Self-Report Measures and Mobile Applications for Mental Health – Focusing on Suicidality and Bipolar Disorder – Including App Development and Measure Validation

by

Eric Christopher Chan

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Psychiatry
University of Alberta

Abstract

This thesis describes studies evaluating the predictive validity of suicide risk assessments, assessment and monitoring of bipolar disorder using mobile app-based self-report questionnaires, the feasibility and criterion validity of novel tools for the evaluation of suicidal ideation and behaviours, otherwise known as suicidality, namely the Suicide Ideation and Behavior Assessment Tool (SIBAT) and mobile applications. As suicidality varies over time and is sensitive to a wide range of factors, the accurate assessment of suicide risk remains a challenge in psychiatry. Self-report scales have been developed for the assessment of suicide risk, but their utility has been debated. As such, I have conducted a systematic review of studies examining the predictive validity of self-report scales in the prediction of future suicide attempts and death by suicide. The results of this review suggest that no existing scale has sufficient validity for routine clinical use and all scales studied had particularly low positive predictive value in the prediction of death by suidide. I also conducted a systematic review of studies with Dr. Sudhakar Sivapalan examining the feasibility and validity of assessment and monitoring of bipolar disorder using mobile app-based self-report tools. At the time that this review was conducted, the data suggested that self-report tools were valid in the assessment of symptoms of mania, but their validity in the assessment of depression was unclear. Our findings were limited by the low number of studies identified for inclusion. The development of novel tools to assess suicide risk may improve our ability to assess suicidality and observe its changes over time and in response to different interventions. The SIBAT is a new tool developed for the evaluation of suicidality that encompasses a wide range of factors associated with suicide risk. It was developed for repeat administration with the goal of measuring changes in suicidality over time. Mobile applications

have also been developed for management of suicidality, but there are few data on the validity of mobile application-based assessment tools for this purpose. In these studies, I have developed a mobile application for the SIBAT. As an add-on study, participants in an addictions study at the University of Alberta were invited to complete the SIBAT using either the mobile application or the Qualtrics interface (programmed by Dr. Bradley Green). Data collected using the SIBAT via mobile device or Qualtrics were pooled and compared to data collected using the Mini International Neuropsychiatric Interview (MINI). Data collected using the SIBAT showed high internal consistency, and the sum of scores from modules 2 and 3 had good concurrent validity with the scores from the MINI and module 5 of the SIBAT. Participants who completed the SIBAT using the application were divided into two groups, one group completed the SIBAT using a mobile device and the other completed the SIBAT using a personal computer. I compared these two groups to assess the concurrent validity of data collected via mobile device with the data collected via personal computer. Participants completing the scale using a mobile device had a higher proportion of scale completion compared to participants completing the scale using a personal computer. A trend toward an increase in disclosed suicidality was also observed in the mobile device group. This suggests that participants may be more willing to report suicidality using their mobile device, which replicates a small prior study; however, replication of this finding using larger populations is needed. These findings indicate that both mobile applications and the SIBAT show promise as tools for the evaluation of suicidality. Further research assessing the administration of these tools over time may improve our understanding of their potential uses in both research and clinical settings.

Preface

This thesis is original work by Eric Chan. The amendments to the research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, as amendments to the Project Name "Genes associated with sexual addiction and related conditions (AddGenes Study)", No. Pro00066552, on December 12, 2017 (SIBAT on Qualtrics) and January 30, 2018 (SIBAT on Appsheets app).

Some of the research conducted for this thesis forms part of an international research collaboration, led by Dr. Katherine J. Aitchison at the University of Alberta. The SIBAT rating scale was developed by a team lead by Dr. Larry Alphs while he was working for Janssen Pharmaceutica. The application for the SIBAT used for data collection in Chapters 4 and 5 was programmed by myself, with permission from Janssen, after being trained to use the AppSheets platform by Dr. Andrius Baskys. The SIBAT, ASRS and IAT were programmed into the Qualtrics interface by Dr. Bradley Green. The MINI was its own software. Chapter 4 used SIBAT data collected via Qualtrics and the SIBAT application in conjunction with MINI data and Chapter 5 used SIBAT data collected via the SIBAT application as well as MINI, ASRS and IAT data.

Ethics approval was obtained by Keanna Wallace, Leslie Roper, Dawon Lee and Dr. Aitchison.

Participant recruitment and data collection was conducted by Keanna Wallace, Esther Yang, Leslie Roper, Garima Aryal, Grace Li, Hana Graham and Narin Sheri (hereafter referred to as the recruitment team) who drafted and sent recruitment emails to potential participants who had been enrolled in the AddGenes study. Instructions on how to access and complete the SIBAT

application (drafted by myself in consultation with Dr. Aitchison) were sent by the recruitment team to AddGenes participants who expressed interest in the SIBAT add-on study. Risk alerts for responses indicative of elevated risk were programmed by myself. Risk thresholds were determined by myself, Dr. Aitchison, and Keanna Wallace, and thresholds were adjusted by the team based on early responses. Keanna Wallace, Leslie Roper and Esther Yang contacted participants whose responses triggered an alert in order to arrange in-person risk assessments, having discussed or corresponded with Dr. Aitchison as required. Risk assessments were performed by Dr. Aitchison, myself, Dr. Rohit Lodhi, and Leslie Roper with some contributions from other members of the psychiatry resident body.

