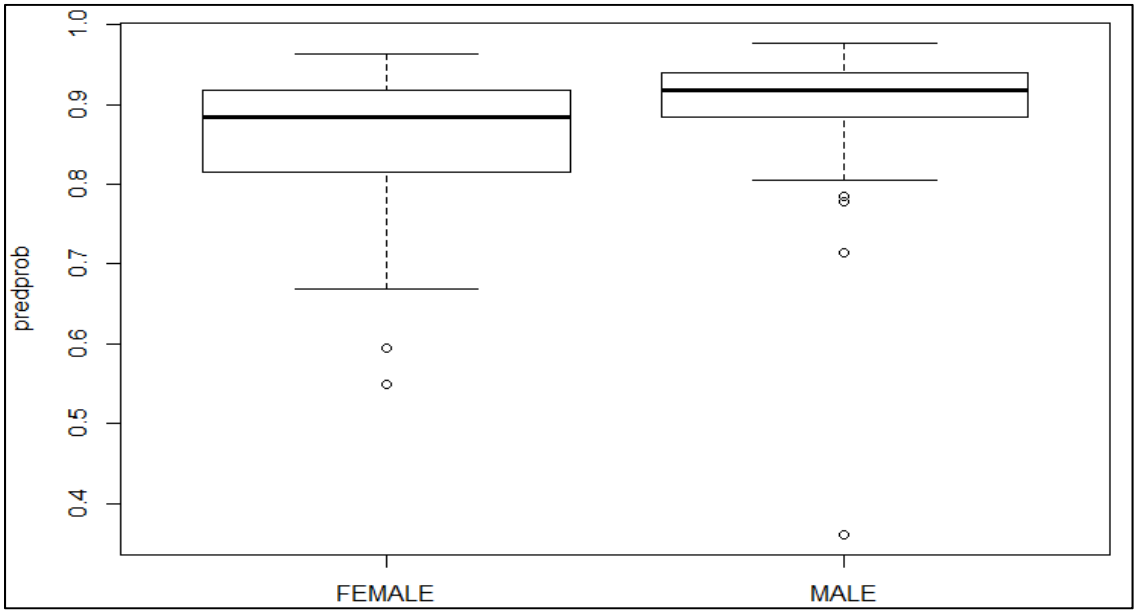


Figure 165 Predicted Probabilities of sadSELFWARE versus Gender

There is a trend that Males are 3.5% more self-aware of their sadness than Females ($p<0.1$).

- Female pr(sadSELFAWARE) = 87.5% (95% CI, 84.1%–90.3%);
- Male pr(sadSELFAWARE) = 91.0% (95% CI, 87.4%–93.6%);

There is no difference in language.

Sadness Empathy (sadEMPATHY)

There is no apparent differences in empathy towards sadness (***sadEMPATHY***) frequencies across groups.

Table 67 Frequency of Sadness Empathy (sadEMPATHY) across Groups

	FALSE	TRUE
GP	11%	89%
AA	12%	88%
SUBX	11%	89%

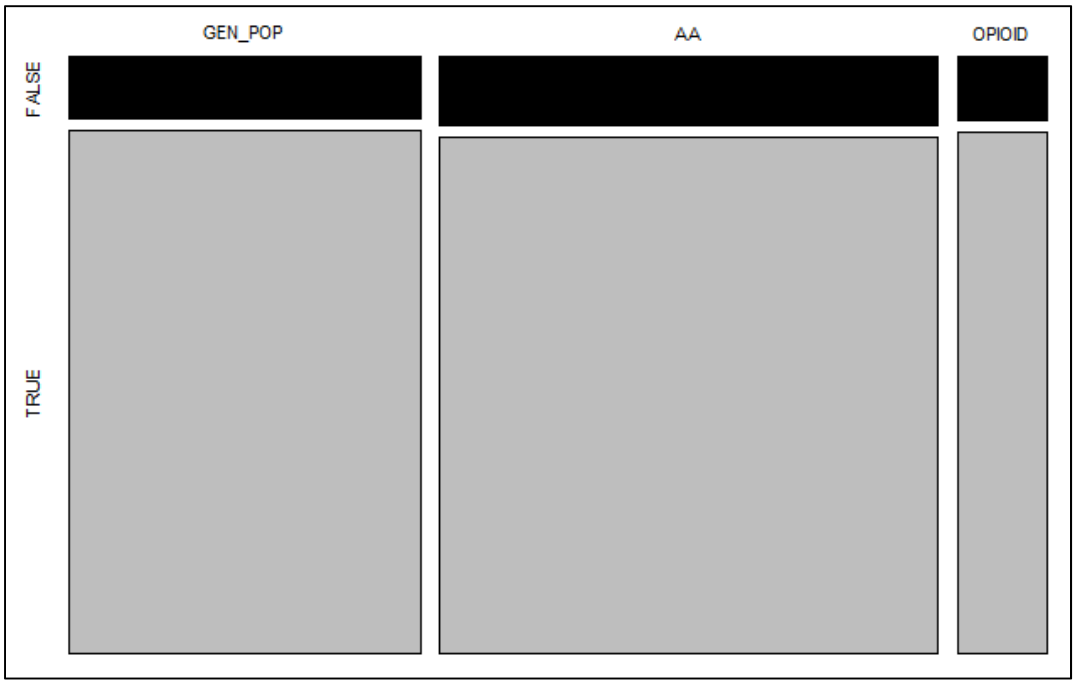


Figure 166 Frequency of Sadness Empathy (sadEMPATHY) across Groups

```
> g1 <- glmer(sadEMPATHY~ (1|p) ,family = binomial, data=EMO)
Generalized linear mixed model fit by the Laplace approximation
Formula: sadEMPATHY ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 9471 9486  -4733     9467
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.089266 0.29877
Number of obs: 13612, groups: p, 112

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.15657    0.04803    44.9   <2e-16 ***
```

Code Snippet 51 Empathy to Sadness Two-Level Null Model in R

The log-odds mean for self-awareness of sadness for an average participant ($\mu_{0j} = 0$) is significant is estimated at 2.15657 that is a probability 89.9% (95% CI, 88.7%–90.5%). The Log Likelihood test statistic between the one-level and two-level null model is 1621.8; evidence that between-participant variance is non-zero.

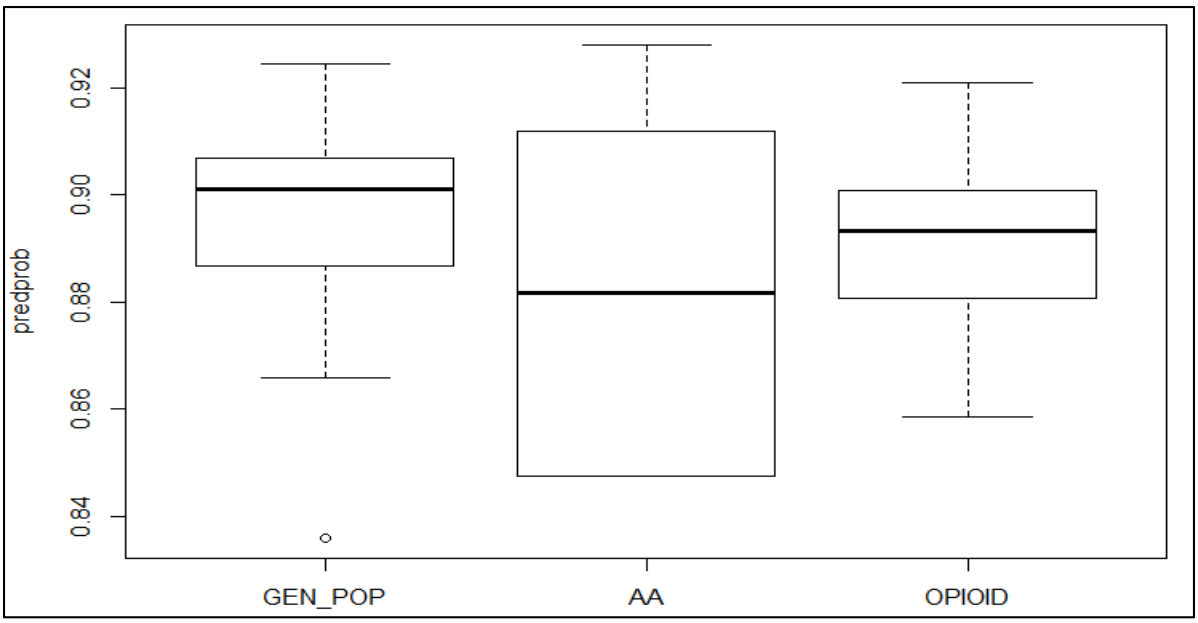


Figure 167 Predicted Probabilities of sadEMPATHY versus Group

Figure 167 indicates AA members have a much wider IQR for empathy of sadness which may be related to the length of their sobriety and/or comorbidity of mood disorders.

APPENDIX F

ANXIETY REGRESSION ANALYSIS

Anxiety (anxCROWD)

AA members see to be more anxious than General population & Opioid addicts.

Table 68 Frequency of Anxiety (anxCROWD) across Groups

	FALSE	TRUE
GP	96%	4%
AA	92%	8%
SUBX	96%	4%

Figure 168 Frequency of Anxiety (anxCROWD) across Groups

```
Formula: anxCROWD ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3222 3236  -1609     3218
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.96622  0.98297
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2155     0.1192  -26.97   <2e-16 ***
```

Code Snippet 52 Anxiety Two-Level Null Model in R

The log-odds mean for sadness for an average participant ($\mu_{0j} = 0$) is significant is estimated at -3.2155 that is a probability 3.9% (95% CI, 3.1%–4.8%). The Log Likelihood test statistic between the one-level and two-level null model is 217 indicating evidence that between-participant variance is non-zero.

```
Formula: anxCROWD ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 2683 2710  -1337     2675
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 1.0428   1.0212
Number of obs: 6650, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.4675     0.2140 -16.203   <2e-16 ***
group3AA       0.4869     0.3044   1.600    0.110
group3OPIOID  -0.3161     0.3646  -0.867    0.386

EMO$group3 <- relevel(EMO$group3,ref="AA")
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.9806     0.2164 -13.772   <2e-16 ***
group3GEN_POP  -0.4869     0.3044  -1.600   0.1096
group3OPIOID   -0.8030     0.3660  -2.194   0.0282 *
```

Code Snippet 53 Anxiety versus Group Model in R

From the coefficient estimates after re-leveling to AA members in Code Snippet 53:

$$logit(\pi_{ij}) = -2.9806 \quad -0.4869D_{i1} - 0.8030D_{i2} + \mu_{0j} \tag{6.8}$$

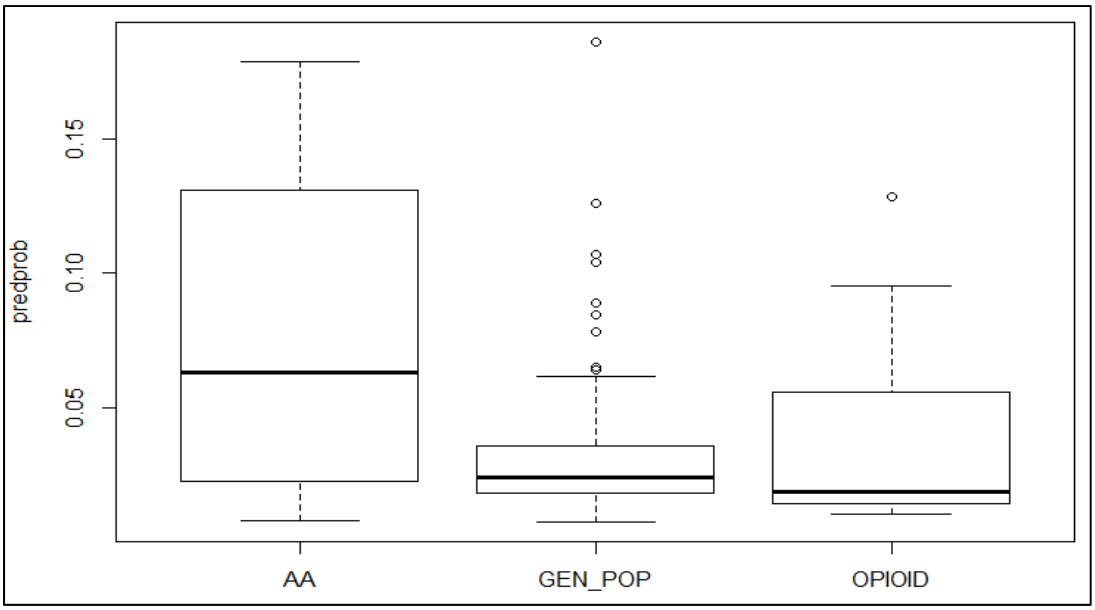


Figure 169 Predicted Probabilities of anxCROWD versus Group

The $R^2_{binomial}$ = 0.020858 indicating that group3 describes a portion of the anxCROWD variance. The ICC is 0.240682 indicating a high degree of correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 543$ on four degress of freedom.

AA members have 2.6% significant probability of being more anxious than the Opioid addicts (p<0.1). AA members have a 4.8% probability of being anxious (95% CI, 1.7%–5.4%). The probability of an Opioid addict being anxious is 2.2% (95% CI, 1.1%–4.5%); Figure 169 indicates a wide IQR for AA members indicating higher variance.

```
Formula: anxCROWD ~ gender + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3085 3105  -1539     3079
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 1.0425   1.021
Number of obs: 7290, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30060    0.19198 -17.192   <2e-16 ***
genderMALE   0.07597    0.25424   0.299    0.765
```

Code Snippet 54 Anxiety versus Gender Model in R

There is no significant difference of anxiety across gender. However, Figure 170 indicate a much higher IQR for males.

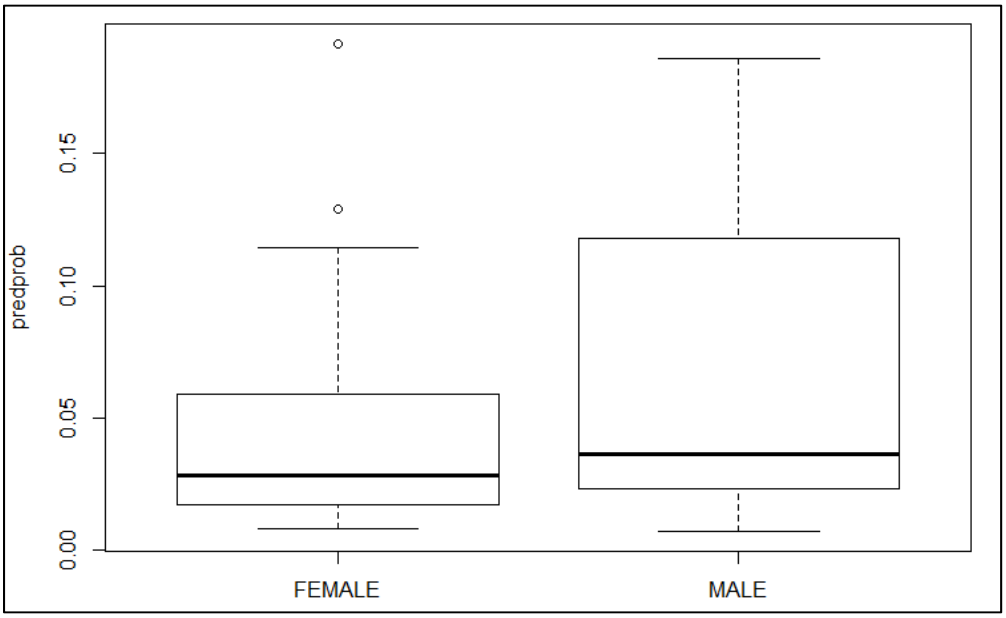


Figure 170 Predicted Probabilities of anxCROWD versus Gender

The General Population is the only group with French participants and the data is therefore filtered before analysis.

```
EMO <- EMO[(EMO$group3 == "GEN_POP"),]
Formula: anxCROWD ~ language + (1 | p)
  Data: EMO
   AIC   BIC logLik deviance
 734.6 751.9 -364.3    728.6
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.90747  0.95261
Number of obs: 2387, groups: p, 44

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.2854     0.2797  -11.74   <2e-16 ***
languageFRENCH  -0.3030     0.3835   -0.79    0.429
```

Code Snippet 55 Anxiety versus Language Model in R

There is no significant difference in anxiety across language. However, there is a larger IQR for English speakers.

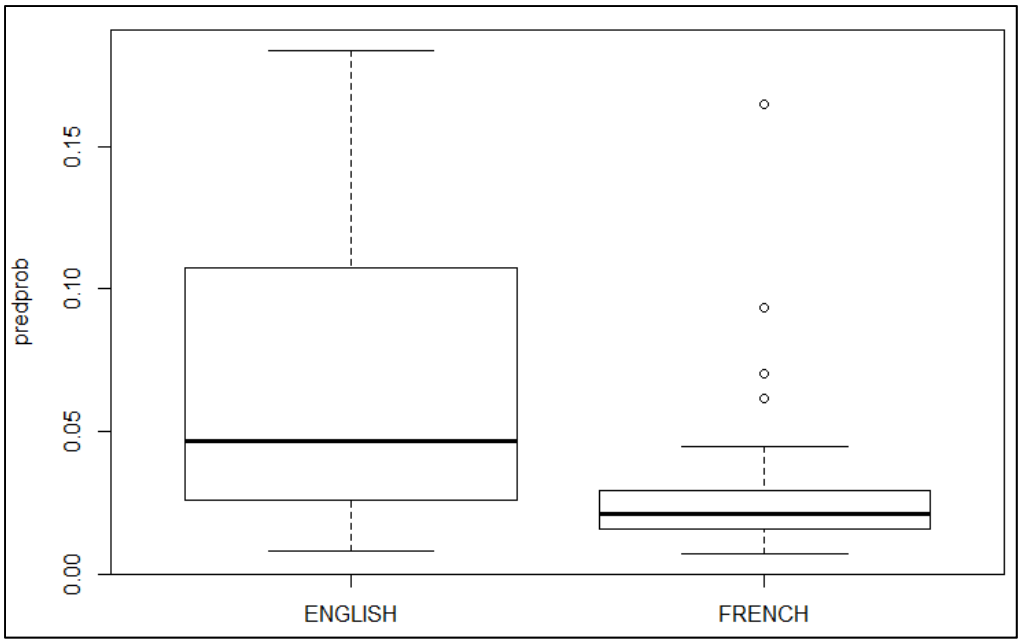


Figure 171 Predicted Probabilities of anxCROWD versus Language

Anxiety Self-Awareness (anxSELFAWARE)

Opioid addicts seem the least self-aware of their anxiety.

Table 69 Frequency of Anxiety Self-Awareness across Groups

	FALSE	TRUE
GP	5%	95%
AA	7%	93%
SUBX	11%	89%

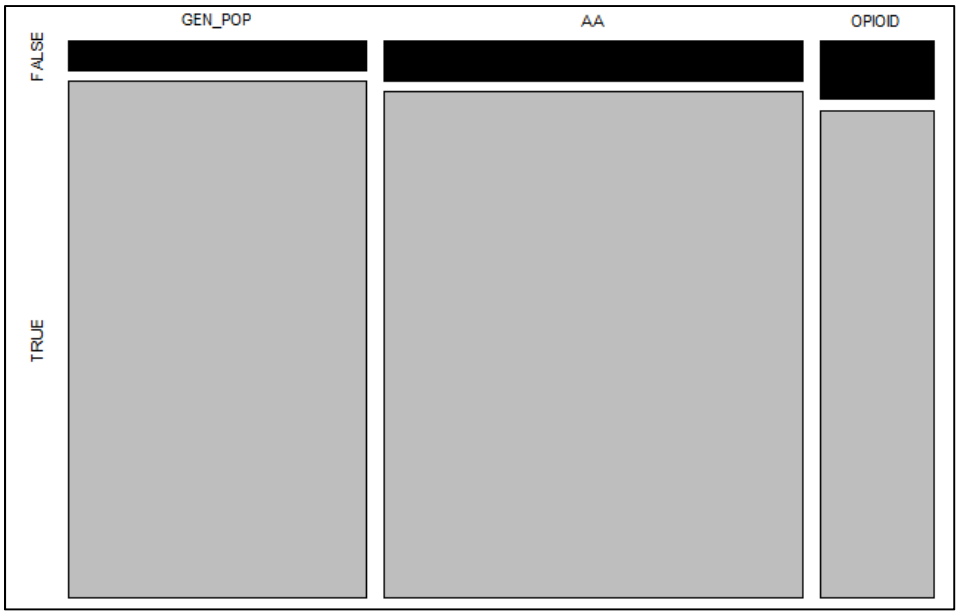


Figure 172 Frequency of Anxiety Self-Awareness across Groups

```
Formula: anxSELFAWARE ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3461 3475  -1729     3457
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 1.0608   1.0299
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9327     0.1177   24.92   <2e-16 ***
```

Code Snippet 56 Anxiety Self-Awareness Two-Level Null Model in R

The log-odds mean for self-awareness of anxiety for an average participant ($\mu_{0j} = 0$) is significant is estimated at 2.9327 which is a probability of 94.9% (95% CI, 93.7%–96.0%). The Log Likelihood test statistic between the one-level and two-level null model is 331; evidence that between-participant variance is non-zero.

```
Formula: anxSELFAWARE ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3111 3139  -1552     3103
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.97305  0.98643
Number of obs: 6650, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.11559    0.19804  15.732   <2e-16 ***
group3AA     -0.04344    0.29074  -0.149    0.881
group3OPIOID -0.70155    0.29903  -2.346    0.019 *
>   EMO$group3 <- relevel(EMO$group3,ref="AA")
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.07215    0.21286  14.433   <2e-16 ***
group3GEN_POP  0.04344    0.29074   0.149   0.8812
group3OPIOID  -0.65810    0.30904  -2.129   0.0332 *
```

Code Snippet 57 Anxiety Self-Awareness versus Group Model in R

From the coefficient estimates with GEN_POP as reference:

$$logit(\pi_{ij}) = 3.11559 - 0.04344D_{i1} - 0.70155D_{i2} + \mu_{0j} \quad (6.9)$$

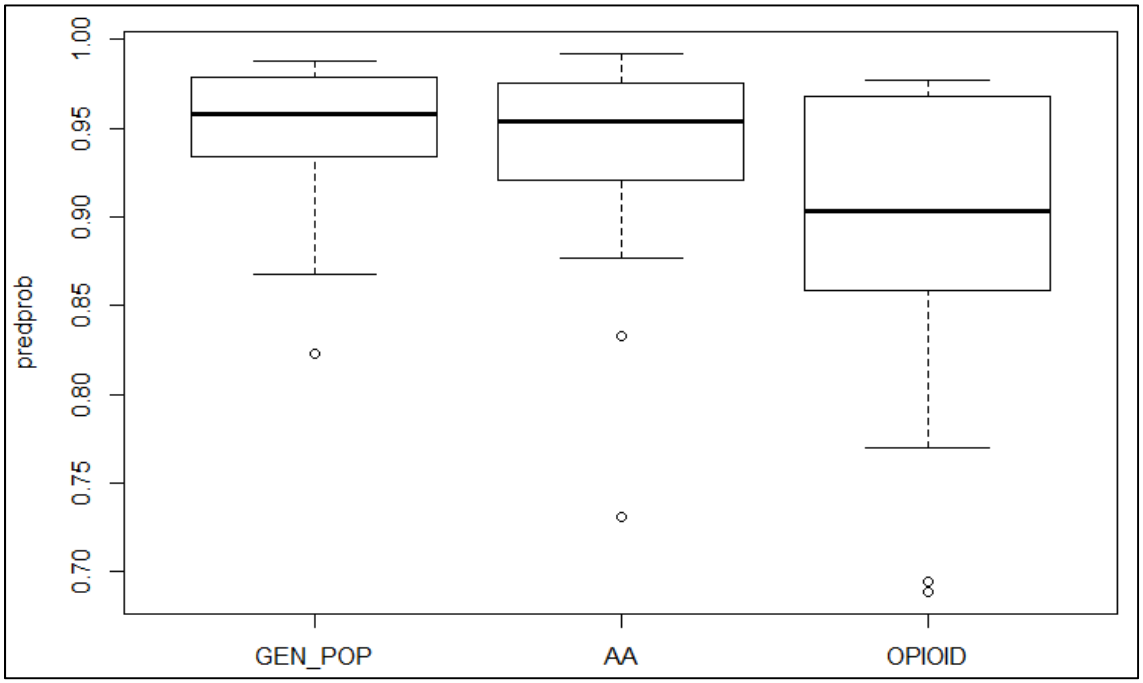


Figure 173 Predicted Probabilities of anxSELFAWARE versus Group

The $R^2_{binomial}$ = 0.012590 indicating that group3 describes some of the anxSELFAWARE variance. The ICC is 0.228258 indicating strong correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 354$ on four degress of freedom.

Opioid Addicts are 3.9% and 4.3% less self-aware of their anxiety than the General Population and AA members respectively ($p<0.05$).

- General Population pr(self-aware) = 95.7% (95% CI, 93.8%–97.1%);
- AA pr(self-aware) = 95.5% (95% CI, 92.3%–97.5%);
- Opioid addict pr(self-aware) = 91.8% (95% CI, 86.0%–95.3%);

There is no significant difference across gender or language.

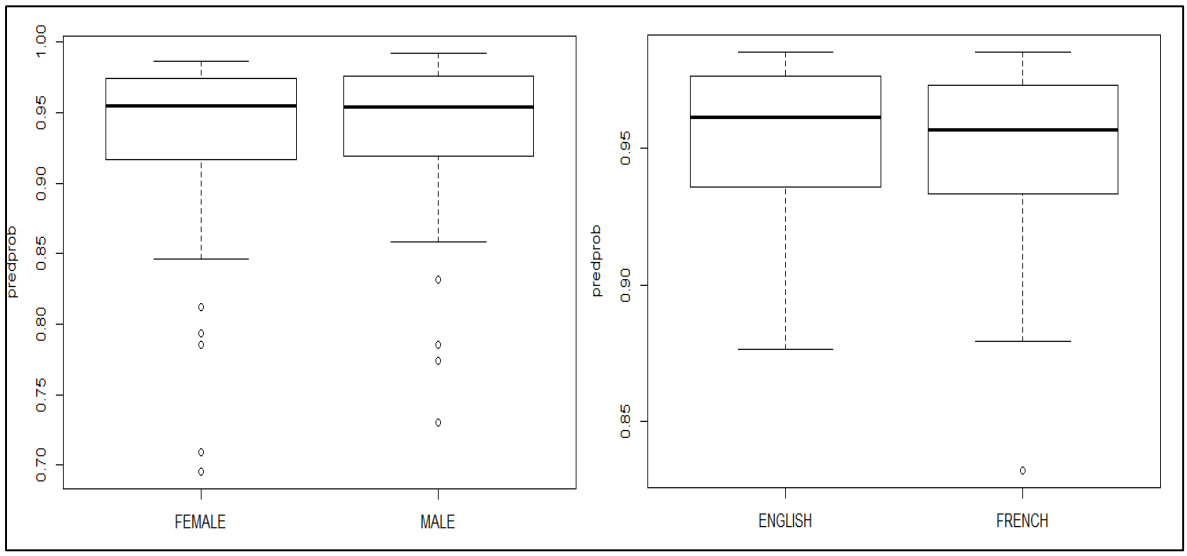


Figure 174 Predicted Prob of anxSELFAWARE versus Gender and Language

Anxious Empathy (anxEMPATHY)

AA members seem less able to empathize with other people's anxiety.

Table 70 Frequency of Empathy to Anxiety (anxEMPATHY) across Groups

	FALSE	TRUE
GP	7%	93%
AA	13%	87%
SUBX	7%	93%

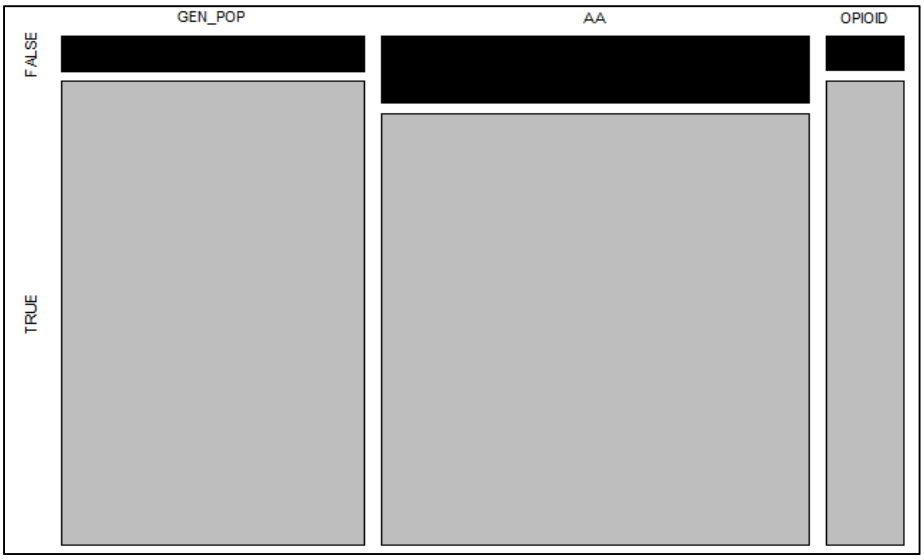


Figure 175 Frequency of Empathy to Anxiety (anxEMPATHY) across Groups

```
Formula: anxEMPATHY ~ (1 | p)
   Data: EMO
   AIC   BIC logLik deviance
 10002 10018  -4999     9998
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.47938  0.69237
Number of obs: 16001, groups: p, 129
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.54248    0.07876   32.28   <2e-16 ***
```

Code Snippet 58 Empathy to Anxiety Two-Level Null Model in R

The log-odds mean for self-awareness of anxiety for an average participant ($\mu_{0j} = 0$) is significant is estimated at 2.54248 which is a probability of 92.7% (95% CI, 91.6%–93.7%). The Log Likelihood test statistic between the one-level and two-level null model is 640; evidence that between-participant variance is non-zero.

```
Formula: anxEMPATHY ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 8523 8553  -4258     8515
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.41731  0.64599
Number of obs: 13612, groups: p, 112
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.2410     0.1349  16.614   <2e-16 ***
group3GEN_POP  0.4211     0.1831   2.300   0.0214 *
group3OPIOID   0.4281     0.2168   1.974   0.0484 *
```

Code Snippet 59 Anxiety Empathy versus Group Model in R

From the coefficient estimates with AA as reference:

$$logit(\pi_{ij}) = 2.2410 + 0.4211D_{i1} + 0.4281D_{i2} + \mu_{0j} \quad (6.10)$$

The $R^2_{binomial}$ = 0.011855 indicating that group3 describes some of the anxEMPATHY variance. The ICC is 0.112568 indicating correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 1483$ on four degress of freedom.

AA Members are 3.1% less empathetic towards the anxiety of others than the General Population and Opioid addicts ($p<0.05$).

- General Population pr(self-aware) = 93.5% (95% CI, 90.9%–95.4%);
- AA pr(self-aware) = 90.4% (95% CI, 87.8%–92.5%);
- Opioid addict pr(self-aware) = 93.5% (95% CI, 90.3%–95.7%);

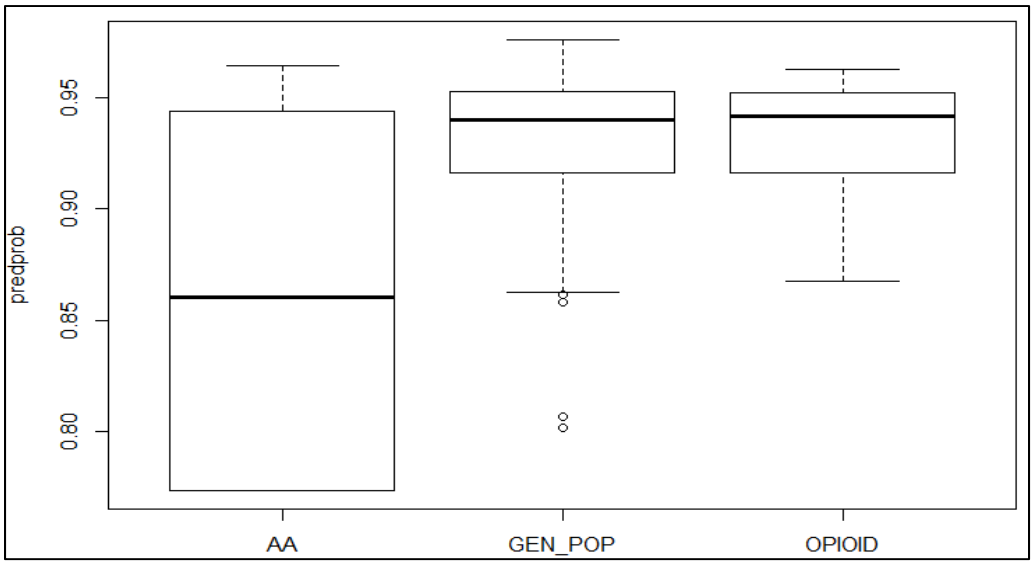


Figure 176 Predicted Probabilities of anxEMPATHY versus Group

Figure 176 indicates AA members have a much wider IQR for empathy of anxiety that may be related to the length of their sobriety and/or comorbidity of mood disorders.

```
Formula: anxEMPATHY ~ gender + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 9341 9364  -4667     9335
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.41903  0.64733
Number of obs: 15216, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.6929     0.1186  22.704   <2e-16 ***
genderMALE   -0.2712     0.1559  -1.739    0.082 .
```

Code Snippet 60 Anxiety Empathy versus Gender Model in R

From the coefficient estimates:

$$logit(\pi_{ij}) = 2.6929 \quad - 0.2712\text{genderMALE} + \mu_{0j} \tag{6.11}$$

The $R^2_{binomial}$ = 0.004502 indicating that group3 describes a small portion of the anxEMPATHY variance. The ICC is 0.112981 indicating correlation within groups. The

deviance from the null 2-level model is highly significant with $X^2 = 663$ on four degress of freedom.

There is a trend indicating Females are 3.1% more empathetic towards the anxiety of others than Males ($p<0.1$).

- Female pr(self-aware) = 94.9% (95% CI, 92.1%–94.9%);
- Male pr(self-aware) = 91.8% (95% CI, 89.2%–93.9%);

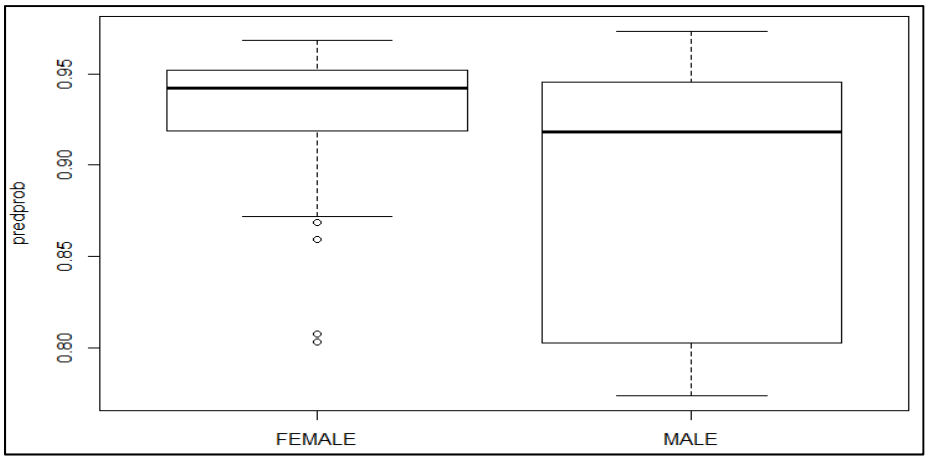


Figure 177 Predicted Probabilities of anxEMPATHY versus Gender

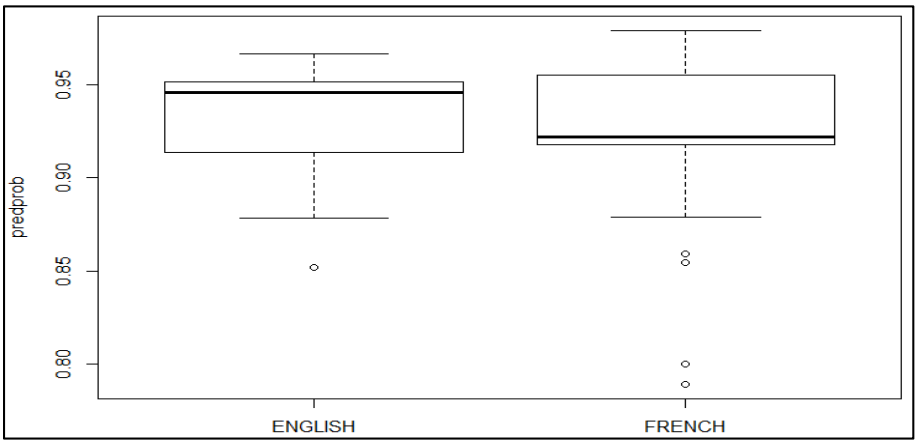


Figure 178 Predicted Probabilities of anxEMPATHY versus Language

There is no effect on language.

APPENDIX G

ANGER REGRESSION ANALYSIS

Anger (angryCROWD)

General population seems the least angry.

Table 71 Frequency of Anger (angryCROWD) across groups

	FALSE	TRUE
GP	96%	4%
AA	90%	10%
SUBX	91%	9%

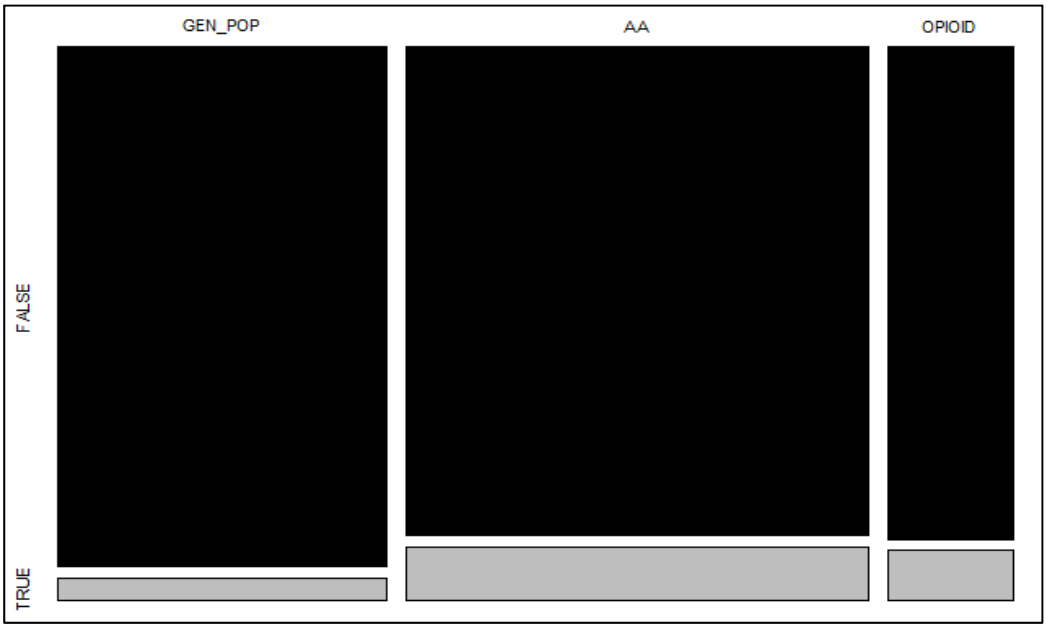


Figure 179 Frequency of Anger (angryCROWD) across groups

The log-odds mean for anger for an average participant ($\mu_{0j} = 0$) is significant is estimated at -3.0792 that is a probability 4.4% (95% CI, 3.4%–5.6%). The Log Likelihood test statistic between the one-level and two-level null model is 442 indicating evidence that between-participant variance is non-zero

```
Formula: angryCROWD ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3651 3665  -1824     3647
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 1.3509   1.1623
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.0792     0.1314  -23.43   <2e-16 ***
```

Code Snippet 61 Anger Two-Level Null Model in R

There are no significant differences in the probability of anger across groups. However, the IQR range is larger for AA members and Opioid addicts.

```
Formula: angryCROWD ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3196 3224  -1594     3188
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 1.3369   1.1563
Number of obs: 6650, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.4341     0.2305 -14.896   <2e-16 ***
group3AA       0.4992     0.3320   1.504    0.133
group3OPIOID   0.5071     0.3539   1.433    0.152
```

Code Snippet 62 Anger versus Group Model in R

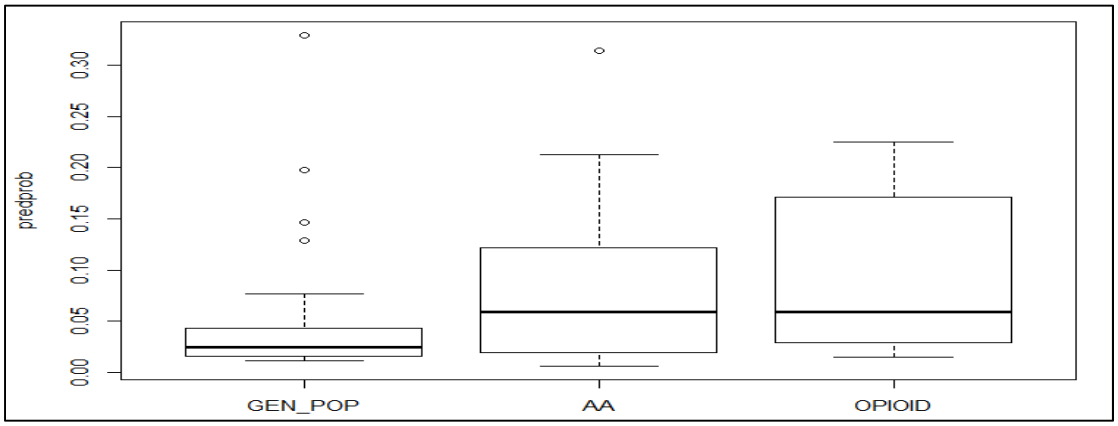


Figure 180 Predicted Probabilities of angryCROWD versus Group

```
Formula: angryCROWD ~ gender + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3474 3494  -1734     3468
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 1.4321   1.1967
Number of obs: 7290, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2737     0.2132 -15.354   <2e-16 ***
genderMALE    0.2472     0.2802   0.883    0.377
```

Code Snippet 63 Anger versus Gender Model in R

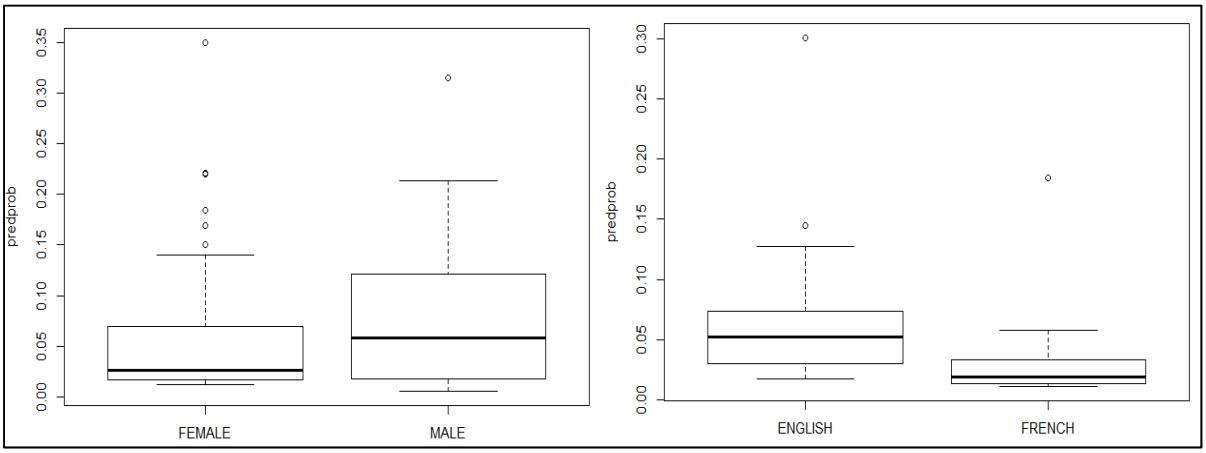


Figure 181 Predicted Prob of angryCROWD versus Gender and Language

There is no significant difference of anxiety across gender or language. However, Figure 181 indicates a much higher IQR for males.

Anger Self-Awareness (angrySELFAWARE)

Opioid addicts seem to be the least aware of their anger.