I was responsible for data analysis and text composition. Dr. Rohit Lodhi contributed to text edits for chapter 5. Dr. Sudhakar Sivapalan provided supervision for the review process and text preparation of chapter 3, including acting as the second reviewer in the search process. Dr. Katherine J. Aitchison was the supervisory author of chapters 4 and 5 and was involved with concept formation and oversaw all aspects of the research in those chapters. Dr. Katherine J. Aitchison also reviewed the entirety of this thesis and provided feedback and guidance contributing to the content of each chapter.

Acknowledgments

I would like to thank the University of Alberta Department of Psychiatry for supporting me financially through payment of tuition and providing some funding through the Department of Psychiatry Resident Research Fund, as well as allowing me to spend time completing my Master's work. I would also like to thank Janssen for their fellowship contributing a small proportion to supporting research assistant time. Furthermore, I would like to thank Janssen for allowing us to use the SIBAT rating scale in this research and Dr. Larry Alphs (while he was working with Janssen) for providing us with the latest version of this measure. I would also like to acknowledge the support I have received through two Rudelle Hall Graduate Scholarships. This award allowed me to better focus on my research and reinforced to me that my work was recognized and important.

I would also like to thank Dr. Larry Alphs for our discussions on risk alert thresholds for the SIBAT as well as potential directions to pursue in the future. In addition, I would like to thank Dr. Richard Isenberg and Dr. Patrick Carnes, both of whom contributed to overall study design of the overall AddGenes study on which this study was built. I would also like to acknowledge the American Foundation for Addiction Research for their contribution to the licenses for the MINI. Furthermore, I would like to acknowledge the AppSheets team, whose platform allowed me to program the SIBAT application used in this research.

Lastly, I would like to recognize the assistance provided to me by the members of my supervisory committee, Dr. Sudhakar Sivapalan and Dr. Vincent Agyapong, and I would like to acknowledge the extensive support provided by my supervisor, Dr. Katherine Jean Aitchison.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADHD	Attention Deficit Hyperactivity Disorder
ASRS	ADHD Adult Self-report Scale
ASRM	Altman Self-Rating Mania Scale
BDI	Beck Depression Inventory
BHS	Beck Hopelessness Scale
BSS	Beck Scale of Suicidal Ideation
CDS	Calgary Depression Scale
C-SSRS	Columbia Suicide Severity Scale
HDRS-17	Hamilton Depression Rating Scale-17
IAT	Internet Addiction Test
IDS-C	Inventory of Depressive Symptomatology, clinician-rated
ISST	InterSePT Scale for Suicidal Thinking
KIVS	Karolinska Interpersonal Violence Scale
Little's MCAR test	Little's Missing Completely at Random test
MADRS	Montgomery Asberg Depression Rating Scale
MDQ	Mood Disorder Questionnaire

MINI	Mini International Neuropsychiatric Interview
MSPS	Modified Sad Persons Scale
NGASR	Nurses' Global Assessment of Suicide Risk Scale
NPV	Negative Predictive Value
PHQ-9	Patient Health Questionnaire-9
PPV	Positive Predictive Value
RDoC	Research Domain Criteria
SCI	Suicide Crisis Inventory
SCS	Suicide Cognitions Scale
SD	Standard Deviation
SIBAT	Suicide Ideation and Behavior Assessment Tool
SIS	Suicide Intent Scale
SPS	Sad Persons Scale
STS-3	Suicide Trigger Scale v.3
SUAS	Suicide Assessment Scale
YMRS	Young Mania Rating Scale

CHAPTER 1: INTRODUCTION

Suicide prevention is a major component of psychiatric care; however, the assessment of suicide risk remains an ongoing challenge due to the multifactorial and dynamic nature of suicidality. While many instruments exist for the prediction of suicide risk, recent studies have shown that none predicted suicide or suicidal behavior with sufficient accuracy for routine clinical use.[1-4] The third edition of the American Psychiatric Association Practice Guidelines indicate that no scale has been shown to produce a clinically useful score for suicide prediction and as such, the development of validated procedures for the determination of suicide risk has been identified as a priority in suicide research by the National Action Alliance for Suicide Prevention.[5,6]

One tool that has the potential to improve our ability to assess suicidality is the newly developed Suicide Ideation and Behavior Assessment Tool (SIBAT). The SIBAT was developed by a group of clinical trial and academic experts in scale development.[7] It was designed to be sensitive to changes over time and to assess a wide range of risk factors that have previously been identified as important in suicide risk assessment. The SIBAT consists of 5 patient-reported modules and 3 clinician-rated modules.[7-12] The patient-reported component assesses factors such as affective states, neurovegetative symptoms, physical and psychological distress, current thoughts of suicide and other factors in addition to previous history of suicidal behaviors. It was designed for assessments on a repeated or serial basis, assessing static risk factors only on the first administration. The ability of the SIBAT to monitor a wide range of risk factors over time differentiates it from previously developed tools for suicide risk assessment, some of which, such as the Beck Suicidal Ideation Scale (SIS) and the Columbia-Suicide Severity Rating Scale (C-SSRS), focus predominantly on demographic factors and the presence of suicidal thoughts or intent and a history of suicidal behaviors.[13-15] It is therefore possible that the SIBAT may

facilitate enhanced longitudinal monitoring of suicidal ideation and behaviours over time and changes in associated risk factors.

Mobile applications may increase the frequency with which suicide risk assessments can be administered, as users are able to complete assessments without having to present to a treatment setting. Interestingly, one study examining the monitoring of symptoms of depression using a mobile application for the PHQ-9 suggested that users may be more willing to disclose suicidality using a mobile device when compared to conventional paper-and-pencil instruments.[16] Though many apps have been developed for suicide prevention, few apps focus on the assessment of suicide risk and at the time of data analysis, there were no data on the validation of mobile apps as tools for suicide risk assessment to our knowledge.[17]

The goal of this research was to evaluate the validity of the self-report component of the SIBAT and a mobile application as tools for suicide assessment. Firstly, we conducted a literature review examining the existing evidence for the predictive validity of scales for suicide risk assessment. Secondly, we performed a literature review of the current evidence examining mobile applications as tools for monitoring symptoms of bipolar disorder. This topic was chosen because, as noted above, we were unable to find any papers examining the validity of mobile applications specifically for suicide risk assessments. We chose to examine data on mobile applications for monitoring symptoms of bipolar disorder given the high rates of suicide attempts and death by suicide in this illness. Thirdly, a mobile application for the self-report component of the SIBAT was developed and university students and trainees participating in a study of addictions were invited to complete the scale as an add-on study. Fourthly, we evaluated the internal consistency of responses obtained using the SIBAT using the Cronbach's alpha measure

and item-total correlations. Fifthly, we compared responses obtained using the self-report component of the SIBAT (administered via application or by the Qualtrics interface) with responses obtained using the Mini International Neuropsychiatric Interview (MINI) in order to assess the concurrent validity of the self-report component of the SIBAT. Finally, we compared responses obtained via mobile device to responses obtained via personal computer to evaluate the concurrent validity of a mobile application as a tool for suicide risk assessment.

CHAPTER 2: A REVIEW OF THE PREDICTIVE VALIDITY OF SUICIDE RISK ASSESSMENTS

2.1 Abstract

Introduction: Suicide is an ongoing issue in Canada and worldwide. In 2016, the suicide rate was 11.0 per 100,000 people in Canada and 10.5 per 100,000 people globally. Detection and assessment of suicide risk is a significant component of effective suicide prevention and intervention. While many instruments have been developed for suicide risk assessment, they have been criticized for perceived inaccuracy in their ability to classify patients into different categories of risk.

Methods: In order to identify data examining the predictive validity of rating scales for suicide assessment, I conducted a search of the Pubmed, Ovid Medline and Embase databases for records evaluating the predictive validity of suicide risk assessment tools from the preceding 10 years. I focused on records reporting the sensitivity, specificity, positive predictive value (PPV) and/or negative predictive value (NPV) of suicide risk assessments using outcomes of suicidal behavior and suicide attempts.

Results: Sixteen records examining a wide range of instruments were identified for inclusion in this review. The psychometric properties of these instruments varied depending on the study method, but overall these instruments showed limited utility in clinical decision making. The PPV of these instruments for death by suicide was particularly low, with none reporting PPV > 20%.

Discussion: I review a number of the difficulties researchers face in the development of instruments for suicide risk assessment including the rarity of suicide as an event, the difficulty

in defining and identifying suicidality, the complex interplay between dynamic factors, and reliance on self-report. I describe different methods in which technology such as repeat measurements, physiologic monitoring, and machine learning could be used to overcome these difficulties.

Conclusion: Current evidence indicates limited clinical utility for the use of rating scales as tools for suicide prediction. Many barriers exist in the development of an accurate, practical tool for suicide prediction; new technologies may allow us to overcome some of these barriers.

2.2 Introduction

According to data from the 2015 Canadian Community Health Survey, about 12.1% of Canadians aged 15 or older reported having seriously contemplated suicide in their lifetime.[18] Amongst those who had ever contemplated suicide, 26.5% reported actually having attempted suicide at some point in their life.[18]

Detection and assessment of suicide risk is a significant component of effective suicide prevention and intervention. In 2014, the National Action Alliance for Suicide Prevention Research Prioritization Task Force in the United States identified the development of validated procedures that can "determine the degree of suicide risk (e.g., imminent, near-term, long-term) among individuals in diverse populations and in diverse settings through feasible and effective screening and assessment approaches" as one of their aspirational goals.[6] According to Claassen et al. (2014), numerous challenges exist in the development of an accurate model for suicide risk assessment including: difficulties in defining "elevated risk;" accurate identification of suicidal "intent" and distinguishing between suicide attempts and non-suicidal self-harm; the complex multifactorial interactions that contribute to suicide risk; the "daunting" and conceptually

"imprecise" number of nonspecific, static risk factors described in the literature; and the rare but significant nature of completed suicide posing problems for many statistical methods of analysis.[19]