Table 72 Frequency of Anger Self-Awareness across Groups

	FALSE	TRUE
GP	4%	96%
AA	6%	94%
SUBX	7%	93%

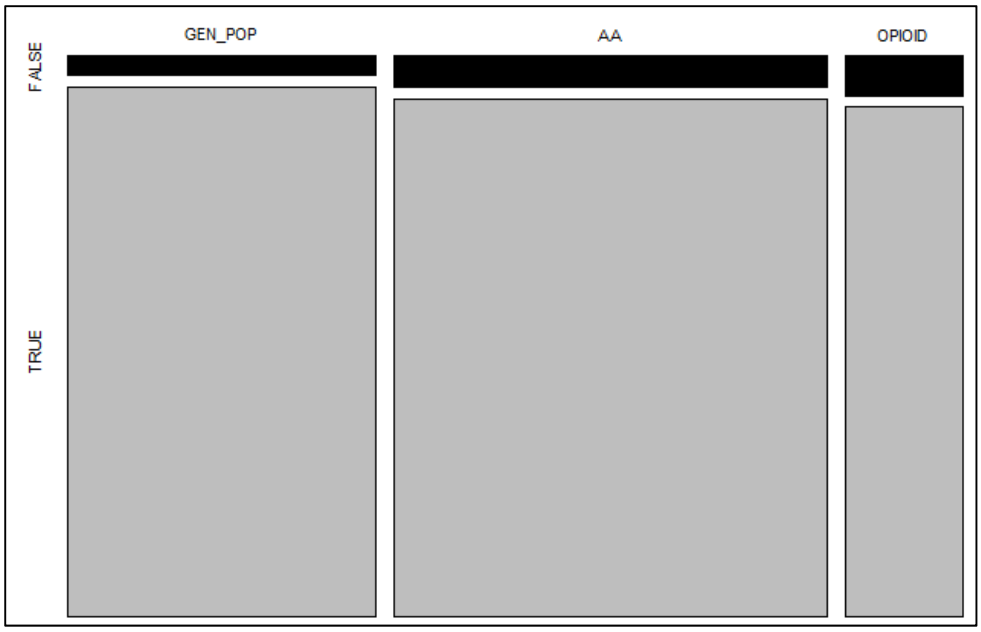


Figure 182 Frequency of Anger Self-Awareness across Groups

```
Formula: angrySELFAWARE ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 2834 2848  -1415     2830
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.91529  0.95671
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.2913     0.1182   27.85   <2e-16 ***
```

Code Snippet 64 Anger Self-Awareness Two-Level Null Model in R

The log-odds mean for self-awareness of anxiety for an average participant ($\mu_{0j} = 0$) is significant is estimated at 3.2913 which is a probability of 96.4% (95% CI, 95.5%–97.1%). The Log Likelihood test statistic between the one-level and two-level null model is 222; evidence that between-participant variance is non-zero.

There are no differences across groups.

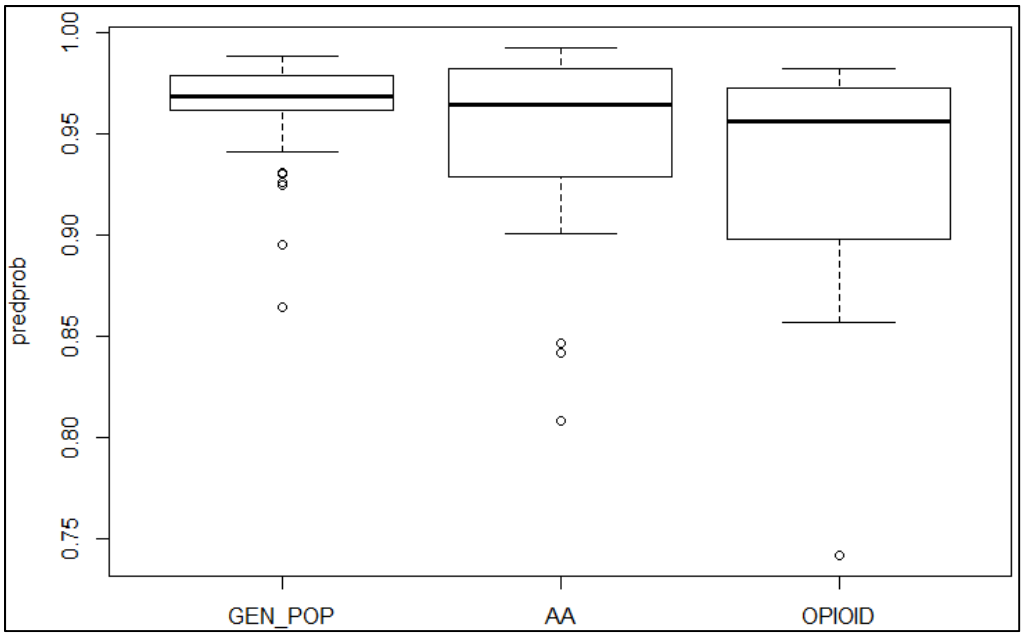


Figure 183 Predicted Prob of angrySELFAWARE versus Group

There is no significant difference across gender or language.

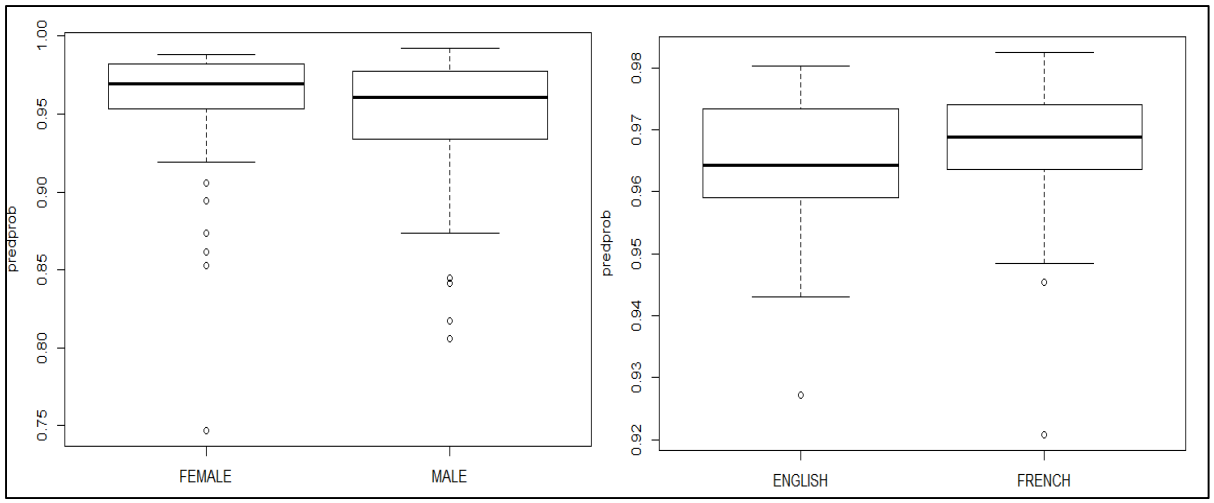


Figure 184 Predicted Prob of angrySELFAWARE versus Gender and Language

Anger Empathy (angryEMPATHY)

There is no difference in ability to empathize with anger.

Table 73 Frequency of Anger Empathy (angryEMPATHY) across groups

	FALSE	TRUE
GP	6%	94%
AA	6%	94%
SUBX	5%	95%

```
Formula: angryEMPATHY ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 7173 7188  -3585     7169
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.18534  0.43051
Number of obs: 16001, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.85543    0.06206   46.01   <2e-16 ***
```

Code Snippet 65 Anger Self-Awareness Two-Level Null Model in R

The log-odds mean for self-awareness of anger for an average participant ($\mu_{0j} = 0$) is significant is estimated at 2.85543 which is a probability of 94.6% (95% CI, 93.9%–95.2%). The Log Likelihood test statistic between the one-level and two-level null model is 78; slight evidence that between-participant variance is non-zero.

There are no differences across groups.

```
Formula: angryEMPATHY ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 6091 6121  -3042     6083
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.19567  0.44235
Number of obs: 13612, groups: p, 112

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.81757    0.10018  28.125   <2e-16 ***
group3AA     0.01751    0.15125   0.116    0.908
group3OPIOID 0.19167    0.18886   1.015    0.310
```

Code Snippet 66 Anger Empathy versus Group Model in R

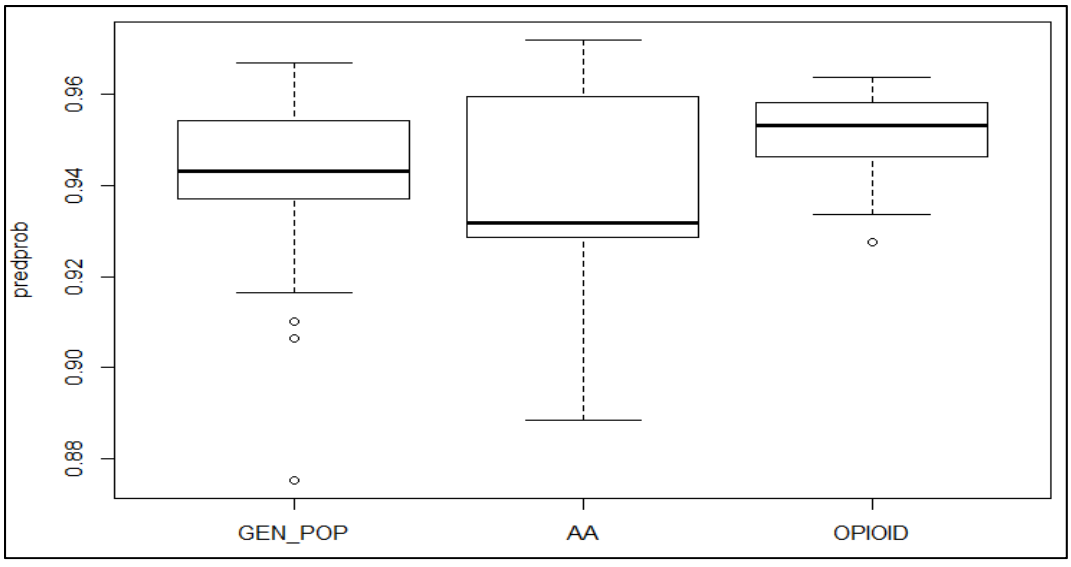


Figure 185 Predicted Probabilities of angryEMPATHY versus Group

Figure 90 indicates AA members have a much wider IQR for empathy of anger that may be related to the length of their sobriety and/or comorbidity of mood disorders.

```
Formula: angryEMPATHY ~ gender + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 6750 6772  -3372     6744
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.17831  0.42227
Number of obs: 15216, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.97635    0.09891  30.092   <2e-16 ***
genderMALE  -0.21181    0.12842  -1.649   0.0991 .
```

Code Snippet 67 Anger Empathy versus Gender Model in R

From the coefficient estimates reference:

$$logit(\pi_{ij}) = 2.97635 \quad -0.21181\text{genderMALE} + \mu_{0j} \tag{6.12}$$

The $R^2_{binomial} =$ 0.002941 indicating that group3 describes a small portion of the angryEMPATHY variance. The ICC is 0.051413 indicating small correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 425$ on four degress of freedom.

There is a trend indicating Females are 1.1% more empathetic towards the anger of others than Males ($p<0.1$).

- Female pr(self-aware) = 95.1% (95% CI, 94.2%–96.0%);
- Male pr(self-aware) = 94% (95% CI, 92.5%–95.6%);

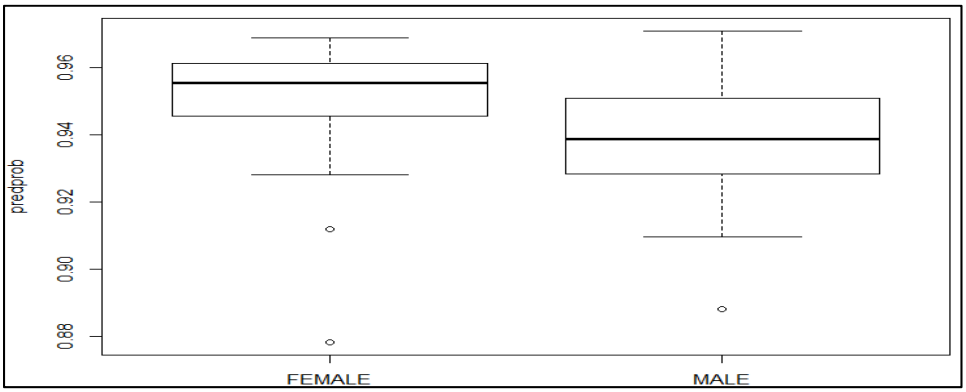


Figure 186 Predicted Probabilities of angryEMPATHY versus Gender

```
EMO <- EMO[(EMO$group3 == "GEN_POP"),] #subset to general population
Formula: angryEMPATHY ~ language + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 2277 2297  -1136     2271
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.18535  0.43052
Number of obs: 5094, groups: p, 44

Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.5882     0.1362  19.008   <2e-16 ***
languageFRENCH   0.4441     0.1882   2.359   0.0183 *
```

Code Snippet 68 Anger Empathy versus Language Model in R

From the coefficient estimates reference:

$$logit(\pi_{ij}) = 2.5882 \quad +0.4441\text{languageFRENCH} + \mu_{0j} \tag{6.13}$$

The $R^2_{binomial}$ = 0.013905 indicating that group3 describes a portion of the angryEMPATHY variance. The ICC is 0.053334 indicating small correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 4898$ on four degress of freedom.

There is a trend indicating French people are 2.4% more empathetic towards the anger of others than English People ($p<0.1$).

- English pr(self-aware) = 93.0% (95% CI, 91.0%–94.6%);
- French pr(self-aware) = 95.4% (95% CI, 93.4%–96.8%);

APPENDIX H

NEUTRAL REGRESSION ANALYSIS

Neutral (okCROWD)

Opioid addicts seem to be the most emotionally neutral.

Table 74 Frequency of Neutral (okCROWD) across Groups

	FALSE	TRUE
GP	54%	46%
AA	53%	47%
SUBX	48%	52%

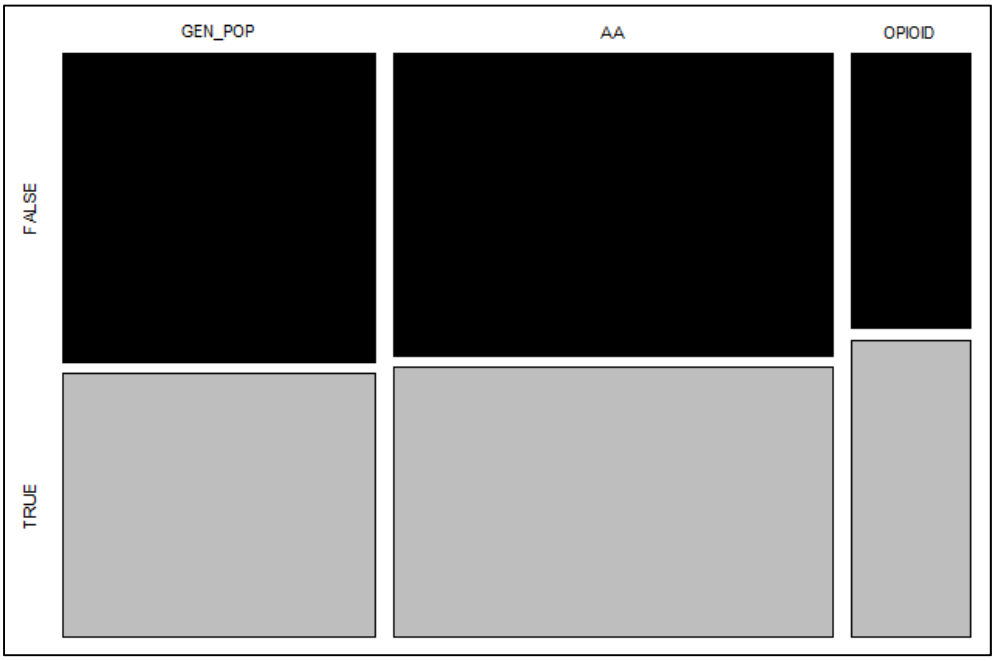


Figure 187 Frequency of Neutral (okCROWD) across Groups

```
Formula: okCROWD ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 9643 9657  -4820     9639
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.50143  0.70811
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.13424    0.07309  -1.836   0.0663 .
```

Code Snippet 69 Neutral Two-Level Null Model in R

There is a trend that the log-odds mean for neutral for an average participant ($\mu_{0j} = 0$) is -0.13424 which is a probability 46.6% (95% CI, 43.0%–50.3%). This is the only emotion where the average participant's emotional probability is not significant. The Log Likelihood test statistic between the one-level and two-level null model is 810 indicating strong evidence that between-participant variance is non-zero.

```
        Shapiro-Wilk normality test

data:  ranef(g1)$p[, 1]
W = 0.9862, p-value = 0.2176
```

Code Snippet 70 Neutral Shapiro-Wilk Normality Test in R

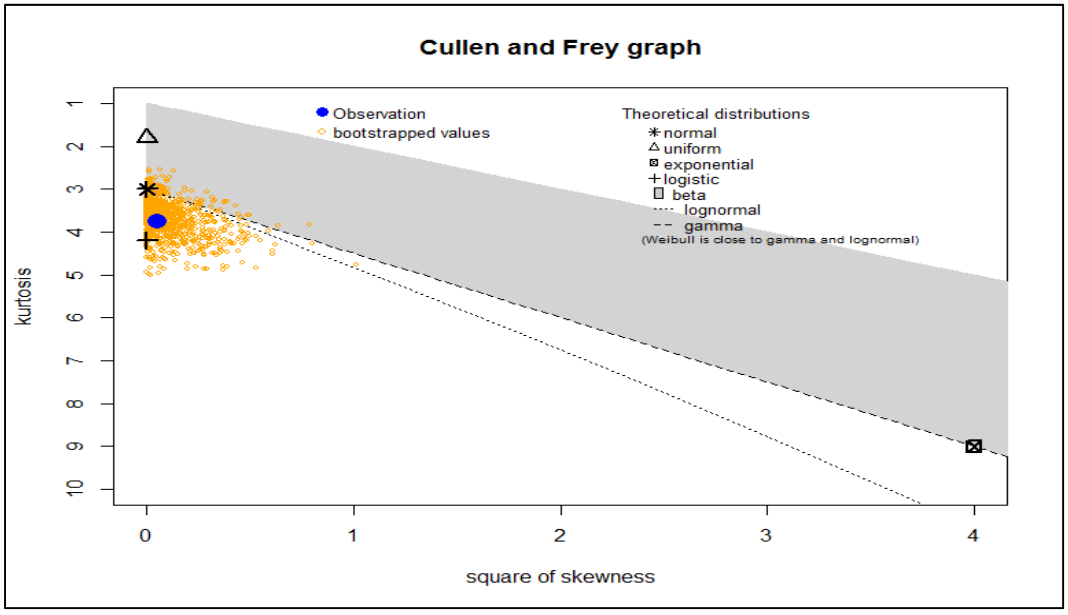


Figure 188 Cullen and Frey Distribution Graph of okCROWD

```
Formula: okCROWD ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 8449 8476  -4221     8441
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.50773  0.71255
Number of obs: 6650, groups: p, 112

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.18845    0.12442  -1.515   0.1299
group3AA      0.03841    0.18549   0.207   0.8360
group3OPIOID  0.35338    0.19479   1.814   0.0697 .
```

Code Snippet 71 Neutral versus Group Model in R

From the coefficient estimates with GEN_POP as reference:

$$logit(\pi_{ij}) = -0.18845 + 0.0384\text{group3AA} + 0.35338\text{group3OPIOID} + \mu_{0j} \quad (6.14)$$

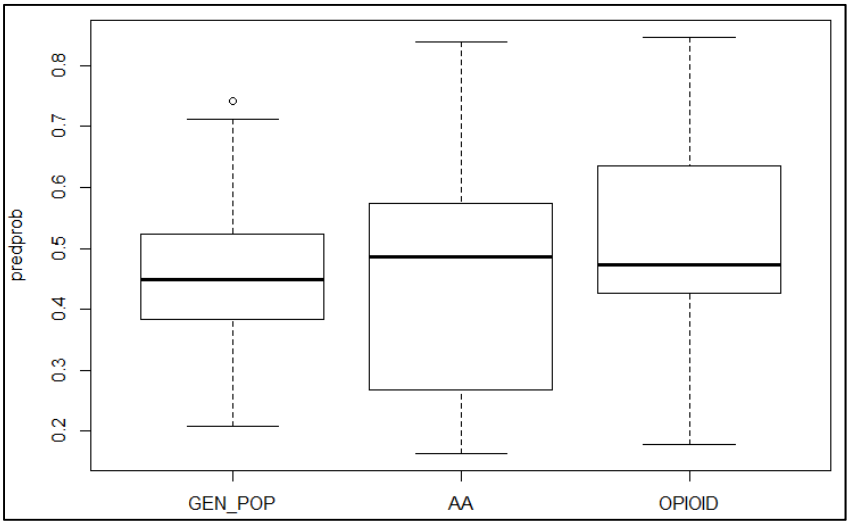


Figure 189 Predicted Probabilities of okCROWD versus Group

The $R^2_{binomial} = 0.003473$ indicating that group3 describes a small amount of the okCROWD variance. The ICC is 0.133697 indicating correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 1197$ on four degress of freedom.

There is a trend that Opioid Addicts are 8.8% more neutral than the General ($p<0.1$).