Different statistical concepts are used in the evaluation of rating scales. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are among the most commonly used. Sensitivity refers to the probability of a test being positive when the disease is present and specificity is the probability of a test being negative when the disease is absent.[20] These measures evaluate the ability of the test to accurately classify an individual with the disease present (sensitivity) or with the disease absent (specificity). Sensitivity and specificity do not vary with the prevalence of the disease of interest. PPV is the probability of a disease being present given a positive test result and NPV is the probability of a disease being absent given a negative test result.[20] Both NPV and PPV vary with the prevalence of illness and, given the rarity of suicide in the population, have more relevance to clinical decision making.[1]

Many instruments have been studied for their ability to assess suicide risk. Examples of these include: the SAD PERSONS Scale[21], the Columbia Suicide Severity Rating Scale (C-SSRS)[13,15], the Beck Hopelessness Scale (BHS)[22], the suicide item of the Beck Depression Inventory (BDI)[23], the suicidality module of the Mini International Neuropsychiatric Interview (MINI)[24] and the Suicide Intent Scale (SIS).[13,15,21,25] These scales have been criticized, however, for their perceived inaccuracy in terms of ability to classify patients into different categories of risk.[26] In order to examine the predictive validity of suicide rating scales, we conducted a review of the data evaluating rating scales as tools predicting suicide risk. Our goal was to identify

records that examined the ability of rating scales to predict suicide or suicidal behavior, as that is often identified as the primary application of these tools.

2.3 Methods

Searches were conducted of the PubMed, Ovid Medline and Embase databases on October 25, 2018 for records from the preceding 10 years. Studies were identified by searching digital databases and reviewing references of extant articles. The search was limited to trials using human adults, and to publications available in English. We defined suicide risk as suicide or suicidal behaviour, with indicators of the latter including repeated suicide attempts and hospital admissions related to such. Records were screened and excluded for the following reasons: did not focus primarily on the assessment of suicide risk, such as papers focused on assessment of nonsuicidal self-injury; did not focus on the validation of a rating scale for suicide risk assessment, such as papers focused primarily on the identification of risk factors for suicide without applying these factors within the framework of a rating scale; the patient population predominantly fell outside of the target age range (18 – 65 years of age); was a protocol paper, case report, or editorial letter.

We then reviewed in-depth the remaining 135 records and records of interest referenced in these articles. We excluded any records that were not longitudinal studies in which a prospective cohort was followed and assessed for clinically relevant outcomes (repeat suicide attempt, hospital admission or suicide completion). We also excluded any papers that did not report results in the form of sensitivity/specificity or PPV/NPV as these were the statistical measures of interest in this study. Some records examined outcomes at multiple time points.[24,27,28] We specify the timelines of interest examined for each study.

2.4 Results

2.4.1 Identified Records

The flow diagram of the search method is depicted in the figure below (Figure 1). Initial searches produced 1949 unique records after removal of duplicates. After excluding papers for the reasons described above, 16 records were identified and their references were also searched for further relevant studies. One study[29] was excluded at this point as the data presented was incorporated into another study identified in this review.[28] The full texts of the identified studies were reviewed and tables describing the ability of the measure to predict suicidal behavior (Table 1) and death by suicide (Table 2) are reported below.

2.4.2 InterSePT Scale for Suicidal Thinking (ISST)

The InterSePT Scale for Suicidal Thinking (ISST) is a 12-item scale that examines suicidal behaviors in the previous 7 days through a semi-structured interview.[30,31] Each item is scored on a scale from 0-2. Each point of increased severity on the ISST total score was associated with a 5% increased risk of suicidal behavior (suicide attempt or hospitalization to prevent imminent risk of suicide as determined by a blinded suicide monitoring board) in individuals at high risk of suicide with schizophrenia or schizoaffective disorder.[31] Ayer et al. (2008), examined the ability of changes in ISST scores to predict future suicide attempts or hospitalizations to prevent suicide attempts.[32] They examined the changes in ISST scores between the assessment immediately preceding the attempt or hospitalization (median 11.5 days, range 1 – 76 days) and the assessment occurring 2 – 12 weeks prior. In keeping with data by Potkin et al. (2003), the absolute scores of the ISST showed a significant difference between the group with suicide attempts/hospitalizations and matched controls.[31] Although change in ISST score also showed

significant between group differences, and increase in ISST score showed a high degree of specificity and PPV for future attempts/hospitalizations, the sensitivity and NPV were poor.[32]

2.4.3 Calgary Depression Scale (CDS)

The same methodology used to examine the ISST described above was also used to examine the predictive value of the CDS, a nine-item measure for assessing depression in schizophrenia.[32] Similar characteristics were found, with an increase in absolute scores amongst individuals with future suicide attempts, or hospitalizations to prevent suicide attempts. With cutoff scores of ≥ 6, the CDS had poor sensitivity (10%) and high specificity (99%) with a PPV of 76% and an NPV of 70%. With cutoff scores of ≥ 0, the sensitivity increased to 20% with a drop in specificity to 90% and PPV and NPV to 49% and 70% respectively.