- General Population pr(okCROWD) = 45.3% (95% CI, 39.2%–51.5%);

- AA pr(okCROWD) = 46.2% (95% CI, 37.3%–55.5%);
- Opioid addict pr(okCROWD) = 54.1% (95% CI, 44.4%–63.5%);

```
Formula: okCROWD ~ gender + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 9272 9293  -4633     9266
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.49921  0.70655
Number of obs: 7290, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2358     0.1125  -2.097    0.036 *
genderMALE    0.2206     0.1503   1.468    0.142
```

Code Snippet 72 Neutral versus Gender Model in R

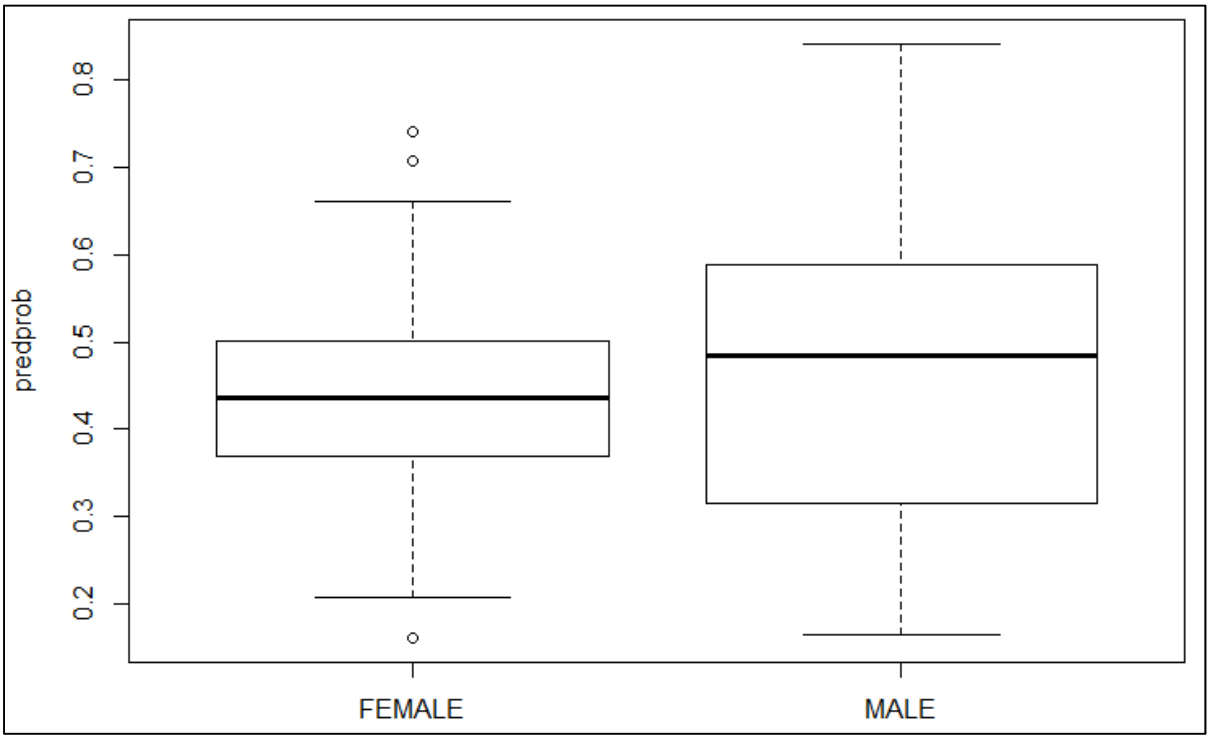


Figure 190 Predicted Probabilities of okCROWD versus Gender

There is no significant difference of neutrality across gender. However, Figure 190 indicates a much higher IQR for males.

```
>    EMO <- EMO[(EMO$group3 == "GEN_POP"),]
Formula: okCROWD ~ language + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 3172 3189  -1583     3166
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.30751  0.55453
Number of obs: 2387, groups: p, 44

Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.3850     0.1415  -2.720  0.00652 **
languageFRENCH   0.3556     0.1818   1.956  0.05052 .
```

Code Snippet 73 Neutral versus Language Model in R

From the coefficient estimates with GEN_POP as reference:

$$logit(\pi_{ij}) = -0.3850 + 0.3556\text{languageFRENCH} + \mu_{0j} \quad (6.15)$$

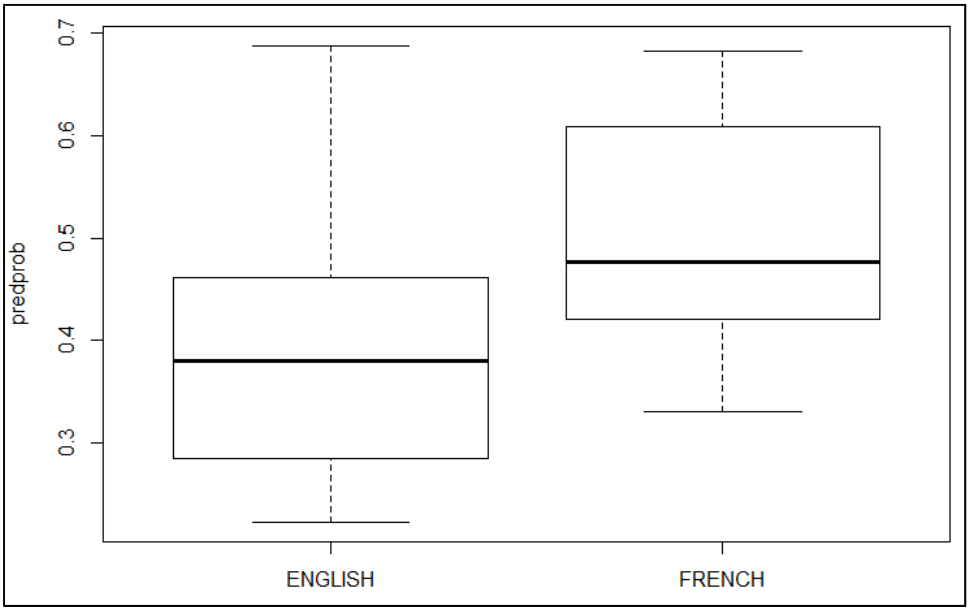


Figure 191 Predicted Probabilities of okCROWD versus Language

The $R^2_{binomial}$ = 0.008354 indicating that language describes a small amount of the okCROWD variance. The ICC is 0.085481 indicating correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 6473$ on four degress of freedom.

There is a trend that French are 8.8% more neutral than the English people (p<0.1).

- English pr(okCROWD) = 40.5% (95% CI, 33.9%–47.5%);
- French pr(okCROWD) = 49.3% (95% CI, 40.3%–58.3%);

Neutral Self-Awareness (okSELFAWARE)

Opioid addicts seem to be the least aware of their anger.

Table 75 Frequency of Neutral Self-Awareness across Groups

	FALSE	TRUE
GP	30%	70%
AA	30%	70%
SUBX	38%	62%

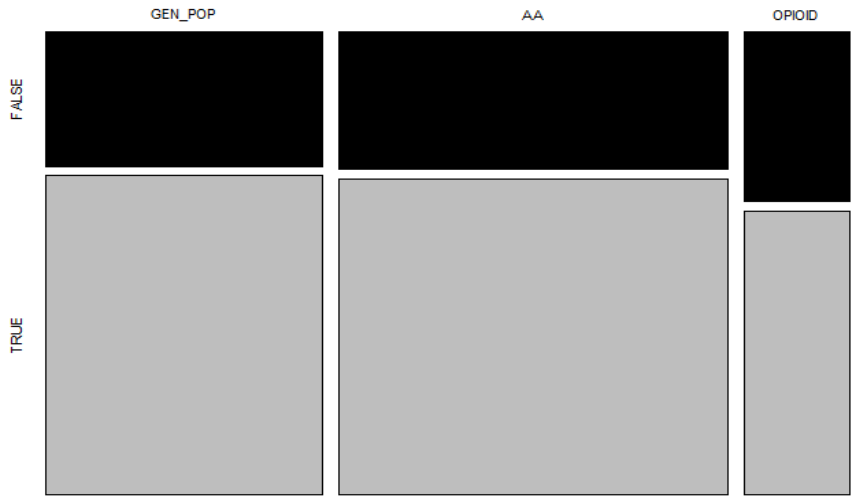


Figure 192 Frequency of Neutral Self-Awareness across Groups

```
Formula: okSELFAWARE ~ (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 9238 9252  -4617     9234
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.1612   0.40149
Number of obs: 7570, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.74382    0.04908   15.16   <2e-16 ***
```

Code Snippet 74 Neutral Self-Awareness Two-Level Null Model in R

The log-odds mean for self-awareness of anxiety for an average participant ($\mu_{0j} = 0$) is significant is estimated at 0.74382 that is a probability of 67.8% (95% CI, 65.6%–69.9%). The Log Likelihood test statistic between the one-level and two-level null model is 157; evidence that between-participant variance is non-zero.

```
Formula: okSELFAWARE ~ group3 + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 8107 8134  -4050     8099
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.14256  0.37757
Number of obs: 6650, groups: p, 112

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.88267    0.08083  10.920  < 2e-16 ***
group3AA     -0.17311    0.11690  -1.481  0.13865
group3OPIOID -0.34280    0.12990  -2.639  0.00832 **
```

Code Snippet 75 Neutral Self-Awareness versus Group Model in R

From the coefficient estimates with GEN_POP as reference:

$$logit(\pi_{ij}) = -0.3428 - 0.17311\text{group3AA} - 0.3428\text{group3OPIOID} + \mu_{0j} \quad (6.16)$$

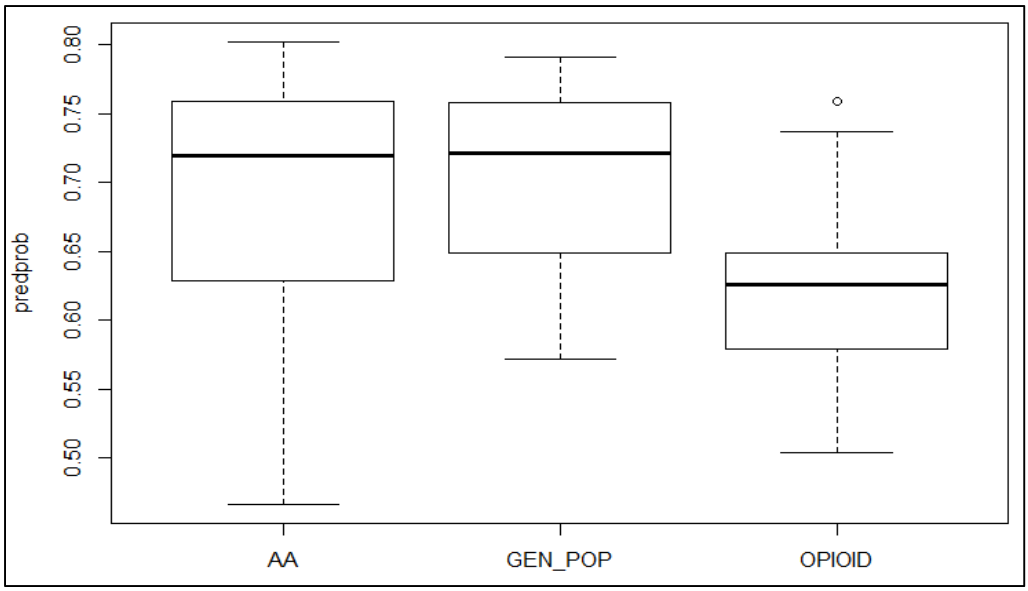


Figure 193 Predicted Probabilities of okSELFAWARE versus Group

The $R^2_{binomial}$ = 0.003825 indicating that language describes a small amount of the okSELFAWARE variance. The ICC is 0.041533 indicating some correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 1135$ on four degress of freedom.

Opioid Addicts are 7.5% less self-aware of their neutrality than the General Population ($p<0.05$).

- General Population pr(okSELFAWARE) = 70.7% (95% CI, 67.3%–74.0%);
- AA Member pr(okSELFAWARE) = 67.0% (95% CI, 61.7%–72.0%);
- Opioid Addict pr(okSELFAWARE) = 63.2% (95% CI, 56.6%–69.0%);

There are no differences in neutrality self-awareness across gender or language. However, Figure 194 indicates a much higher IQR for French people.

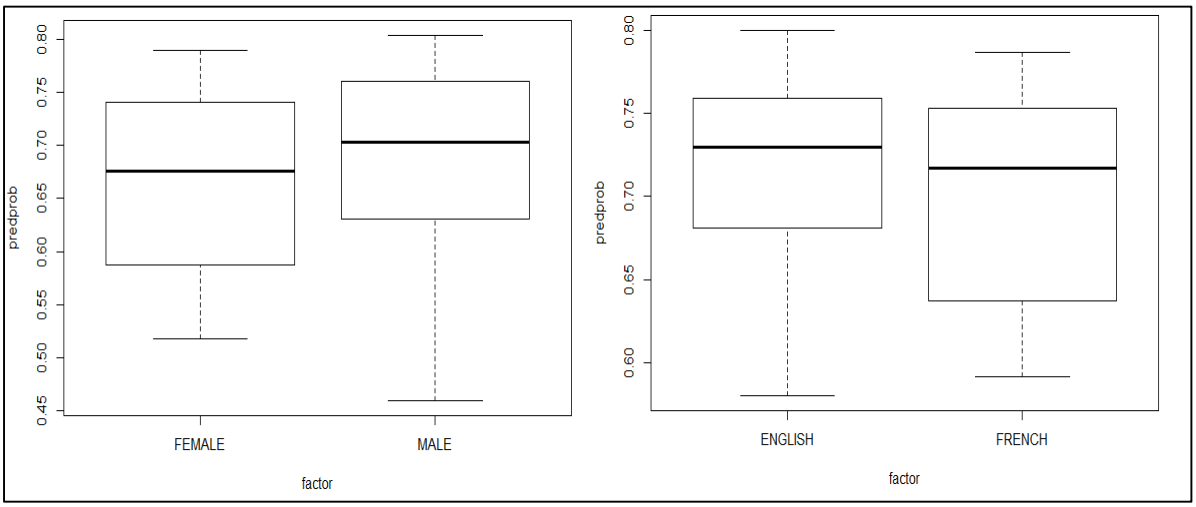


Figure 194 Predicted Prob of okSELFAWARE versus Gender and Language

Neutral Empathy (okEMPATHY)

Opioid addicts seem to emphasize others as neutral more than the General Population and AA members.

Table 76 Frequency of Empathy to Neutral (okEMPATHY) across Groups

	FALSE	TRUE
GP	26%	74%
AA	30%	70%
SUBX	23%	77%

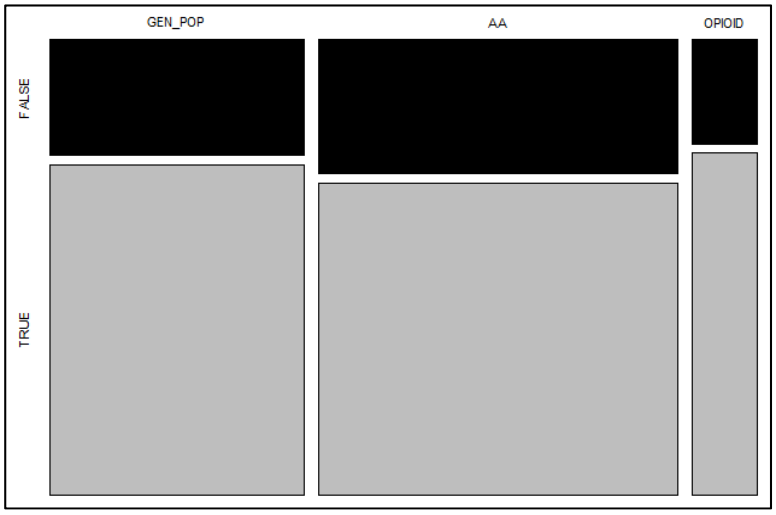


Figure 195 Frequency of Empathy to Neutral (okEMPATHY) across Groups

```
Formula: okEMPATHY ~ (1 | p)
   Data: EMO
   AIC   BIC logLik deviance
 18668 18683  -9332    18664
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.07375  0.27157
Number of obs: 16001, groups: p, 129

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.02857    0.03582   28.71   <2e-16 ***
```

Code Snippet 76 Empathy to Neutral Two-Level Null Model in R

The log-odds mean for self-awareness of ok for an average participant ($\mu_{0j} = 0$) is significant is estimated at 1.02857 which is a probability of 73.7% (95% CI, 72.3%–75.0%). The Log Likelihood test statistic between the one-level and two-level null model is 179; some evidence that between-participant variance is non-zero.

There are no differences across groups.

```
Formula: okEMPATHY ~ group3 + (1 | p)
   Data: EMO
   AIC   BIC logLik deviance
 15941 15971  -7967    15933
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.078548 0.28026
Number of obs: 13612, groups: p, 112

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.92789    0.06624  14.008   <2e-16 ***
group3GEN_POP 0.09875    0.08848   1.116   0.2644
group3OPIOID  0.25076    0.10970   2.286   0.0223 *
```

Code Snippet 77 Neutral Empathy versus Group Model in R

From the coefficient estimates with GEN_POP as reference:

$$logit(\pi_{ij}) = 0.92789 + 0.09875\text{group3AA} + 0.25076\text{group3OPIOID} + \mu_{0j} \quad (6.17)$$

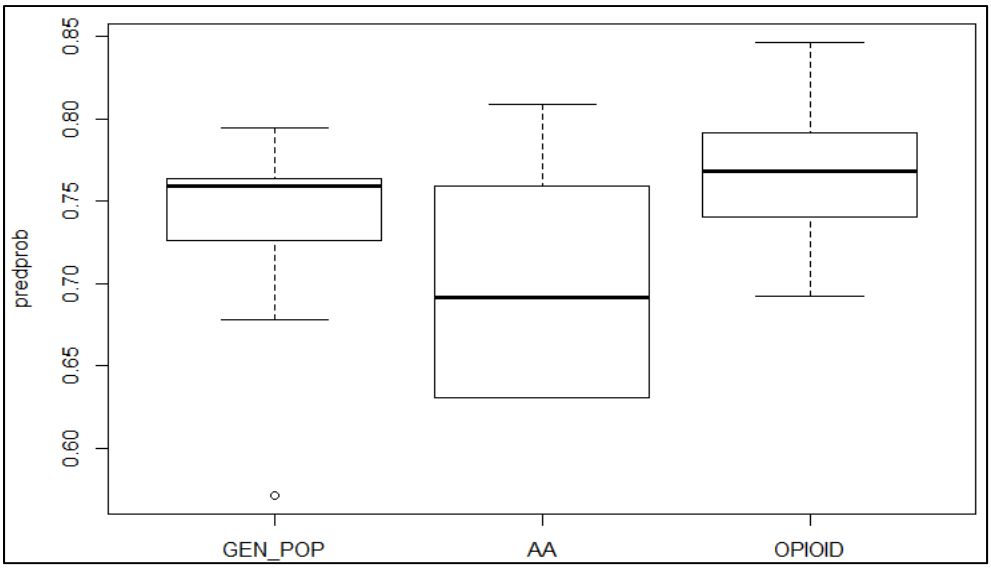


Figure 196 Predicted Probabilities of okEMPATHY versus Group

Figure 123 indicates AA members have a much wider IQR for neutral empathy that may be related to the length of their sobriety and/or comorbidity of mood disorders.

The $R^2_{binomial}$ = 0.023319 indicating that language describes some amount of the okEMPATHY variance. The ICC is 0.023319 indicating some correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 2730$ on four degress of freedom.

Opioid Addicts are 7.5% less self-aware of their neutrality than the General Population ($p<0.05$).

- General Population pr(okEMPATHY) = 73.6% (95% CI, 70.0%–76.9%);
- AA Member pr(o okEMPATHY) = 71.7% (95% CI, 68.9%–74.2%);
- Opioid Addict pr(o okEMPATHY) = 76.5% (95% CI, 72.3%–80.2%);

There are no differences across gender or language

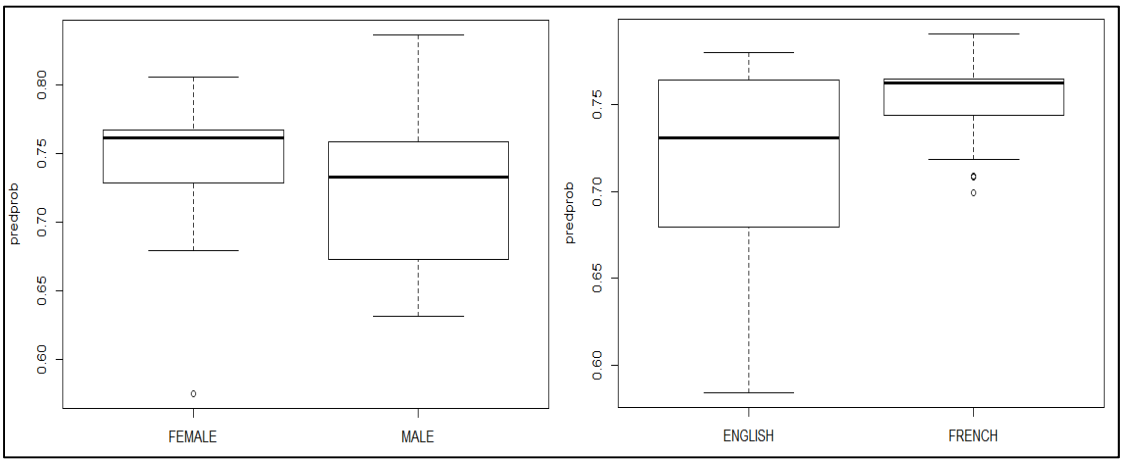


Figure 197 Predicted Prob of okEMPATHY versus Gender and Language

Figure 197 indicates English people and males have a much wider IQR for empathy of ok than French people and Females respectively.

```
Formula: okEMPATHY ~ gender + (1 | p)
   AIC   BIC logLik deviance
 17722 17745  -8858    17716
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.079476 0.28191
Number of obs: 15216, groups: p, 122

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.06271    0.05716  18.592   <2e-16 ***
genderMALE  -0.05146    0.07586  -0.678    0.497
```

Code Snippet 78 Neutral Empathy versus Gender Model in R

From the coefficient estimates reference:

$$logit(\pi_{ij}) = 2.97635 \quad -0.21181\text{genderMALE} + \mu_{0j} \tag{6.18}$$

The $R^2_{binomial}$ = 0.002941 indicating that group3 describes a small portion of the angryEMPATHY variance. The ICC is 0.051413 indicating small correlation within groups. The deviance from the null 2-level model is highly significant with $X^2 = 425$ on four degress of freedom.

There is a trend indicating Females are 1.1% more empathetic towards the anger of others than Males ($p<0.1$).

- Female pr(self-aware) = 95.1% (95% CI, 94.2%–96.0%);
- Male pr(self-aware) = 94% (95% CI, 92.5%–95.6%);

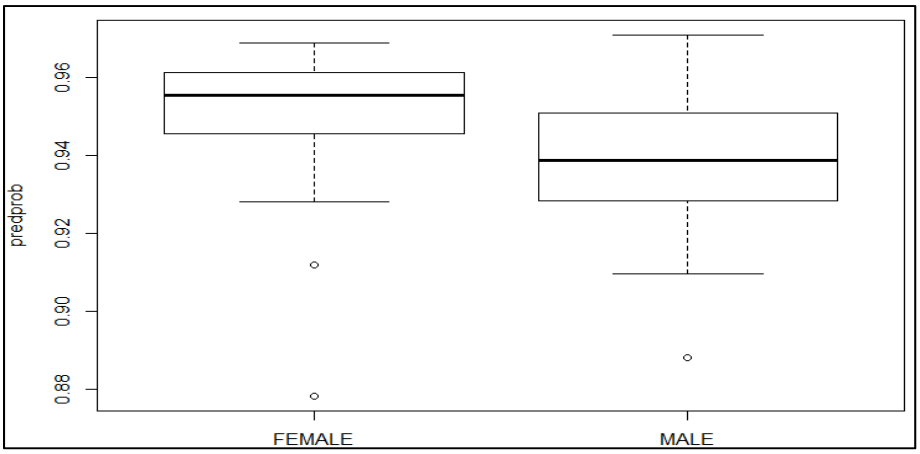


Figure 198 Predicted Probabilities of okEMPATHY versus Gender

```
EMO <- EMO[(EMO$group3 == "GEN_POP"),] #subset to general population
Formula: angryEMPATHY ~ language + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance
 5829 5848  -2911     5823
Random effects:
 Groups Name        Variance Std.Dev.
 p      (Intercept) 0.061484 0.24796
Number of obs: 5094, groups: p, 44

Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.95241    0.07803  12.205   <2e-16 ***
languageFRENCH  0.13875    0.10411   1.333    0.183
```

Code Snippet 79 Neutral Empathy versus Language Model in R

APPENDIX I

EXPRESSIVENESS ANALYSIS THROUGH LENGTH OF SPEECH

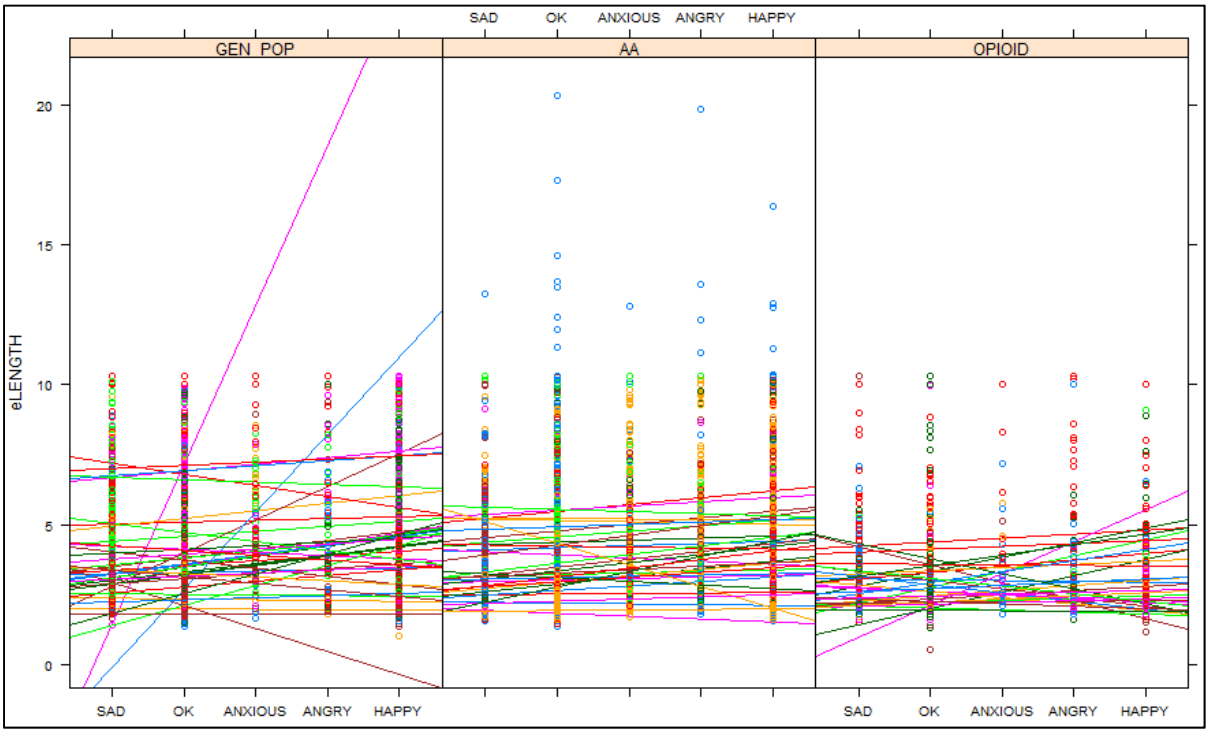


Figure 199 Length of Speech versus Emotion

Figure 199 is a visualization of speech length versus emotional truth for each trial group. The circles are data samples, and the lines represent the intercept and slope for each participant. Speech length may vary by emotion. The expression of Happiness seems to be consistently longer in duration than the other emotions. Opioid addicts seem to talk for shorter lengths than the General Population or AA members, and seem to have little variation of response length between emotions.

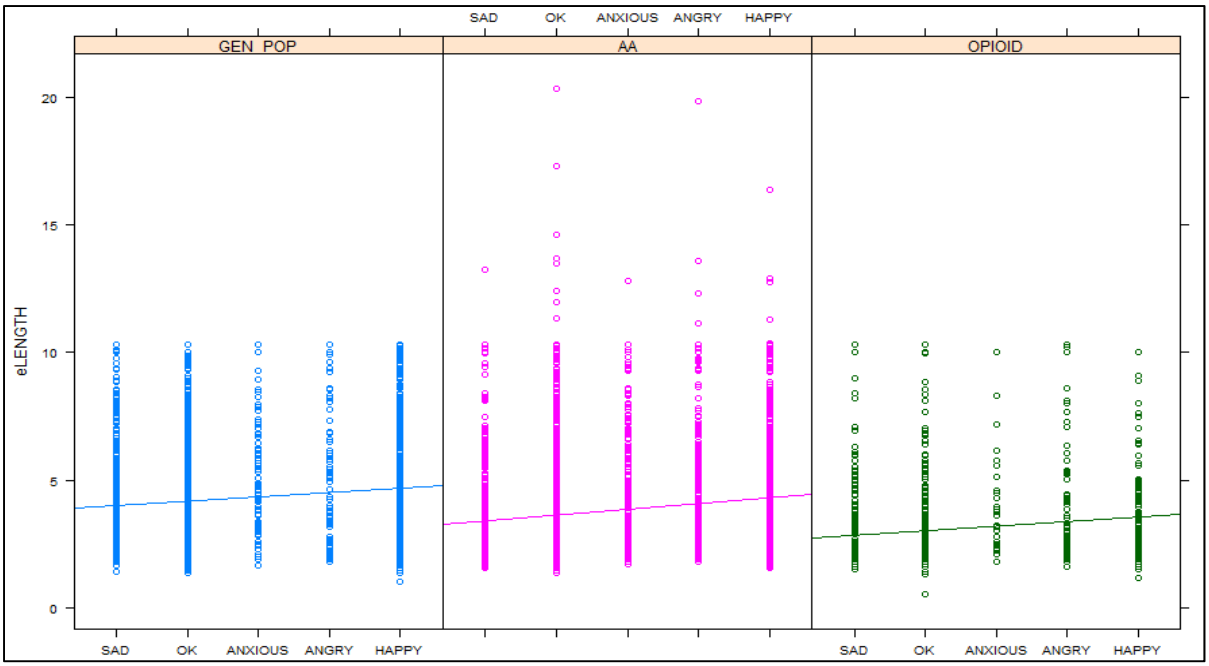


Figure 200 Length of Speech versus Emotion with Regression Lines

Figure 200 is a visualization of speech length versus emotional truth for each trial group with regression approximations.

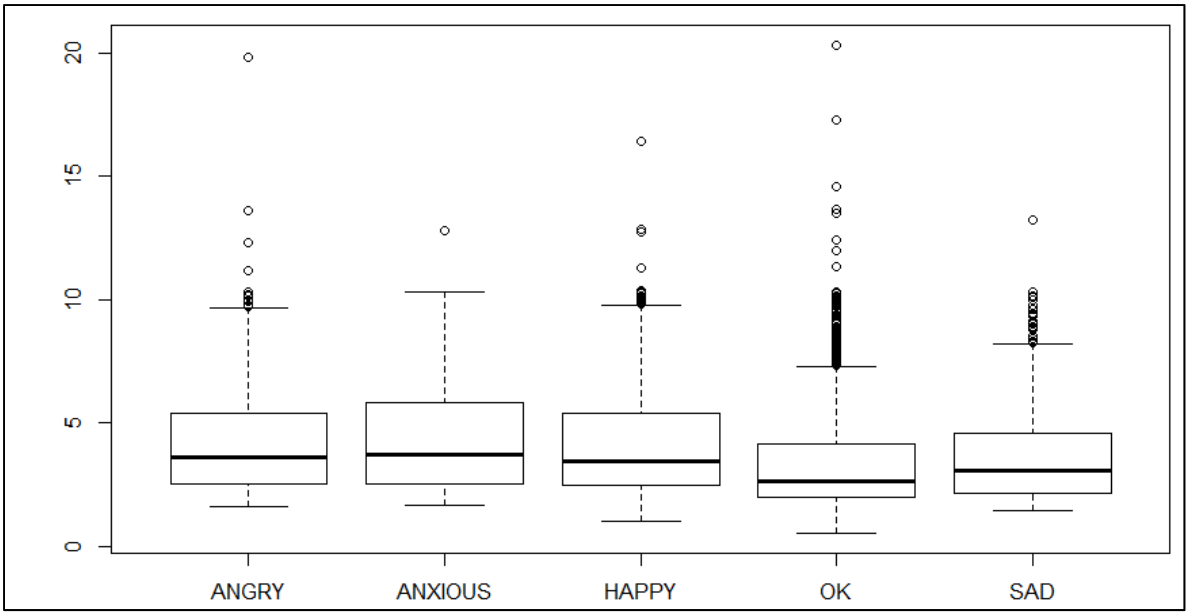


Figure 201 Predicted Probabilities of Length-of-Speech versus Emotion

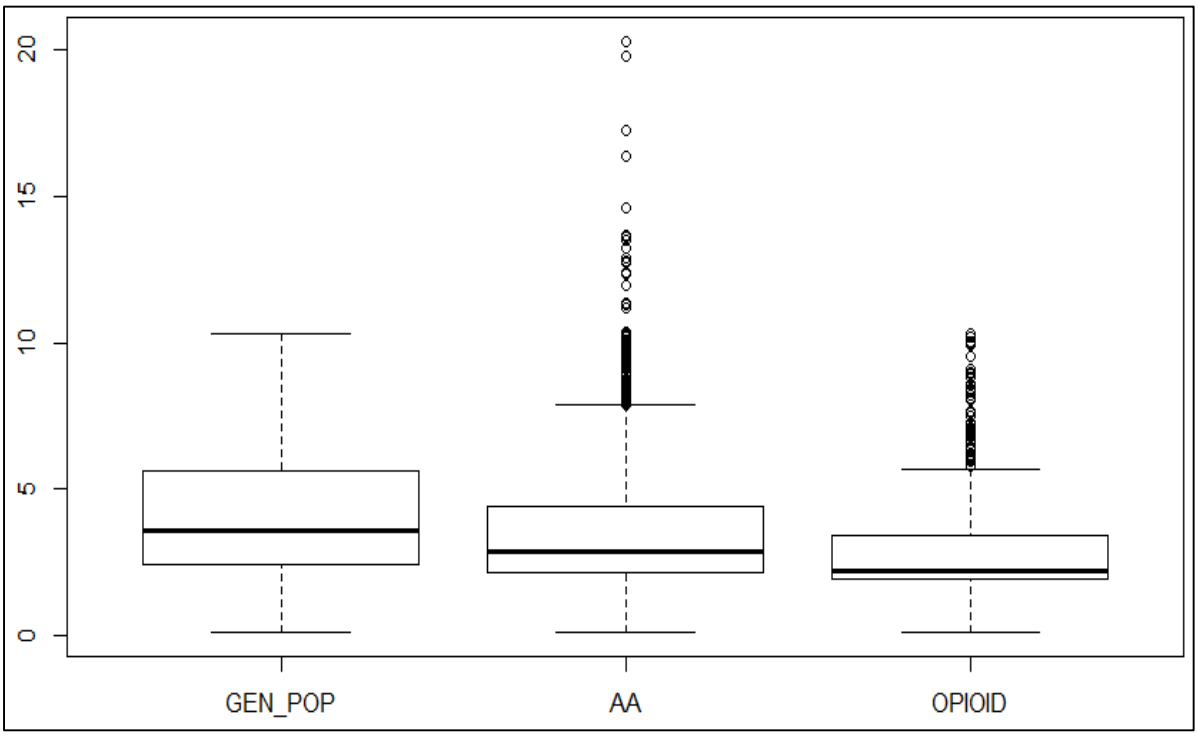


Figure 202 Predicted Probabilities of Length-of-Speech versus Group

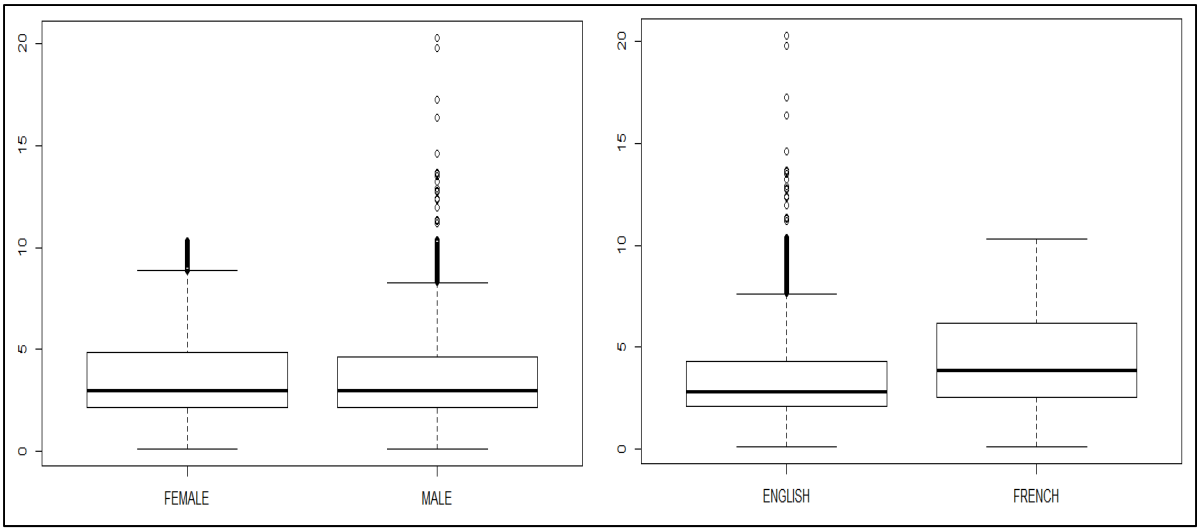


Figure 203 Predicted Prob of Length-of-Speech versus Gender and Language

```
> library(fitdistrplus)
> descdist(EMO$eLENGTH,boot=1000, obs.col="blue",boot.col="orange"
+ )
summary statistics
------
min:  0.09575   max:  20.3057
median:  2.97575
mean:  3.791971
estimated sd:  2.301589
estimated skewness:  1.539065
estimated kurtosis:  5.166731
```

Code Snippet 80 Length-of-Speech Distribution Statistics in R

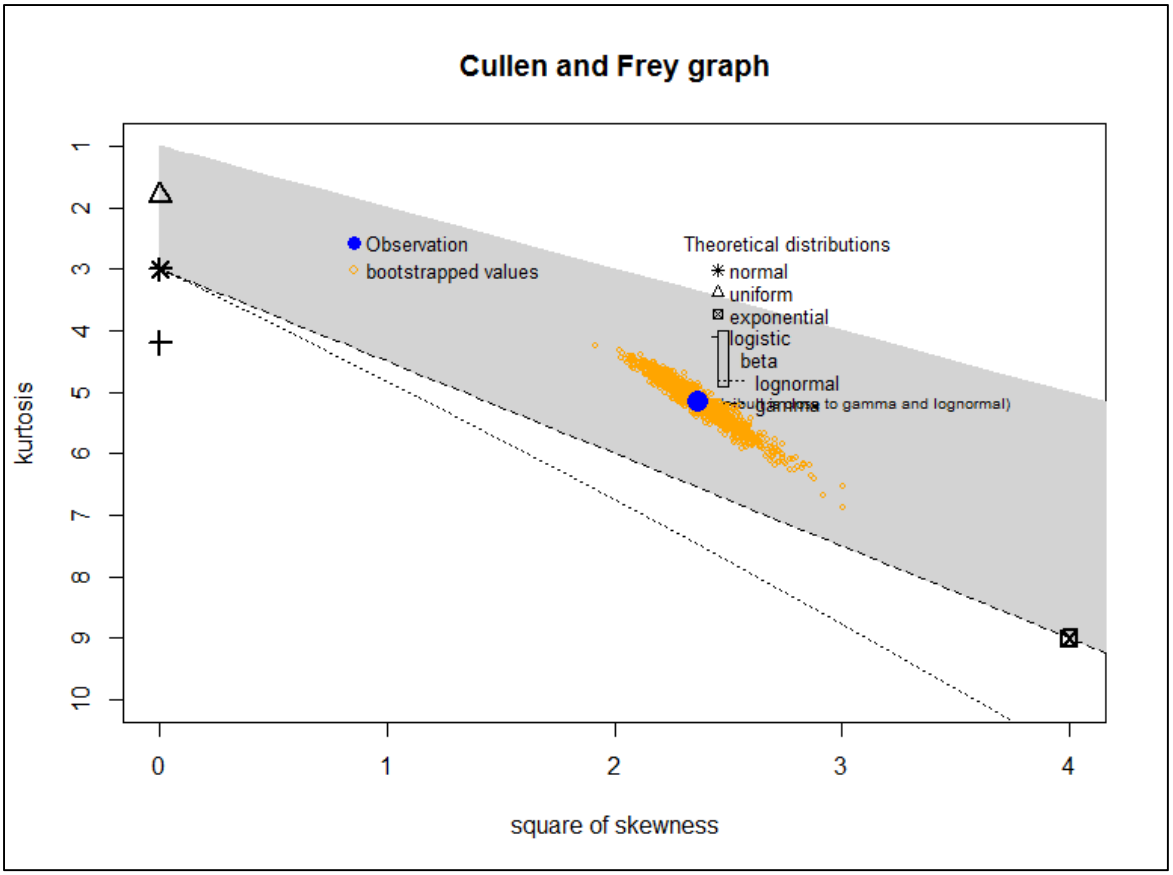


Figure 204 Cullen and Frey graph of log normalized Length of Speech

```
> nullmodel <- lmer(eLENGTH ~(1|p), data=EMO, REML=FALSE)
> #residuals
> THemoDIST(nullmodel)

        Shapiro-Wilk normality test

data:  lme4::ranef(g1)$p[, 1]
W = 0.9331, p-value = 7.597e-06
```

Code Snippet 81 Shapiro-Wilk Normality Test of Length-of-Speech in R

Length-of-Speech is not normally distributed.

```
> library(mosaic)
> #histogram of speechsecs with a gamma distr curve
> xhistogram(~eLENGTH,data=EMO, fit='gamma', groups = eLENGTH > 1 )
```

Code Snippet 82 Histogram to Investigate Length-of-Speech Distribution in R

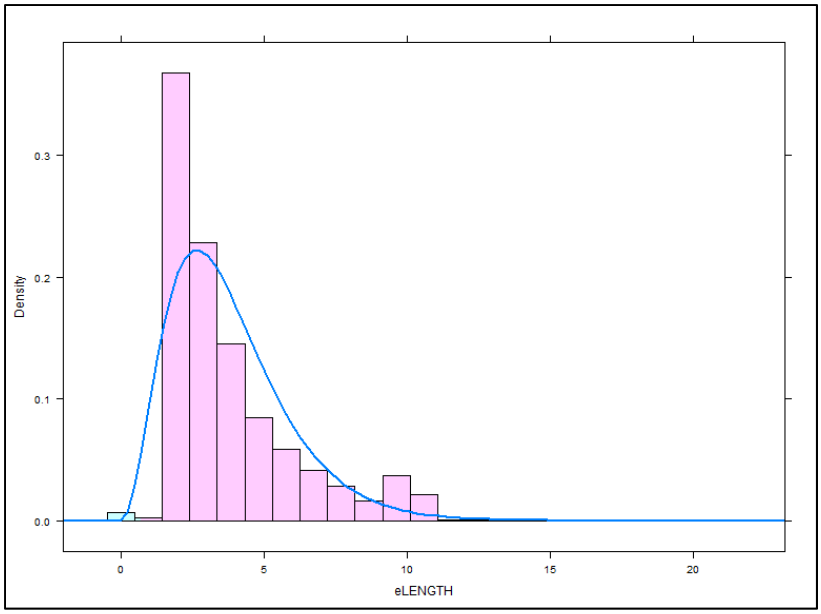


Figure 205 Gamma Distribution of Length of Speech

```
> # log normalize eLength
> EMO$eLENLOG = log(EMO$eLENGTH)
```

Code Snippet 83 Log-Normalization of Length-of-Speech in R

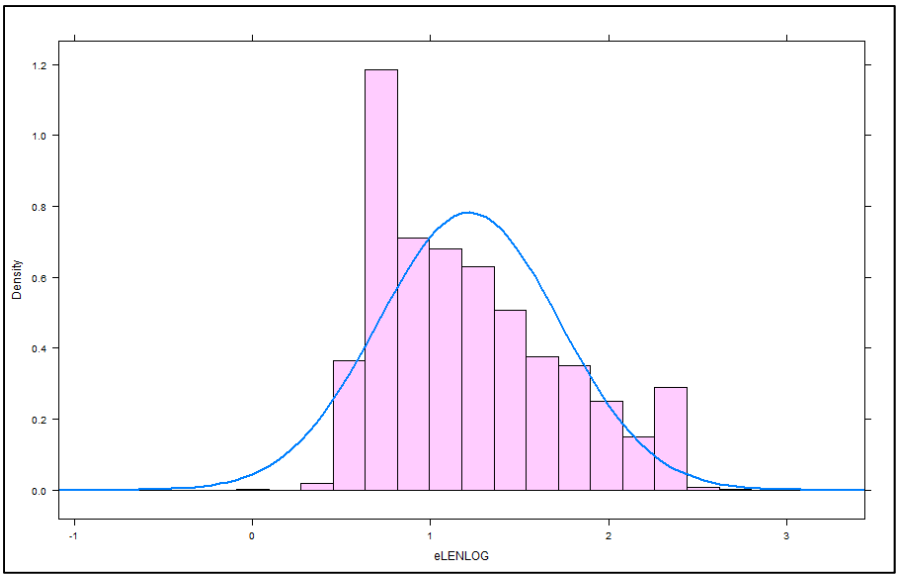


Figure 206 Log-Normalized Length of Speech Distribution

The log-normalized Length-of-Speech is a better fit; almost normal.

```
        Shapiro-Wilk normality test

data:  lme4::ranef(g1)$p[, 1]
W = 0.9784, p-value = 0.03742
```

Code Snippet 84 Shapiro-Wilk Normality Test of Log-Norm Length-of-Speech in R

The null hypothesis (p –value < 0.05) is not rejected. The log-normalized Length-of-Speech is a better fit.