2.4.4 SAD PERSONS and Modified SAD PERSONS Scales

The SAD PERSONS scale (SPS) is a mnemonic that uses ten major risk factors for suicide to estimate suicide risk.[21] The modified SAD PERSONS scale (MSPS) replaces the "sickness" item of the SAD PERSONS scale with "stated intent" and scores four of the items as two points instead of one point.[27] Three studies examined the predictive validity of the SPS and MSPS in patients referred to psychiatry or to a specialist self-harm team after presenting to general hospital emergency departments.[27,33,34] Bolton et al. (2012) and Saunders et al. (2014) examined presentations who re-presented to the emergency department with a suicide attempt in 6 months.[33,34] Katz et al. (2017) used death by suicide at 6 months, 1 year and 5 years (as determined by administrative data) as the outcome of interest.[27]

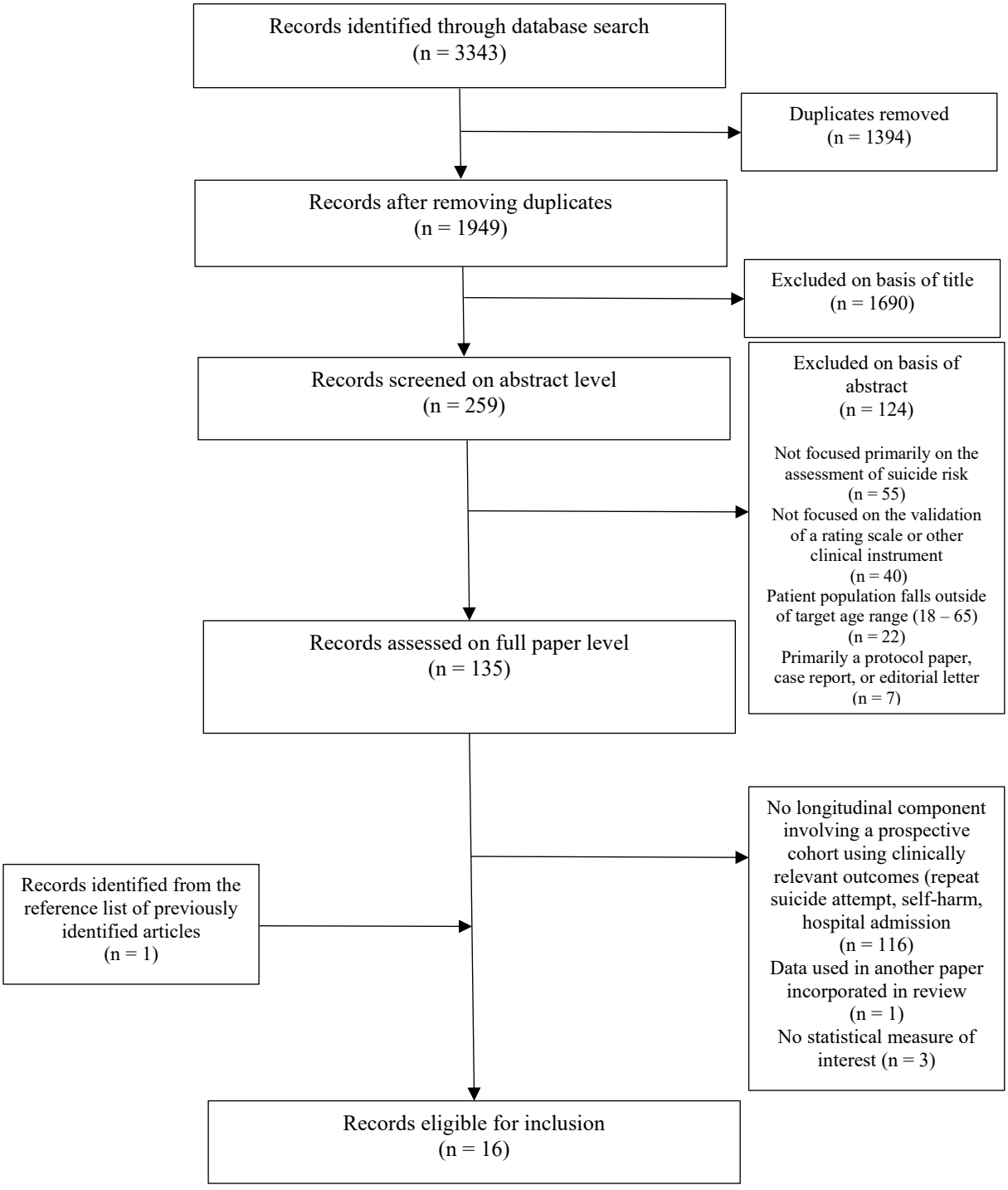

Figure 1. Flow diagram of search process to identify papers reporting on the ability of rating scales to predict future suicidal behavior

Table 1. Predictive validity of rating scales in the prediction of suicidal behavior