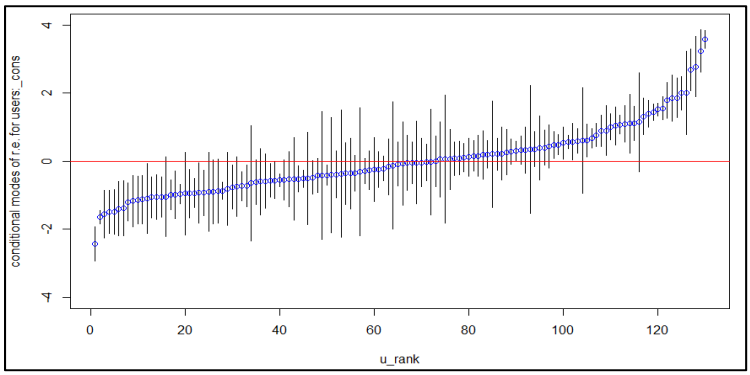


Figure 207 Estimates of Log-Norm Residuals $\hat{\mu}_{0j}$ for each $participant_j$.

```
> nullmodel <- lmer(eLENLOG ~(1|p), data=EMO, REML=FALSE)
> fit <- lm(eLENLOG ~ 1, data = EMO)
> print(2*(logLik(nullmodel)-logLik(fit)))
'log Lik.' 2901.626 (df=3)
```

Code Snippet 85 Two-and One-Level Log-Norm Model Comparison

The Log Likelihood test statistic is 2901 indicating strong evidence that between-participant variance of Length-of-Speech is non-zero.

```
Formula: eLENLOG ~ group3 + (1 | p)
   Data: EMO
   AIC   BIC logLik deviance REMLdev
 10004 10039  -4997     9994   10007
Random effects:
 Groups   Name        Variance Std.Dev.
 p        (Intercept) 0.087806 0.29632
 Residual             0.219088 0.46807
Number of obs: 7342, groups: p, 113

Fixed effects:
            Estimate Std. Error t value
(Intercept)   1.24093    0.04744  26.159
group3AA     -0.04422    0.07162  -0.617
group3OPIOID -0.37016    0.07212  -5.133

Correlation of Fixed Effects:
            (Intr) grp3AA
group3AA    -0.662
grop3OPIOID -0.658  0.436

> anova(g.group3,nullmodel)
Data: EMO
Models:
nullmodel: eLENLOG ~ (1 | p)
g.group3: eLENLOG ~ group3 + (1 | p)
          Df   AIC   BIC  logLik  Chisq Chi Df Pr(>Chisq)
nullmodel  3 11343 11364 -5668.5
g.group3   5 10004 10039 -4997.2 1342.7      2  < 2.2e-16 ***

> print(2*(logLik(g.group3)-logLik(nullmodel)))
'log Lik.' 1342.673 (df=5)> wald(g.group3)

> THwaldCI2exp(g.group3,"Group","eLENLOG","EXPRESS","eCROWD")

                 coef    lower    upper          pval effect trend
(Intercept)  3.458838 3.145765 3.803070 7.761745e-151   TRUE FALSE
group3AA     3.309224 2.867579 3.818888  5.369778e-01  FALSE FALSE
group3OPIOID 2.388764 2.067924 2.759383  2.854156e-07   TRUE FALSE
```

Code Snippet 86 Log-Normalized Length-of-Speech versus Group (GP ref) in R

There is an effect with the SUBX group (group3OPIOID). This is to say that SUBX patients are less expressive than the GP.

```
> EMO$group3 <- relevel(EMO$group3,ref="AA")
> g.group3AA <- lmer(eLENLOG ~group3+(1|p), data=EMO, REML=FALSE)
> THwaldCI2exp(g.group3AA,"Group","eLENLOG","EXPRESS","eCROWD")
                  coef    lower    upper          pval effect trend
(Intercept)   3.309224 2.972467 3.684133 3.574437e-110   TRUE FALSE
group3GEN_POP 3.458838 2.997226 3.991545  5.369778e-01  FALSE FALSE
group3OPIOID  2.388764 2.050473 2.782868  1.965016e-05   TRUE FALSE
```

Code Snippet 87 Log-Normalized Length-of-Speech versus Group (AA ref) in R

We relevel to see if there is in an effect between AA and SUBX. There is an effect with the SUBX group (group3OPIOID). This is to say that SUBX patients are less expressive than the AA.

```
Formula: eLENLOG ~ eCROWD + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance REMLdev
 8999 9047  -4492     8985    9015
Random effects:
 Groups   Name        Variance Std.Dev.
 p        (Intercept) 0.066241 0.25737
 Residual             0.183684 0.42858
Number of obs: 7570, groups: p, 129

> anova(g.eCROWD,nullmodel)
Data: EMO
Models:
nullmodel: eLENLOG ~ (1 | p)
g.eCROWD: eLENLOG ~ eCROWD + (1 | p)
          Df     AIC     BIC  logLik  Chisq Chi Df Pr(>Chisq)
nullmodel  3 11343.0 11364.1 -5668.5
g.eCROWD   7  8998.6  9047.1 -4492.3 2352.4      4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

'log Lik.' 2352.434 (df=7)

0.056994  0.015915 Inf  3.581039 0.00034   0.025800   0.088188

> THwaldCI2exp(g.eCROWD,"MALE","eLENLOG","EXPRESS","eCROWD")

                  coef    lower    upper         pval effect trend
(Intercept)   2.971771 2.826486 3.124524 0.000000e+00   TRUE FALSE
eCROWDHAPPY   3.358802 3.273292 3.446546 2.207338e-21   TRUE FALSE
eCROWDSAD     3.146064 3.047499 3.247816 3.422300e-04   TRUE FALSE
eCROWDANGRY   3.414152 3.276826 3.557233 1.375864e-11   TRUE FALSE
eCROWDANXIOUS 3.597344 3.440407 3.761439 1.075222e-17   TRUE FALSE
```

Code Snippet 88 Log-Normalized Length versus Emotional Truth Model in R

There are statistically significant effects of Log-Normalized Length-of-Speech with Neutral compared to Anger, Anxiety, Happy and Sad. This is to say, that the Length-of-Speech varies with emotion.

We now investigate differences of cross effects of groups and emotions.

```
> g.eg <- lmer(eLENLOG ~eCROWD*group3+(1|p), data=EMO, REML=FALSE)
> THwaldCI2exp(g.eg,"MALE","eLENLOG","EXPRESS","eCROWD")

                                coef    lower    upper          pval effect trend
(Intercept)                 3.228943 2.954317 3.529097 2.725489e-153   TRUE FALSE
eCROWDANGRY                 3.386781 3.200952 3.583398  9.075525e-02  FALSE  TRUE
eCROWDANXIOUS               3.637619 3.422174 3.866628  9.463711e-05   TRUE FALSE
eCROWDHAPPY                 3.531944 3.394709 3.674728  5.996862e-06   TRUE FALSE
eCROWDSAD                   3.127758 2.968815 3.295210  2.221061e-01  FALSE FALSE
group3GEN_POP               3.173170 2.815539 3.576228  7.707290e-01  FALSE FALSE
group3OPIOID                2.422389 2.127519 2.758127  9.491619e-06   TRUE FALSE
eCROWDANGRY:group3GEN_POP   3.914329 3.508247 4.367416  4.399187e-04   TRUE FALSE
eCROWDANXIOUS:group3GEN_POP 3.854489 3.435546 4.324520  2.083672e-03   TRUE FALSE
eCROWDHAPPY:group3GEN_POP   3.454913 3.257922 3.663814  2.120098e-02   TRUE FALSE
eCROWDSAD:group3GEN_POP     3.904305 3.620224 4.210677  4.950965e-07   TRUE FALSE
eCROWDANGRY:group3OPIOID    3.747641 3.318066 4.232832  1.439360e-02   TRUE FALSE
eCROWDANXIOUS:group3OPIOID  4.008367 3.374741 4.760958  1.195945e-02   TRUE FALSE
eCROWDHAPPY:group3OPIOID    3.295787 3.002139 3.618156  6.605623e-01  FALSE FALSE
eCROWDSAD:group3OPIOID      3.772414 3.419728 4.161474  1.525871e-03   TRUE FALSE
```

Code Snippet 89 Log-Norm Length across effects of group and emotion (ref GP) in R

SUBX patients are significantly lower than GP across all emotions.

```
> EMO$group3 <- relevel(EMO$group3,ref="AA")
> g.eg <- lmer(eLENLOG ~eCROWD*group3+(1|p), data=EMO, REML=FALSE)
> THwaldCI2exp(g.eg,"CROSS","eLENLOG","EXPRESS","eCROWD")

                                coef    lower    upper          pval effect trend
(Intercept)                 3.173170 2.929245 3.437407 2.592025e-183   TRUE FALSE
eCROWDANGRY                 4.034754 3.673241 4.431846  3.088406e-07   TRUE FALSE
eCROWDANXIOUS               4.267334 3.870797 4.704495  1.237672e-09   TRUE FALSE
eCROWDHAPPY                 3.713843 3.556420 3.878235  3.727239e-13   TRUE FALSE
eCROWDSAD                   3.716631 3.518959 3.925407  7.244004e-09   TRUE FALSE
group3AA                    3.228943 2.865026 3.639085  7.707290e-01  FALSE FALSE
group3OPIOID                2.422389 2.140163 2.741832  1.306810e-05   TRUE FALSE
eCROWDANGRY:group3AA        2.617558 2.346006 2.920542  4.399187e-04   TRUE FALSE
eCROWDANXIOUS:group3AA      2.658195 2.369276 2.982345  2.083672e-03   TRUE FALSE
eCROWDHAPPY:group3AA        2.965627 2.796535 3.144944  2.120098e-02   TRUE FALSE
eCROWDSAD:group3AA          2.624279 2.433334 2.830207  4.950965e-07   TRUE FALSE
eCROWDANGRY:group3OPIOID    3.038044 2.633237 3.505081  5.427665e-01  FALSE FALSE
eCROWDANXIOUS:group3OPIOID  3.299848 2.733962 3.982863  6.772905e-01  FALSE FALSE
eCROWDHAPPY:group3OPIOID    3.027020 2.752850 3.328497  3.205700e-01  FALSE FALSE
eCROWDSAD:group3OPIOID      3.065978 2.775587 3.386751  4.897454e-01  FALSE FALSE
```

Code Snippet 90 Log-Norm Length across effects of group and emotion (ref AA) in R

SUBX patients are significantly lower than AA across all emotions except Happiness.

```
Formula: eLENLOG ~ gender + (1 | p)
   Data: EMO
   AIC   BIC logLik deviance REMLdev
 10917 10945  -5455    10909   10918
Random effects:
 Groups   Name        Variance Std.Dev.
 p        (Intercept) 0.10816  0.32888
 Residual             0.21573  0.46446
Number of obs: 8089, groups: p, 123

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.10000    0.04652  23.646
genderMALE   0.04158    0.06290   0.661

Correlation of Fixed Effects:
           (Intr)
genderMALE -0.740
> wald(g.gender)
 numDF denDF F.value p.value
     2   Inf 643.131 <.00001

Coefficients  Estimate Std.Error  DF   t-value p-value Lower 0.95 Upper 0.95
  (Intercept) 1.100004  0.046519 Inf 23.646486 <.00001   1.008829   1.191179
  genderMALE  0.041580  0.062899 Inf  0.661051 0.50858  -0.081701   0.164860
```

Code Snippet 91 Log-Normalized Length-of-Speech versus Gender Model in R

There is no effect on gender.

```
> EMO <- EMO[(EMO$group3 == "GEN_POP"),]
> EMO <- EMO[!is.na(EMO$language),]
> g.lang <- lmer(eLENLOG ~language+(1|p), data=EMO, REML=FALSE)
  Linear mixed model fit by maximum likelihood
Formula: eLENLOG ~ language + (1 | p)
   Data: EMO
  AIC  BIC logLik deviance REMLdev
 3182 3206  -1587     3174    3182
Random effects:
 Groups   Name        Variance Std.Dev.
 p        (Intercept) 0.077093 0.27766
 Residual             0.204976 0.45274
Number of obs: 2440, groups: p, 44
Fixed effects:
              Estimate Std. Error t value
(Intercept)     1.18744    0.05570  21.317
languageFRENCH  0.09881    0.06147   1.607
Correlation of Fixed Effects:
            (Intr)
langgFRENCH -0.600
                  coef    lower    upper          pval effect trend
(Intercept)    3.278675 2.933022 3.665064 7.848627e-101  FALSE FALSE
languageFRENCH 3.619187 3.200507 4.092637  1.079566e-01  FALSE FALSE
```

Code Snippet 92 Log-Normalized Length-of-Speech versus Language Model in R

There is a no effect of length of speech across language, within the general population.

APPENDIX J

EXPRESSIVENESS ANALYSIS THROUGH CONFUSABILITY

Figure 138 is a visualization of confidence score versus emotional truth for each trial group. The circles are data samples, and the lines represent the intercept and slope for each participant.

Opioid-Suboxone patients seem to have consistently lower confidence scores. AA members seem to have the highest confidence score for happiness.

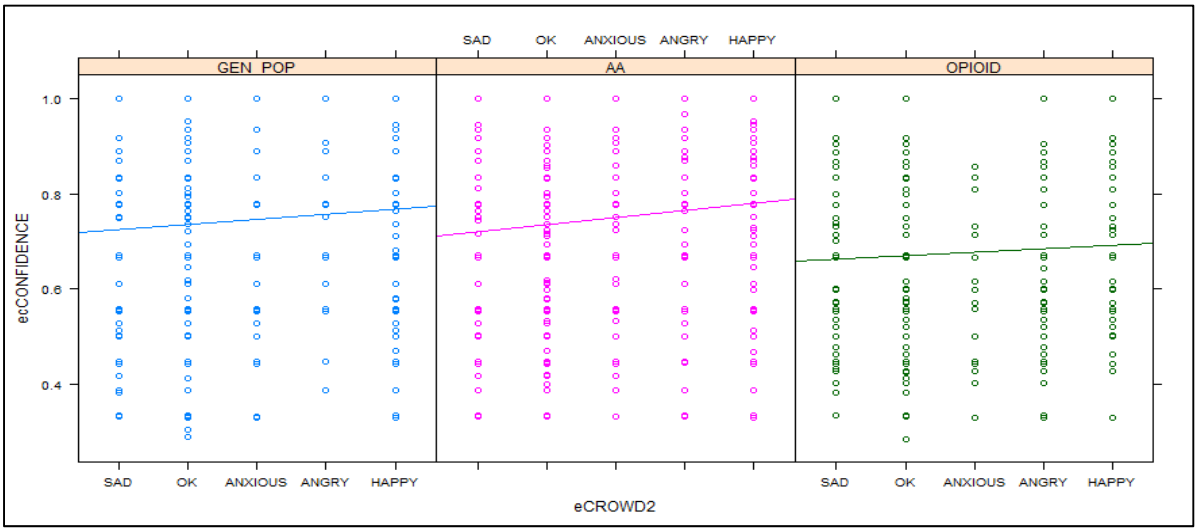


Figure 208 Confusability versus Emotion with Regression Lines

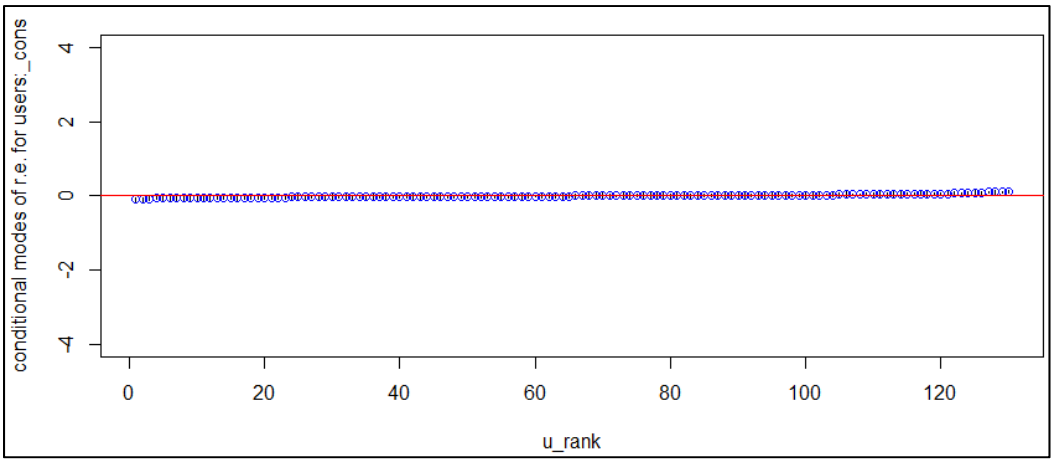


Figure 209 Estimates of Confusability residuals $\hat{\mu}_{0j}$ for each $participant_j$

Confusability residuals are homoscedastic. There is little variation in confusability across participants.

```
        Shapiro-Wilk normality test

data:  lme4::ranef(g1)$p[, 1]
W = 0.9739, p-value = 0.01364
```

Code Snippet 93 Normality Test of Confusability Residuals in R

The p=0.01364 < 0.05 for Shapiro-Wilk. The residuals are not normally distributed.

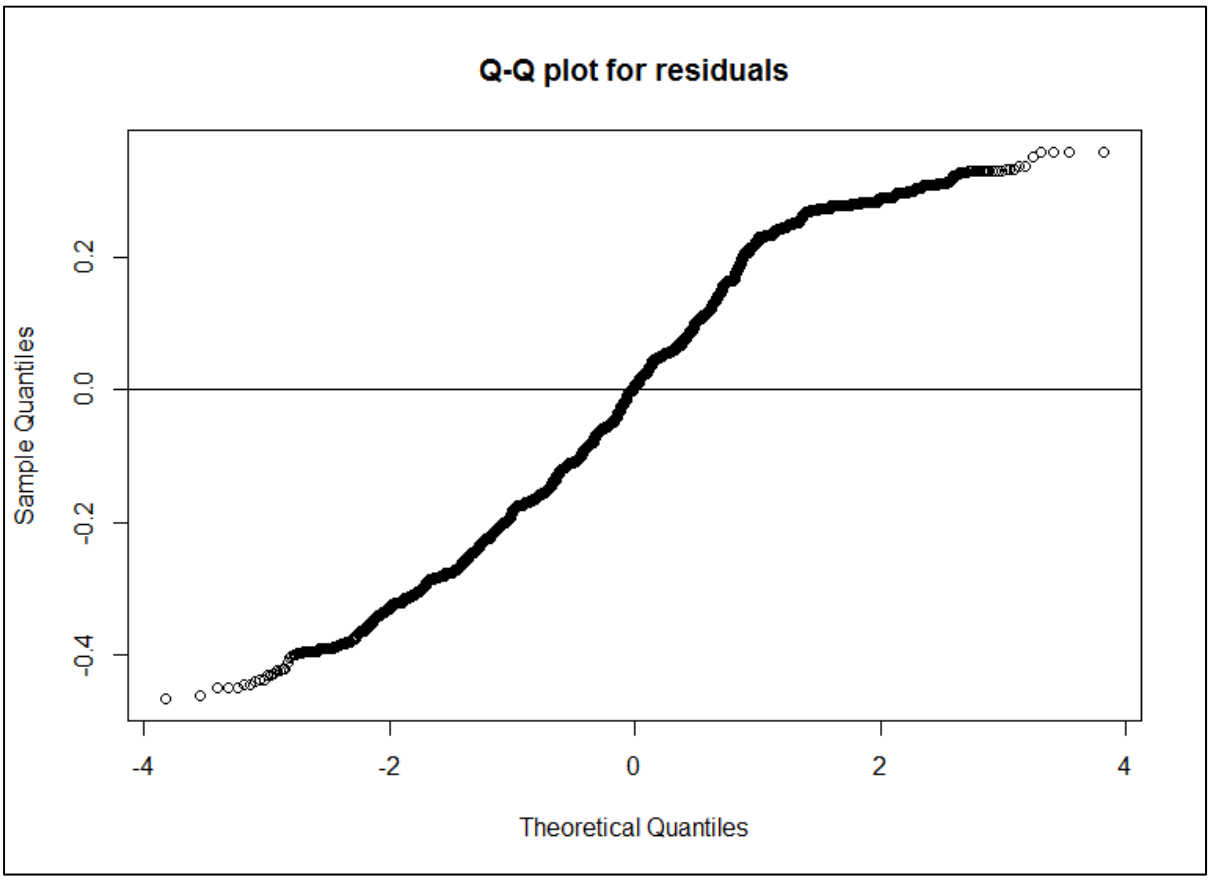


Figure 210 Q-Q Plot depicting Lack of Normality in R

The Q-Q plot is not linear.

```
> descdist(EMO$ecCONFIDENCE,boot=1000, obs.col="blue",boot.col="orange")
summary statistics
------
min:  0.286   max:  1
median:  0.777
mean:  0.7636728
estimated sd:  0.1911443
estimated skewness:  -0.2328627
estimated kurtosis:  1.90543
```

Code Snippet 94 Confidence Score (Confusability) Distribution Statistics in R

The negative skew of -0.23 indicates that the tail on the left side of the probability density function is longer than the right side and the bulk of the values (possibly including the median) lie to the right of the mean.