Rating Scale	Reference	Cut-off Score	Sens.	Spec.	PPV	NPV
ISST	Ayer et al., 2008	Increase ≥ 6 after 2 – 12 weeks	9%	100%	100%	70%
		Increase ≥ 0 after 2 – 12 weeks	19%	97%	73%	71%
CDS	Ayer et al., 2008	≥ 6	10%	99%	76%	70%
		≥ 0	20%	90%	49%	70%
SPS	Bolton et al., 2012	≥ 7	19.6%	90.7%	-	-
	Saunders et al., 2014	≥ 7	6.6%	96.8%	-	-
	Bolton et al., 2012	≥ 5	48.8%	69.2%	3.8%	97.7%
MSPS	Bolton et al., 2012	≥ 9	40.0%	90.5%	-	-
	Bolton et al., 2012	≥ 6	56.7%	60.8%	4.2%	98.0%
STS-3	Yaseen et al., 2014	≥ or ≤ one standard deviation from mean	69.2%	68.3%	40.9%	87.5%
SCI	Galynker et al., 2017	> 114	63.6%	88.2%	-	-
BDI	Green et al., 2015	3	25%	88%	47%	73%
		2 or 3	61%	60%	40%	78%
KIVS	Haglund et al., 2016	≥ 6	62%	53%	-	-
		≥ 6 (for violent means)	81%	52%	-	-
C-SSRS	Mundt et al., 2013	Suicidal ideation with intent to act & suicide attempts	67%	76%	-	-
	Madan et al., 2016	≥ 23	69.4%	65.2%	-	-
BHS	Madan et al., 2016	≥ 13	63.3%	56.1%	-	-
BSS	Madan et al., 2016	≥ 14	58.1%	65.0%	-	-
SCS	Madan et al., 2016	≥ 55	58.1%	67.8%	-	-
PHQ-9 Item 9	Madan et al., 2016	≥ 3	86.1%	33.1%	-	-
MINI Suicidal Scale	Roaldset et al., 2012	≥ 10	61%	75%	43%	86%
		≥ 6	73%	62%	39%	88%
SUAS	Waern et al., 2010	≥ 24	61%	60%	-	-

Sens.: Sensitivity; Spec.: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; ISST: InterSePT Scale for Suicidal Thinking; CDS: Calgary Depression Scale; SPS: Sad Persons Scale; MSPS: Modified Sad Persons Scale; STS-3: Suicide Trigger Scale; SCI: Suicide Crisis Inventory; BDI: Beck Depression Inventory; KIVS: Karolinska Interpersonal Violence Scale; C-SSRS: Columbia Suicide Severity Rating Scale; BHS: Beck Hopelessness Scale; BSS: Beck Scale of Suicidal Ideation; SCS: Suicide Cognitions Scale; PHQ-9: Patient Health Questionnaire-9; MINI: Mini International Neuropsychiatric Interview; SUAS: Suicide Assessment Scale;

Table 2. Predictive validity of rating scales in the prediction of death by suicide

Rating Scale	Reference	Cut-off Score	Sens.	Spec.	PPV	NPV
SPS	Katz et al., 2017	≥ 5	48.1%	60.0%	6%	99.6%
MSPS	Katz et al., 2017	≥ 6	57.7%	59.6%	7%	99.6%
BDI Item 9	Green et al., 2015	3	12%	99%	8%	99%
		2 or 3	28%	93%	3%	99%
PHQ-9 Item 9	Simon et al., 2016	Either of two highest responses	35%	-	3%	-
	Simon et al., 2016	Any positive response	64%	-	< 2%	-
SIS	Stefansson et al., 2012	> 16	100%	52%	16.7%	-
SIS + KIVS	Stefansson et al., 2015	SIS > 16, KIVS ≥ 6	100%	63%	18.8%	-

Sens.: Sensitivity; Spec.: Specificity; PPV: Positive Predictive Value; NPV: Negative Predictive Value; SPS: Sad Persons Scale; MSPS: Modified Sad Persons Scale; BDI: Beck Depression Inventory; PHQ-9: Patient Health Questionnaire-9SIS: Suicide Intent Scale; KIVS: Karolinska Interpersonal Violence Scale;

2.4.5 Suicide Trigger Scale (STS-3) and Suicide Crisis Inventory (SCI)

The Suicide Trigger Scale v.3 (STS-3) is a 42-item psychological test that includes measures of suicidality, depression affective intensity, impulsivity, and attachment.[35] Yaseen et al. (2014) examined the predictive validity of the STS-3 in high-risk psychiatric inpatients using suicidal behavior (as determined by participant self-report and medical record/death registry review) as the outcome. The predictive validity for the transformed score (⩾ or ⩽ one standard deviation from mean) of the STS-3 was reported for suicidal behavior at 2 – 6 months post-discharge with reported sensitivity of 69.2%, specificity of 68.3%, PPV of 40.9% and NPV of 87.5%.[35]

In a separate study, the authors also reported on the Suicide Crisis Inventory (SCI), a 49-item version of the STS-3 with increased Likert scale response range from three to five points.[36] The SCI was administered to psychiatric inpatients and suicidal behavior at 4 – 8 weeks post discharge was determined by participant self-report at follow-up and medical record review. The SCI at discharge of adults admitted for high suicide risk showed better performance that the SCI

at admission, with sensitivity 63.6% and specificity 88.2% for predicting suicidal behavior 4 – 8 weeks post-discharge using the optimal cut-off score of 114.[36]