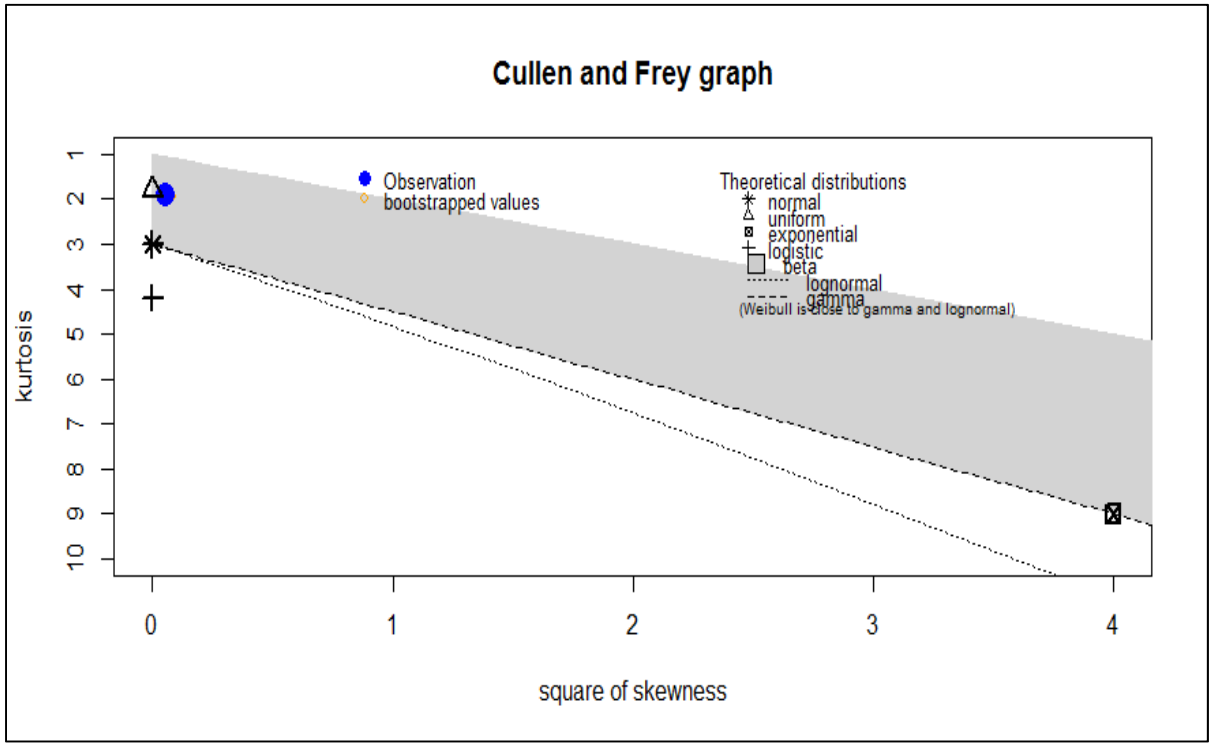


Figure 211 Cullen and Frey Graph of Confusability

The distribution is somewhere with beta (lognormal or gamma)

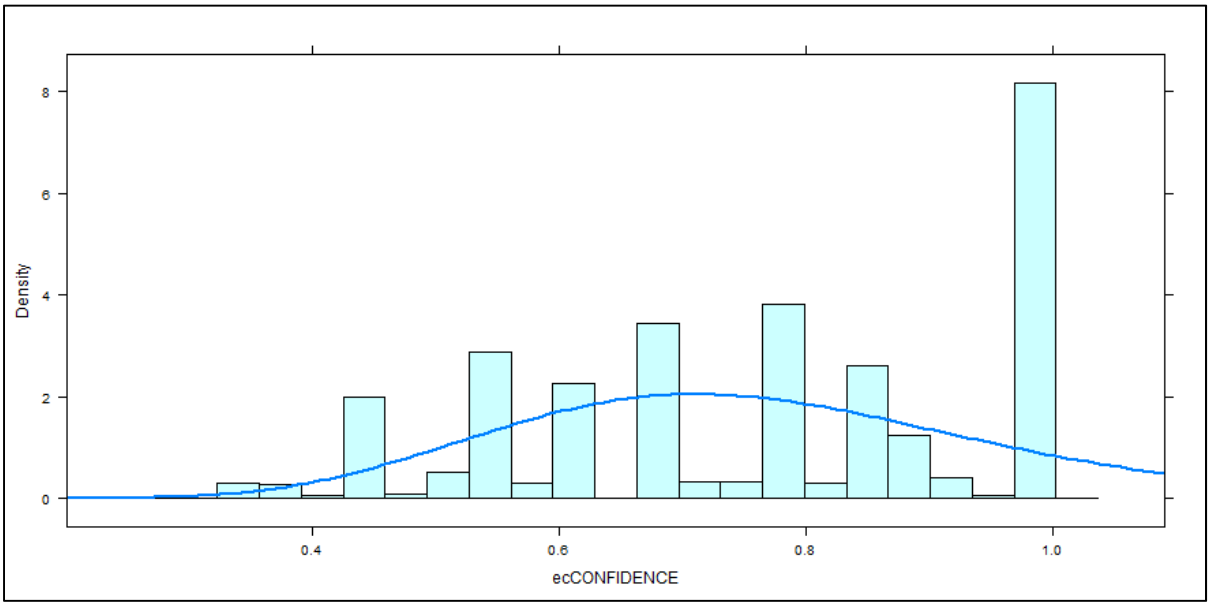


Figure 212 Confusability Distribution

We will attempt to normalize CONFIDENCE score.

```
> #attempt to normalize
> pow <- powerTransform(EMO$ecCONFIDENCE)
> pow$lambda

EMO$ecCONFIDENCE
        1.276283
> transformed_dv <- EMO$ecCONFIDENCE^(pow$lambda)

> descdist(transformed_dv,boot=1000, obs.col="blue",boot.col="orange")
summary statistics
------
min:  0.2023809   max:  1
median:  0.72468
mean:  0.7169064
estimated sd:  0.2237387
estimated skewness:  -0.1388364
estimated kurtosis:  1.803554
```

Code Snippet 95 Attempt to Normalize the Confidence Score in R

Skewness was improved to -0.13, but it is still negative

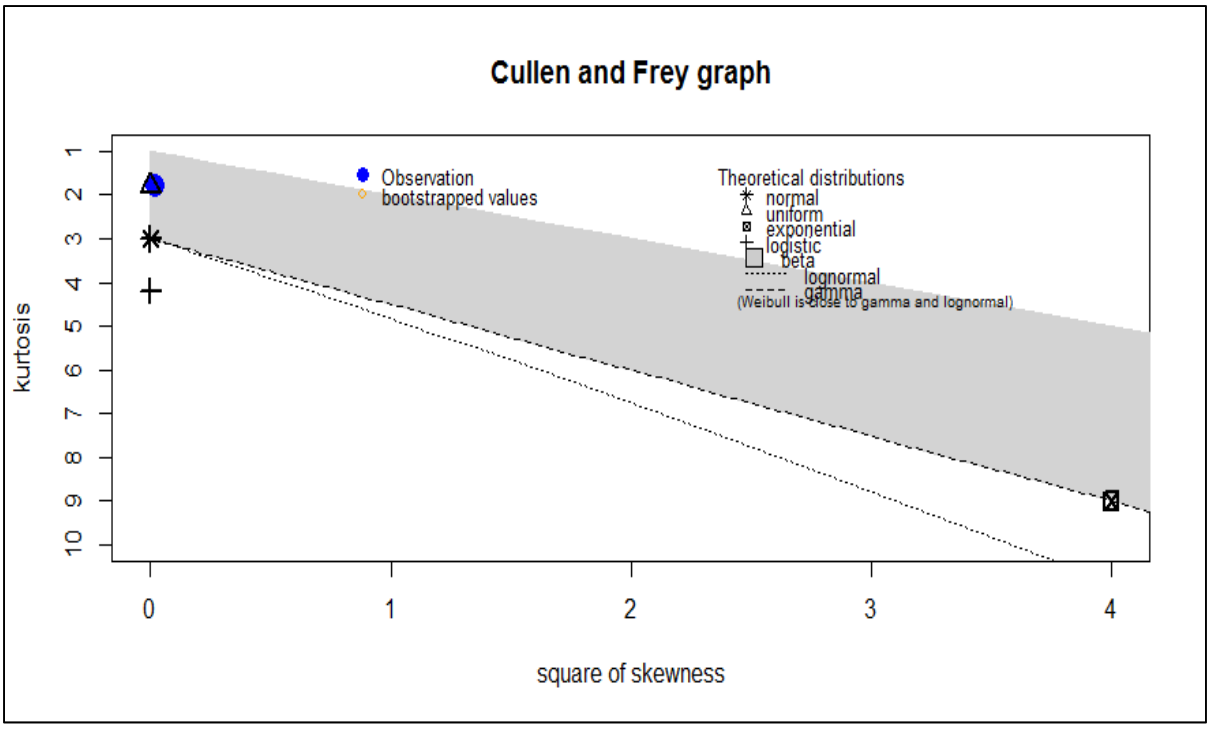


Figure 213 Cullen and Frey Graph of Power Transformed Confusability

The power-transformed confusability is still a Beta distribution, no closer to normal than the original data.

```
        Shapiro-Wilk normality test

data:  lme4::ranef(g1)$p[, 1]
W = 0.9739, p-value = 0.0135
```

Code Snippet 96 Attempt to Normalize with Power Transform Fails in R

We will attempt to log-normalize

```
        Shapiro-Wilk normality test

data:  lme4::ranef(g1)$p[, 1]
W = 0.9763, p-value = 0.02327
```

Code Snippet 97 Attempt to Log-Normalize Fails in R

$p=0.02327 < 0.05$ for Shapiro-Wilk which indicates non-normalcy, but improved.

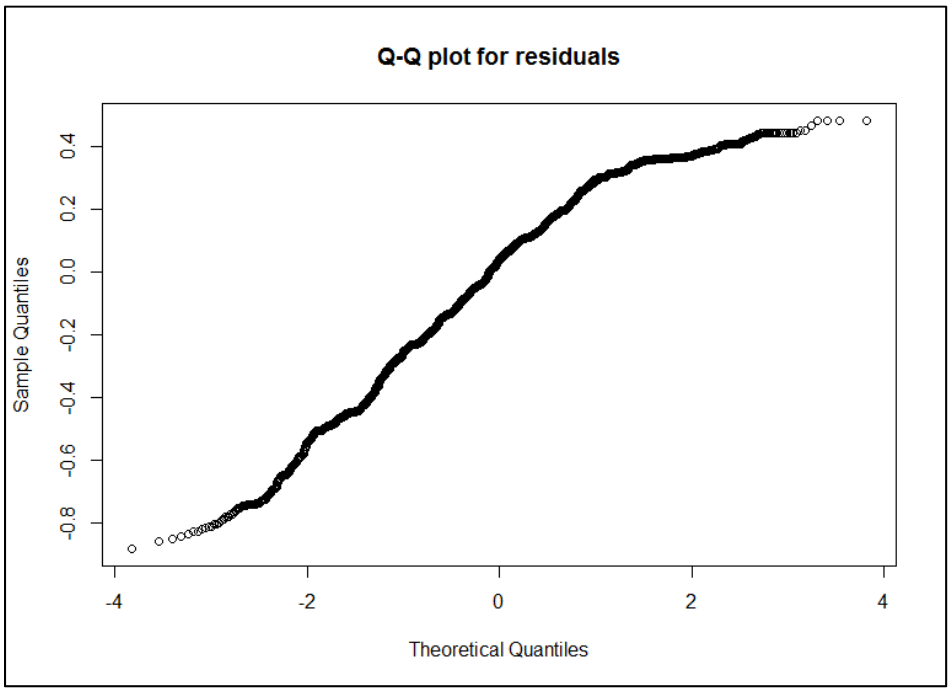


Figure 214 QQ Plot of Residuals of Log-Normalized Confusability

The Q-Q plot indicates we are closer to normal than Figure 210. We will proceed with analysis but provide a caveat on the violation of normalcy.

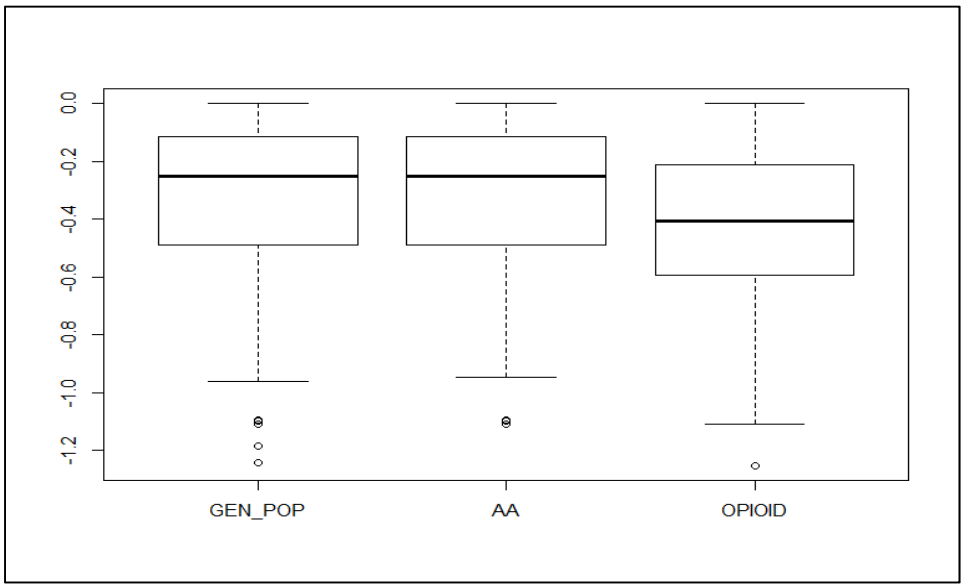


Figure 215 Predicted Probabilities of Confusability versus Group

```
> g.group3 <- lmer(ecCONFIDENCE ~group3+(1|p), data=EMO, REML=FALSE)
> THwaldCI2exp(g.eg,"eCROWD","ecCONFIDENCE","EXPRESS","eCROWD")

                  coef     lower     upper          pval effect trend
(Intercept)  0.7139849 0.7001313 0.7281126 3.865761e-259   TRUE FALSE
group3AA     0.7096297 0.6896590 0.7301788  6.681589e-01  FALSE FALSE
group3OPIOID 0.6498749 0.6294028 0.6710128  4.137989e-09   TRUE FALSE
```

Code Snippet 98 Confusability versus Group Model in R

There are significant differences between OPOIOD (SUBX) and the GP. However, the boxplot of Figure 215 and the non-normal residuals leaves this conclusion suspect at best. We relevel to see if there are differences between AA and SUBX.

```
> EMO$group3 <- relevel(EMO$group3,ref="AA")
> g.group3 <- lmer(ecCONFIDENCE ~group3+(1|p), data=EMO, REML=FALSE)
> summary(g.group3)
                  coef     lower     upper          pval effect trend
(Intercept)   0.7096297 0.6950500 0.7245153 1.765402e-239   TRUE FALSE
group3GEN_POP 0.7139849 0.6938916 0.7346601  6.681589e-01  FALSE FALSE
group3OPIOID  0.6498749 0.6289457 0.6715005  7.688622e-08   TRUE FALSE
```

Code Snippet 99 Confusability versus Group Model in R (ref AA)

There are significant differences between OPOIOD (SUBX) and the AA.

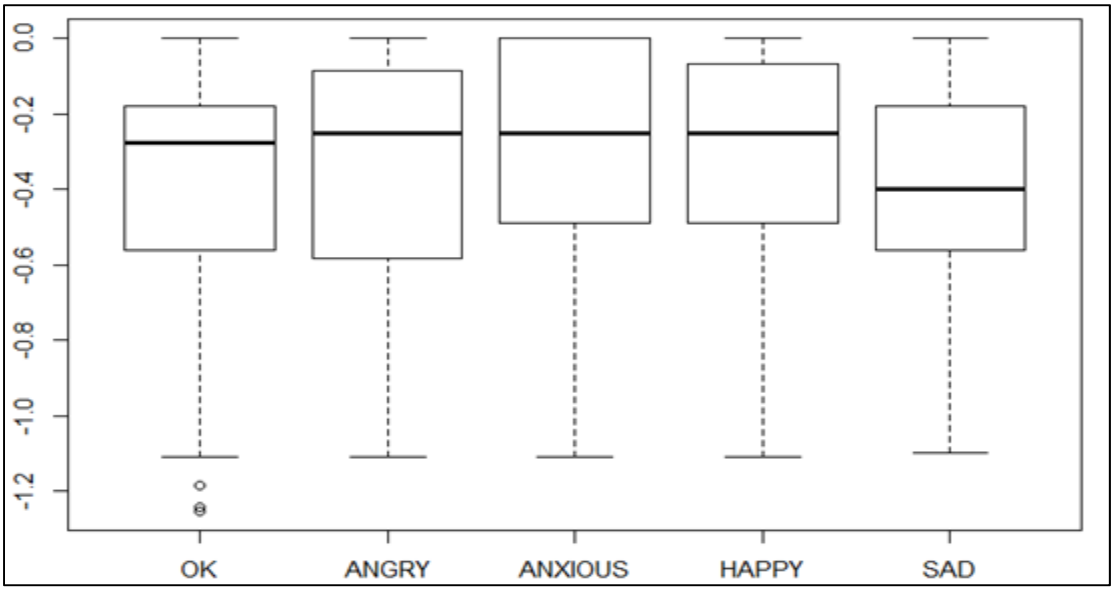


Figure 216 Predicted Probabilities of Confusability versus emotion

```
> EMO <- THloadEMO() # load the data
> EMO <- EMO[(!is.na(EMO$eCROWD)),]
```

```
> EMO$ecCONFIDENCE = log(EMO$ecCONFIDENCE)
> EMO$eCROWD <- relevel(EMO$eCROWD,ref="OK")
> g.eg <- lmer(ecCONFIDENCE~eCROWD+(1|p), data=EMO, REML=FALSE)
> THwaldCI2exp(g.eg,"eCROWD","ecCONFIDENCE","EXPRESS","eCROWD")
                    coef     lower     upper          pval effect trend
(Intercept)    0.6922661 0.6815289 0.7031724  0.000000e+00   TRUE FALSE
eCROWDANGRY    0.7028378 0.6857663 0.7203343  2.176852e-01  FALSE FALSE
eCROWDANXIOUS  0.7116853 0.6928822 0.7309987  3.878685e-02   TRUE FALSE
eCROWDHAPPY    0.7267683 0.7156490 0.7380603  2.804674e-10   TRUE FALSE
eCROWDSAD      0.6821038 0.6692697 0.6951841  1.194428e-01  FALSE FALSE
```

Code Snippet 100 Confusability versus Emotion Model in R

There are significant differences between Neutral, Anxious, and Happy.

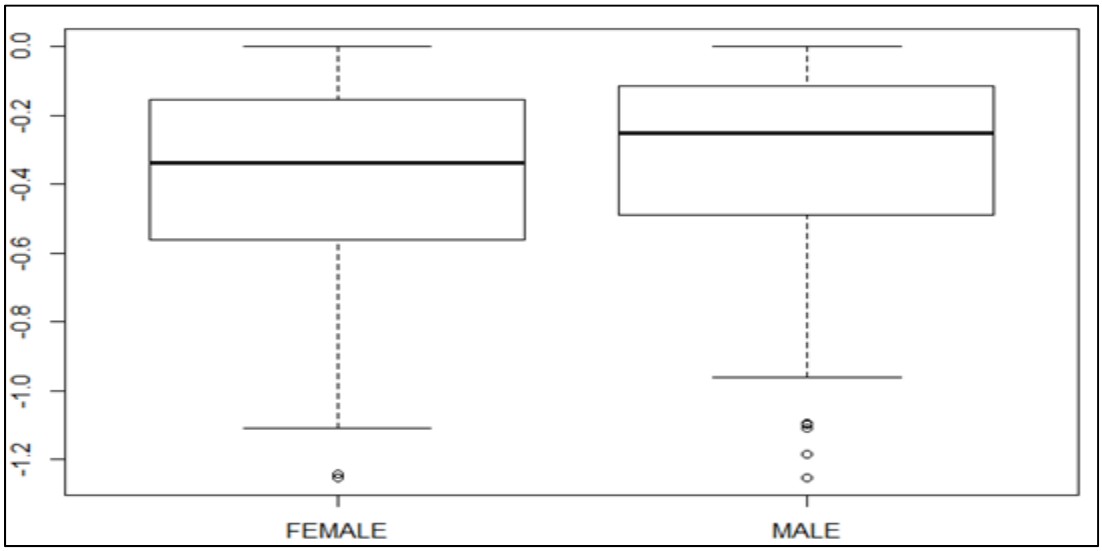


Figure 217 Predicted Probabilities of Confusability versus Gender

```
 > g.gender <- lmer(ecCONFIDENCE ~gender+(1|p), data=EMO, REML=FALSE)
> g5<-THwaldCI2(g.gender,"eCROWD","ecCONFIDENCE","EXPRESS","eCROWD")
                 coef      lower      upper          pval effect trend
(Intercept) -0.3734756 -0.3945502 -0.3524009 3.708692e-275   TRUE FALSE
genderMALE  -0.3498834 -0.3778265 -0.3219403  9.129946e-02  FALSE  TRUE

move back to linear domain

                 coef      lower      upper          pval effect trend
(Intercept)  0.6883378  0.6739831  0.7029982 3.708692e-275   TRUE FALSE
genderMALE   0.7047703  0.6853494  0.7247415  0.09129946e  FALSE  TRUE
```

Code Snippet 101 Confusability versus Gender Model in R

There is a trend that confusability differs between Males and Females.

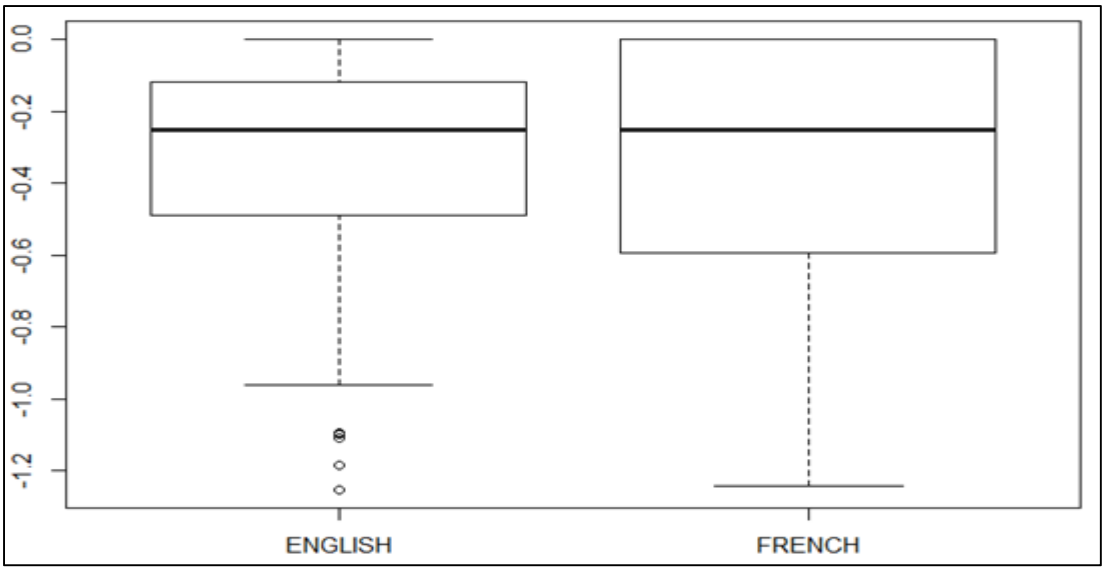


Figure 218 Predicted Probabilities of Confusability versus Language

```
> g.lang <- lmer(ecCONFIDENCE ~language+(1|p), data=EMO, REML=FALSE)
> THwaldCI2exp(g.lang,"eCROWD","ecCONFIDENCE","EXPRESS","eCROWD")
                    coef     lower     upper       pval effect trend
(Intercept)    0.6925572 0.6821161 0.7031581 0.00000000   TRUE FALSE
languageFRENCH 0.7188053 0.6959066 0.7424575 0.02155932   TRUE FALSE
```

Code Snippet 102 Confusability versus Language Model in R

APPENDIX K

ETHICS APPROVAL

Université du Québec
École de technologie supérieure
1100, rue Notre-Dame Ouest
Montréal (Québec) H3C 1K3
Téléphone : (514) 396-8800
Télécopieur : (514) 396-8950

03 août 2010

M. Edward Hill
M. Pierre Dumouchel
Département de génie logiciel et des TI

Objet : Approbation de votre projet de recherche intitulé « *An evidence-based toolset to measure and assess emotional health* ».

Messieurs,

Les modifications et précisions demandées par le CÉR dans sa lettre du 18 juin 2010 ayant été apportées adéquatement, votre projet peut aller de l'avant.