2.4.6 Beck Depression Inventory (BDI) Suicide Item

The BDI is a 21-item self-report measure that includes an item (item 9) that assesses the severity of suicidal thoughts rated on a 4-point scale.[23] Green et al. (2015) examined the predictive validity of the BDI suicide item in two samples. In the first sample, using death by suicide for up to 20 years post-assessment as an outcome, the authors reported sensitivity 12%, specificity 99%, PPV 8% and NPV 99% for a score of 3 (the highest score) and sensitivity 28%, specificity 93%, PPV 3% and NPV 99% for scores of 2 or 3. Using death by suicide (0.8% of sample) and repeat suicide attempts (30.3% of sample) at 18 months as an outcome in the second sample, the authors note increased sensitivity and PPV, but lower specificity and NPV at the same thresholds (sensitivity 25%, specificity 88%, PPV 47%, NPV 73% for a score of 3; sensitivity 61%, specificity 60%, PPV 40% and NPV 78% for scores ≥2).[23] This is consistent with expectations: the inclusion of suicide attempts would be expected to increase sensitivity and PPV, and the shorter timeframe to decrease specificity and NPV due to the number of individuals who may not engage in suicidal behaviors until after the time period observed.

2.4.7 Karolinska Interpersonal Violence Scale (KIVS)

The Karolinska Interpersonal Violence Scale is an instrument that measures expressions of and exposure to violence as a child and expressions of and exposure to violence as an adult.[37] Haglund et al. (2016) studied the psychometric properties of the KIVS in participants who presented to hospital with suicide attempts. They used repeat suicide attempt at six months as the outcome of interest and differentiated between violent (all methods except poisoning) and non-violent attempts. Using the identified optimal cut-off of 6 points, the KIVS predicted repeat

suicide attempt with a sensitivity of 62% and specificity of 53%. Using the same threshold, the KIVS predicted repeat suicide attempt using a violent method with a sensitivity of 81% and specificity of 52%.[37]

2.4.8 Columbia Suicide Severity Rating Scale (C-SSRS)

The Columbia Suicide Severity Rating Scale (C-SSRS) is a semistructured, rater-based interview used in the assessment of suicidal ideation and behaviors. The US Food and Drug Administration (FDA) has identified the C-SSRS as the recommended instrument for the prospective evaluation of suicidality in antidepressant clinical trials.[38-40] Sensitivity and specificity of 67% and 76% respectively have been reported for the electronic C-SSRS in predicting subsequent suicidal behavior with a mean follow-up period of 64 days.[39] Another study reported sensitivity and specificity values of 69.4% and 65.2% respectively for the C-SSRS total score in predicting suicidal behavior based on self-report using the C-SSRS at 6 months.[40]

2.4.9 Beck Hopelessness Scale (BHS), Beck Scale of Suicidal Ideation (BSS), and Suicide Cognitions Scale (SCS)

In addition to evaluating the C-SSRS, Madan et al. (2016) reported the sensitivity and specificity of the Beck Hopelessness Scale (BHS), Beck Scale of Suicidal Ideation (BSS) and Suicide Cognitions Scale (SCS) using the same outcome that they used for the C-SSRS described above.[40] The BHS is a 20-item self-report instrument intended to measure negative future thinking.[40] The BSS is a 21-item self-report instrument examining suicidal thoughts and behaviors.[40] The SCS is an 18-item self-report instrument constructed to examine schemas of unbearability and unlovability.[41] The reported sensitivities and specificities were, respectively: 63.3% and 56.1% for the BHS (≥ 13); 58.1% and 65.0% for the BSS (≥ 14); and 58.1% and 67.8% for the SCS (≥ 55).[40]

2.4.10 PHQ-9 Suicide Item 9

The Patient Health Questionnaire - 9 (PHQ-9) is a tool used to assess depression in the preceding 2 weeks; its ninth question asks about the frequency of suicidal/self-injurious thoughts in that timeframe.[28] Simon et al. (2016) investigated the predictive validity of this question by identifying deaths by suicide identified on medical records 7 and 30 days after assessment. The authors note the trade-off in sensitivity and PPV; any positive response to item 9 would have a sensitivity of 64% for identifying death by suicide over 1 month, but a PPV of less than 2%. They note that increasing the threshold to any number greater than 1 would increase the PPV to approximately 3%, but reduce sensitivity to approximately 35%.[28] Madan et al. (2016) also examined the psychometric properties of the PHQ-9 item 9 in their study of the C-SSRS described above. Using the outcome of self-reported suicidal behavior reported on the C-SSRS at 6 months, they report a sensitivity of 86.1% and specificity of 33.1% for the PHQ-9 item 9 (score ≥ 3).[40]

2.4.11 Mini International Neuropsychiatric Interview (MINI) suicidal scale

The suicidality module of the Mini International Neuropsychiatric Interview 5.0.0 (MINI) is a short questionnaire that explores suicide ideation and previous suicide attempts.[24] One study examined the ability of the MINI suicidal scale to predict threats and acts of suicidal behavior or nonsuicidal self-injury in admitted psychiatric patients at 12 months. This was determined using a combination of assessments by a therapist and a nurse and by review of the medical record. Using the high-risk threshold of 10, the scale had a sensitivity of 61% and specificity of 75%, with PPV 43% and NPV 86%. Moderate risk thresholds (≥6) had a sensitivity of 73% and specificity of 62%, with PPV 39% and NPV 88%.[24]