Veuillez toutefois noter que cette approbation n'est valable que pour une année. Vous devrez donc annuellement demander le renouvellement de l'approbation au Comité, sans quoi le projet sera considéré comme terminé. Également, nous attendons le rapport final de votre projet pour le **31 décembre 2010**. Vous trouverez le formulaire nécessaire à l'adresse suivante :

http://www.etsmtl.ca/zone2/administration/decanats/recherche/humains/formulaire_rapport.doc

Veuillez agréer, messieurs, l'expression de mes sentiments les meilleurs.

Liliana Guédez M.Env
Secrétaire
Comité d'éthique de la recherche

cc.Claude Bédard, Doyen à la recherche et au transfert technologique

ÉTS
Le génie pour l'industrie

Comité d'éthique de la recherche
École de technologie supérieure

Date : 29 septembre 2011

OBJET : **Titre du projet :** An evidence-based toolset to measure and assess emotional health.

Responsable du projet : Pierre Dumouchel

Décision : APPROBATION FINALE

Monsieur,

La demande de renouvellement pour le projet de recherche mentionné en rubrique a été déposée le 12 juillet 2011 pour évaluation par le comité d'éthique de la recherche de l'ÉTS. La présente lettre est pour vous informer que le CÉR de l'ÉTS a procédé le 28 septembre 2011, à l'évaluation du dossier en comité restreint.

Liste des documents soumis pour évaluation :

- Demande de renouvellement

J'ai le plaisir de vous informer que suite à l'analyse des documents que vous nous avez soumis, le projet a été **accepté sans condition.**

Veuillez toutefois noter que cette approbation n'est valable que pour une année. Vous devrez donc annuellement demander le renouvellement de l'approbation au Comité, sans quoi le projet sera considéré comme terminé. Dans la perspective où il devait être terminé, vous devrez fournir un rapport final suivant le modèle disponible sur la page Internet de l'ÉTS. Ce rapport est attendu pour le 31 mai 2013.

Veuillez agréer, Monsieur, l'expression de mes sentiments les meilleurs.

Paul V. Gervais, Ing., M.Eng.
Président
Comité d'éthique de la recherche

c.c. : Claude Bédard, Doyen à la recherche et au transfert technologique

BIBLIOGRAPHY

1. *Cogito Behavioral data systems*. 2014; Available from: http://www.cogitocorp.com/.

2. Scott R.L. K.M.F., Coombs R.H., *Affect-Regulation Coping-Skills Training: Managing Mood Without Drugs*, in *Addiction Recovery Tools: A Practical Handbook*, R.H. Coombs, Editor. 2001, Thousand Oaks, CA: Sage. p. 191-206.

3. Russell J.A., Bachorowski J.-A.,Fernández-Dols J.-M., *Facial and vocal expressions of emotion.* Annual review of psychology, 2003. 54(1): p. 329-349.

4. Banse R., *Acoustic Profiles in Vocal Emotion Expression.* Journal of Personality and Social Psychology, 1996. 70(3): p. 614-636.

5. Arnott M.I., *Towards the simulation of emotion in synthetic speech.* Jornal Acoustical Society of Speech, 1993. 93(2): p. 1097-1108.

6. Morrison D., *Ensemble methods for spoken emotion recogntion in call-centres.* Speech Communication, 2007. 49: p. 98-112.

7. Raquel Tato R.S., Ralf Kompe, J.M. Pardo. *EMOTIONAL SPACE IMPROVES EMOTION RECOGNITION*. in *ICSLP*. 2002.

8. Drake R.E., Goldman H.H., Leff H.S., Lehman A.F., Dixon L., Mueser K.T.,Torrey W.C., *Implementing evidence-based practices in routine mental health service settings.* Psychiatric services, 2001. 52(2): p. 179-182.

9. Boyd J.W., Linsenmeyer A., Woolhandler S., Himmelstein D.U.,Nardin R., *The crisis in mental health care: a preliminary study of access to psychiatric care in Boston.* Annals of emergency medicine, 2011. 58(2): p. 218.

10. Mark Olfson S.C.M., *National Trends in Outpatient Psychotherapy.* The American Journal of Psychiatry, 2010. 167(12).

11. Rehm J., *The Costs of Substance Abuse in Canada 2002*, C.C.o.S.A. (CCSA), Editor. 2006.

12. Savvas S.M., Somogyi A.A.,White J.M., *The effect of methadone on emotional reactivity.* Addiction, 2012. 107(2): p. 388-92.

13. Carson A.,Sabol J., *Prisoners in 2011*, U.S. Department of Justice, Office of Justice Programs,, Bureau of Justice Statistics, Editor. 2011.

14. National Institute of Mental Health. *The Numbers Count: Mental Disorders in America*. 2013; Available from: http://www.nimh.nih.gov/health/publications/the-numbers-count-mental-disorders-in-america/index.shtml.

15. American Psychiatric Association, *DSM-IV (Diagnostic and Statistical Manual of Mental Disorders)*. 2004.

16. Grant B.F., *Prevalence co-occurrence of substance use disorders and independent mood and anxiety disorders.* Archives of General Psychiatry, American medical Association: Results from the national Epidemiologic survey on alcohol and related conditions, 2004. 61(AUG 2004): p. 807-816.

17. Center for Behavioral Health Statistics and Quality, *National Survey of Substance Abuse Treatment Services (N-SSATS): 2011, Data on Substance Abuse Treatment Facilities*, Department Of Health And Human Services, Substance Abuse and Mental Health Services Administration, Editor. 2012.

18. Anonymous A., *2011 Membership Survey.* 2011.

19. E. Hill D.H., P. Dumouchel, N. Dehak, T. Quatieri, C. Moehs, M. Oscar-Berman, J. Giordano, T. Simpatico, K. Blum, *Automatic Detection of Flat Affect in Long-Term Suboxone™ Patients.* PLOS ONE (pending), 2013.

20. Policy O.o.N.D.C., *The Economic Costs of Drug Abuse in the United States, 1992-2002 . Washington, DC: Executive Office of the President (Publication No. 207303).* 2004.

21. Rehm J M.C., Popova S, Thavorncharoensap M, Teerawattananon Y, Patra J. , *Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders.* Lancet, 2009. 373(9682): p. 2223–2233.

22. Center for Behavioral Health Statistics and Quality, *Results from the 2011 National Survey on Drug Use and Health: National Findings*, Department Of Health And Human Services, Substance Abuse and Mental Health Services Administration Editor. 2012.

23. Lyubomrsky S. K.L., Diener E., *The Benefits of frequent positive affect: Does happiness lead to success?* The American psychological Association: Psychological Bulletin, 2005. 131(6): p. 803 -- 855.

24. Connell J., Brazier J., O'Cathain A., Lloyd-Jones M.,Paisley S., *Quality of life of people with mental health problems: a synthesis of qualitative research.* Health and quality of life outcomes, 2012. 10(1): p. 1-16.

25. Graham C., Eggers A.,Sukhtankar S., *Does happiness pay?: An exploration based on panel data from Russia.* Journal of Economic Behavior & Organization, 2004. 55(3): p. 319-342.

26. Michele M. Tugade B.L.F., Lisa F. Barret, *Psychological resilience and positive and motion granularity: Examining the benefits of positive emotions on coping and health.* National Institute of health: NIH public access, 2004. 72(6): p. 1161-1190.

27. Fredrickson B.L., *The Role of Positive Emotions in Positive Psychology: The Broaden-and-Build Theory of Positive Emotions.* American psychologist, 2001. 56(3): p. 218-226.

28. Dodge R., Sindelar J.,Sinha R., *The role of depression symptoms in predicting drug abstinence in outpatient substance abuse treatment.* J Subst Abuse Treat, 2005. 28(2): p. 189-96.

29. Wurmser L., *Psychoanalytic considerations of the etiology of compulsive drug use.* American Psychoanalytic Association, 1974. 22(4): p. 820-843.

30. Blum K., Chen A.L., Chen T.J., Bowirrat A., Downs B.W., Waite R.L.,Savarimuthu S., *Genes and happiness.* Gene Therapy and Molecular Biology, 2009. 13: p. 91-129.

31. National Institute on Drug Abuse, *Principles of Drug Addiction Treatment: A Research Based Guide*, N.I.o.H. National Institute on Drug Abuse, U.S. Department of Health and Human Services, Editor. 2009.

32. Dodge R., *The role of depression symptoms in predicting drug abstinence in outpatient substance abuse treatment.* Journal of Substance Abuse Treatment, 2005. 28: p. 189–96.

33. Miller G.A., *The Magical Number Seven, Plus or Minus Two.* Psychological Review 1956. 63(2): p. 81-97.

34. IEEE Signal Processing Society. *The INTERSPEECH 2009 Emotion Challenge: Results and Lessons Learnt*. 2009; Available from: http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2009-10/interspeech-emotion-challenge/.

35. Scherer K.R., *What are emotions? And how can they be measured?* Social Science Information, 2005. 44: p. 695-729.

36. Goleman D., *Emotional Intelligence*. 1995.

37. Ekman P., *Basic Emotions*. The Handbook of Cognition and Emotion. 1999: T. Dalgleish and T. Power

38. Tracy J.L.,Randles D., *Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt.* Emotion Review, 2011. 3(4): p. 397-405.

39. Öhman A., *Fear and anxiety.* Handbook of emotions, 2008: p. 709-729.

40. Michelle Patterson PhD S.F.U., *CBT in Practice: Part science, part art.* Visions, BC's Mental Health and Addictions Journal, 2009. 6(1).

41. Olson C., *Is 80% accuracy good enough.* Pecora 17 "The Future of Land Imaging… Going Operational", 2009: p. 18-21.

42. Nwe T.L., Foo S.W.,De Silva L.C., *Speech emotion recognition using hidden Markov models.* Speech communication, 2003. 41(4): p. 603-623.

43. Steidl S. L.M., Batliner A. , Noth E. ,Niemann H. . *"Of all things the measure is man" automatic classification of emotions and inter-labeler consistency*. in *ICASSP*. 2005.

44. Pierre Dumouchel N.D., Yazid Attabi1, Reda Dehak, Narjes Boufaden *Cepstral and Long-Term Features for Emotion Recognition*, in *INTERSPEECH 2009*. 2009.

45. Stone A.A., *Historical roots and rationale of ecological momentary assessment (EMA)*. The science of real-time data capture: self-reports in health research. 2007.

46. Arthur Stone S.S., *Capturing Momentary, Self-Report Data: A Proposal for Reporting Guidelines*, in *The Science of Real-Time Data Capture*. 2007.

47. Hufford M., *Special Methodological Challenges and Opportunities in Ecological Momentary Assessment*, in *Science of Real-Time Data Capture: Self-Reports in Health Research*. 2007.

48. Meeker M.,Wu L. *KPCB Internet Trends 2013*. in *Internet Trends D11 Conference*. 2013. Kleiner Perkins Caufield & Byers.

49. Bhaskara S., *Setting Benchmarks and Determining Psychiatric Workloads in Community Mental Health Programs.* Psychiatric Services, 1999. 50: p. 695-697.

50. Manlandro J.J., *Using buprenorphine for outpatient opioid detoxification.* JAOA: Journal of the American Osteopathic Association, 2007. 107(suppl 5): p. ES11-ES16.

51. Therapy B.I.f.C.B. *What is cognitive behavior therapy?* 2012 [cited 2012 July 13]; Available from: http://www.beckinstitute.org/cognitive-behavioral-therapy/.

52. Harley Therapy, Psychotherapy & Counselling London. *London Cognitive-Behavioural Therapy (CBT) Clinic*. 2013; Available from: http://www.harleytherapy.co.uk/cognitive-behavioural-therapy-london.htm.

53. Norcross J.C.,Wampold B.E., *Evidence-based therapy relationships: Research conclusions and clinical practices.* Psychotherapy relationships that work: Evidence-based responsiveness, 2011. 2: p. 423-430.

54. Jensen-Doss A.,Hawley K.M., *Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools.* Journal of Clinical Child & Adolescent Psychology, 2010. 39(6): p. 885-896.

55. Carroll K.M., *A vision of the next generation of behavioral therapies research in the addictions. Addiction.*, 2007. 102(6): p. 850-869.

56. Carroll K.M., *Computer-Assisted Delivery of Cognitive-Behavioral Therapy for Addiction Has Enduring Effects.* Drug and alcohol dependence, 2009. 100(1-2): p. 178-181.

57. Substance Abuse and Mental Health Services Administration, *Illness Management & Recovery: Evidence Based Practices: Monitoring Client Outcomes*, Department of Health and Human Services, SAMHSA Substance Abuse and Mental Health Services Administration, Editor. 2003.

58. Stuyt E.B., *Recovery Rates After Treatment for Alcohol/Drug Dependence: Tobacco Users vs. Non-Tobacco Users.* American Journal on Addictions, 1997. 6(2): p. 159-167.

59. Schumacher J.E., *Abstinent-contingent housing and treatment retention among crack-cocaine-dependent homeless persons.* Journal of Substance Abuse Treatment 2000. 19: p. 81-88.

60. Jan B., *WEIGHTED KAPLAN MEIER ESTIMATION OF SURVIVAL FUNCTION IN HEAVY CENSORING.* Pakistan Journal of Statistics, 2005. 21(1): p. 55-63.

61. Montgomery D., *Design and Analysis of Experiments: Sixth Edition*. 2005: John Wiley & Sons Inc.

62. Team R.D.C., *R: A Language and Environment for Statistical Computing*. 2012.

63. Steidl S. *FAU Aibo Emotion Corpus*. 2010; Available from: http://www5.informatik.uni-erlangen.de/our-team/steidl-stefan/fau-aibo-emotion-corpus/.

64. Statistics.com, The Institute for Statistics Education,. *Dependent and Independent Variables (Glossary of statistical terms)*. Available from: http://www.statistics.com/index.php?page=glossary&term_id=744.

65. StatsDirect. *Statisical Software*. 2013; Available from: http://www.statsdirect.co.uk/.

66. Loftus G.R.,Masson M.E., *Using confidence intervals in within-subject designs.* Psychonomic Bulletin & Review, 1994. 1(4): p. 476-490.

67. Whitley E.,Ball J., *Statistics review 2: Samples and populations.* Critical Care, 2002. 6(2): p. 143-148.

68. T Snijders R.B., *Multilevel Analysis : An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. 2012.

69. Steenbergen M. *Why Multilevel Analysis? Multilevel Data Structures and the Limits of Regression*. in *ELECDEM*. 2012.

70. Bates D. *Computational methods for mixed models*. 2012.

71. John Fox S.W., *An R COMPANION to APPLIED REGRESSION*. 2nd ed. 2011.

72. Monette G. *SPIDA 2012: Mixed Models with R*. 2012; Available from: http://scs.math.yorku.ca/index.php/SPIDA_2012:_Mixed_Models_with_R.

73. Zuur A.F., *Mixed effects models and extensions in ecology with R*. 2009: Springer.

74. NIST/SEMATECH. *e-Handbook of Statistical Methods*. 2012; Available from: http://www.itl.nist.gov/div898/handbook/.

75. Bradley K.D. *Chapter 8; Residual Analysis*. EPE & EDP 660 Research Design and Analysis in Education 2013 June 22, 2013]; Available from: http://www.uky.edu/~kdbrad2/.

76. Delignette-Muller M.L., Pouillot R., Denis J.-B.,Dutang C., *fitdistrplus: help to fit of a parametric distribution to non-censored or censored data*. 2013.

77. Hausman J.,McFadden D., *Specification tests for the multinomial logit model.* Econometrica: Journal of the Econometric Society, 1984: p. 1219-1240.

78. RABE-HESKETH S., *MULTILEVEL LOGISTIC REGRESSION FOR POLYTOMOUS DATA AND RANKINGS.* PSYCHOMETRIKA, 2003. 68(2): p. 267-287.

79. Hadfield J.D., *MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package.* Journal of Statistical Software, 2010. 33(2).

80. Team J.P.a.D.B.a.S.D.a.D.S.a.R.C., *nlme: Linear and Nonlinear Mixed Effects Models*. 2012.

81. Douglas Bates M.M., Ben Bolker, *lme4: Linear mixed-effects models using S4 classes*. 2011.

82. Ruiter S.,Van Tubergen F., *Religious Attendance in Cross-National Perspective: A Multilevel Analysis of 60 Countries1.* American Journal of Sociology, 2009. 115(3): p. 863-895.

83. Monette G., *spidadev: Collection of miscellaneous functions for mixed models etc. prepared for SPIDA 2009+ (development version)*. 2012.

84. Kirkman T.W. *Statistics to Use*. 1996 June 21, 2013]; Available from: http://www.physics.csbsju.edu/stats/.

85. Baroni M. *Regression 3: Logistic Regression (Practical Statistics in R)*. 2012; Available from: http://cogsci.uni-osnabrueck.de/~severt/SIGIL/#sigil.

86. Krause H. *datassit: making sense of your data*. 2012 [cited 2012; Available from: http://datassist.ca/.

87. Camille Szmaragd G.L., *Module 7: Multilevel Models for Binary Responses*, in *R Practical*.

88. Diez D.M., *Survival Analysis in R*. 2012.

89. Therneau T., *A Package for Survival Analysis in S*. 2012.

90. Sarkar D., *Statistics with R - Survival Analysis*, in *Summer Institute for Training in Biostatistics (2005)*. 2005, University of Wisconsin – Madison.

91. Diez D., *OIsurv: Survival analysis supplement to OpenIntro guide*. 2012.

92. Ahn L. V. M.B., McMillen C., Abraham D., Blum M., *reCAPTCHA: Human-Based Character Recognition via Web Security Measures*. Science, 2008. 321(5895): p. 1465 - 1468.

93. Morrison D., *Ensemble methods for spoken emotion recogntion in call-centres*. Speech Communication, 2006. 49: p. 98-112.

94. Ren C. Lua C.Y., *Multisensor Fusion and Integration: Approaches, Applications and Future Directions*. IEEE Sensors Journal, 2002.

95. Mohamed S. Kamel N.M.W. *Data Dependence in Combining Classifiers*. in *MCS'03 Proceedings of the 4th international conference on Multiple classifier systems*. 2003. Berlin.

96. Zeng Z., *A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 2009.

97. Marcel Kockmann L.B., Jan Honza Cernocky, *Brno University of Technology System for Interspeech 2009 Emotion Challenge*, in *INTERSPEECH 2009*. 2009.

98. Attabi Y., <*Memoire_Yazid_Attabi_2008_final.pdf*>. 2008.

99. Batliner A. *Combining Efforts for Improving Automatic Classification of Emotional User States*. in *Fifth Slovenian and First International Language Technologies Conference*. 2006. Ljubljana, Slovenia.

100. University M.S., *MSU Endpointer under LPGL license*, M.S.U. Institute for Signal & Information Processing, Editor. 2008.

101. al Y.e., *The HTK Book.* 2006.

102. Reynolds D.A., Quatieri T.F.,Dunn R.B., *Speaker Verification Using Adapted Gaussian Mixture Models☆☆☆.* Digital Signal Processing, 2000. 10(1-3): p. 19-41.

103. Gales M.J.,Woodland P., *Mean and variance adaptation within the MLLR framework.* Computer Speech and Language, 1996. 10(4): p. 249-264.

104. Kohavi R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. 1996.

105. Reynolds D.A., *Speaker Verification Using Adapted Gaussian Mixture Models.* Digital Signal Processing, 2000. 10: p. 19-41.

106. Darrah B., *Proc::ParallelLoop - Parallel looping constructs for Perl programs*. 2002.

107. Kuhn M. W.J., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., *caret: Classification and Regression Training*. 2013.

108. Bland M. *Measurement in Health and Disease: Cohen's Kappa*. 2008.

109. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages.* Psychometrika, 1947. 12(2): p. 153–157.

110. Ogura L.,Nussbaum D., *Measuring the Validity of the Automatic Emotion Detector in Detecting Mood Predictive of Performance on the Iowa Gambling Task*, in *Canadian Pyshcological Association*. 2013: Quebec City, Quebec, Canada.

111. Carlozzi A., *Urine Drug Monitoring*. 2008: Anesthesiology News.

112. Verstraete A.G., *Detection times of drugs of abuse in blood, urine, and oral fluid.* Therapeutic drug monitoring, 2004. 26(2): p. 200-205.

113. Blum K., Chen T.J., Bailey J., Bowirrat A., Femino J., Chen A.L., Simpatico T., Morse S., Giordano J.,Damle U., *Can the chronic administration of the combination of buprenorphine and naloxone block dopaminergic activity causing anti-reward and relapse potential?* Molecular neurobiology, 2011. 44(3): p. 250-268.

114. Madrid G., MacMurray J., Lee J., Anderson B.,Comings D., *Stress as a mediating factor in the association between the DRD2 Taqi polymorphism and alcoholism.* Alcohol, 2001. 23(2): p. 117-122.

115. Wooley C.N., Rogers R., Fiduccia C.E.,Kelsey K., *The Effectiveness of Substance Use Measures in the Detection of Full and Partial Denial of Drug Use.* Assessment, 2012.

116. Blum K.,Han D., *Genetic Addiction Risk Scores (GARS) coupled with Comprehensive Analysis of Reported Drugs (CARD) provide diagnostic and outcome data supporting the need for dopamine D2 agonist therapy: Proposing a holistic model for Reward deficiency Syndrome (RDS).* Addiction Research & Therapy 2012. 3(4).

117. S. Rabe-Hesketh A.S., A. Pickles, *Generalized Linear Latent And Mixed Models (GLLAMM)* 2012.

118. Jaeger F. *multinomial random effects models in r*. 2009; Available from: http://hlplab.wordpress.com/2009/05/07/multinomial-random-effects-models-in-r/.

119. Li B., Lingsma H.F., Steyerberg E.W.,Lesaffre E., *Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes.* BMC Med Res Methodol, 2011. 11: p. 77.

120. N. Kazantzis F.D., K. Ronan, *Homework Assignments in Cognitive and Behavioral Therapy: A Meta-Analysis.* Clinical Psychology: Science and Practice, 2006. 7(2): p. 189-202.

121. Vahabzadeh M. L.J.-L., Mezghanni M., Epstein D.H., Preston K.L., *Automation in an Addiction Treatment Research Clinic: Computerized Contingency Management, Ecological Momentary Assessment, and a Protocol Workflow System*, in *Practice management Conference*, M.G.M. Associtiona, Editor. 2010: New Jersey. p. 3-11.

122. Stritzke W. D.L., Durkin K., Houghton S., *Use of interactive voice response (IVR) technology in health research with children.* Behavior Research Methods, 2005. 37(1): p. 119-126.

123. Drupal. *The Drupal overview*. 2010; Available from: http://drupal.org/getting-started/before/overview.

124. W3C. *Voice Browser Call Control: CCXML Version 1.0*. 2010; Available from: http://www.w3.org/TR/ccxml/.

125. W3C. *Voice Extensible Markup Language (VoiceXML) Version 2.0*. 2004; Available from: http://www.w3.org/TR/voicexml20/.