2.4.12 Suicide Intent Scale (SIS)

The Beck's Suicide Intent Scale (SIS) is a 15-item questionnaire designed to examine the factual aspects of the suicide attempt preceding assessment.[25] Using committed suicide (as determined by the Swedish national register) at 10 – 15 years post-assessment as the outcome of interest, Stefansson et al., (2012) identified an optimal cut-off score of 16, which gave specificity of 52% and sensitivity of 100%. The PPV at this threshold was 16.7%.[25]

The authors of the study conducted a follow-up analysis using the same population in which they examined the predictive validity of the SIS and KIVS combined.[42] When combined, the instruments had maintained the high sensitivity (100%) and had a higher specificity (63%) and PPV (18.8%).[42]

2.4.13 Nurses' Global Assessment of Suicide Risk scale (NGASR)

The Nurses' Global Assessment of Suicide Risk Scale (NGASR) is a 15-item scale intended to act as a rapid assessment of suicide risk.[43] When examining suicidality at 6-month follow-up, no significant association with NGASR scores was found.[43]

2.4.14 Suicide Assessment Scale (SUAS)

The Suicide Assessment Scale (SUAS) is made up of 20 items that cover five areas, with each item being rated on a five-point scale.[44] In a study of patients admitted to emergency wards after suicide attempt, the SUAS had a sensitivity of 61% and a specificity of 60% in predicting repetition of suicidal behavior in the following 3 years, as determined by medical and census records.[44]

2.5 Discussion

The results of this review demonstrate that, even though many instruments have been studied in their ability to predict suicidal behavior, the development of an instrument with a high level of clinical utility remains elusive. Multiple factors contribute to the challenge of predicting suicide.

2.5.1 Rarity of suicide as an event

Suicide is a rare event and this has been identified by numerous authors[1,45,46] as a major barrier to the development of an effective instrument. Given the rarity of suicide, even a highly specific test will result in high rates of false positives. This results in low PPVs even in tests with moderate sensitivities and specificities, as evidenced by our results, in which no study reported a PPV for death by suicide >20%. As many of these studies examined high-risk populations, the PPV of instruments would drop even lower when applied to a more general population.

2.5.2 Identification of suicidality

Criterion contamination is a major factor that makes evaluation of suicide assessment tools difficult. In one study, the C-SSRS was evaluated against follow-up assessments using the same instrument.[40] Difficulties in defining suicidality remain another barrier in the development of a tool for detection of those at risk. Many of the papers identified in this review used different criteria to identify those that engaged in future self-harm. Some studies[39,40] used self-reported suicidal behavior at a later time point as their outcome of interest. As a result, individuals who engage in nonsuicidal self-injury may report these behaviors as suicide attempts and be included as a positive outcome despite the lack of suicide intent in the period of interest. Parasuicidal gestures (in which the individual engages in suicidal behaviors without intent to die) interfere with the validity of repeat emergency department presentations as outcomes. This factor is particularly problematic for large-scale studies examining suicide attempts, as more

comprehensive assessment is required in each case to distinguish parasuicidal gestures from true suicide attempts.

Furthermore, death by suicide is likely caused by distinct clinical entities. Yaseen et al. (2014) noted a bimodal distribution in scores amongst the population who later died by suicide.[35] When accounting for both high and low scores on the STS-3, they observed improvement in the predictive validity of this instrument. This suggests that this instrument was identifying, at minimum, two different processes that predispose individuals toward suicidality. Given the potential heterogeneity amongst the populations that go on to attempt/die by suicide, evaluation of clinical instruments remains difficult as they need to identify multiple different processes simultaneously in order to have high predictive validity.

The use of hospitalization in order to prevent imminent risk of suicide also presents problems. Due to ethical considerations, individuals deemed at imminent risk of suicide on assessment by mental health clinicians are admitted to hospital. As a result, any tools using hospitalization as an outcome are inherently limited by the validity of clinician assessment. Studies have attempted to evaluate the predictive validity of clinician assessment on suicide risk;[47,48] however these studies have used future admissions and future assessments using rating scales as outcome measures. This results in criterion contamination, as future hospitalization is influenced primarily by clinician assessment. Admission to hospital in order to prevent suicide may also affect rates of death by suicide, as individuals who present to hospital with a high number of apparent risk factors are likely to be admitted, confounding the use of death by suicide (and excluding hospitalized patients) as an outcome measure. As the decision to admit a patient to hospital is dependent on the clinician assessing the patient, as well as the availability of community

resources, the use of admission to hospital as a surrogate measure has limitations as well. The lack of an outcome measure that can be intrinsically linked to suicide risk poses a major problem in developing new instruments. Even if a new tool were to outperform currently accepted instruments, its validity as determined through statistical measures would be limited by the validity of other instruments, and attempts to improve the new tool's performance in comparison to current instruments may, in fact, negatively impact its clinical performance.

2.5.3 Complex interplay between dynamic factors

Suicide is a complex behavior influenced by numerous different biological, psychological, interpersonal, socioeconomic and environmental factors that may all interact bidirectionally both with each other and with the outcome(s) of interest. This poses problems for both data collection and data synthesis. As evidenced when comparing data from Stefansson et al. (2012) and Stefansson et al. (2015), the combination of different assessment tools may provide more accurate predictive data than a tool in isolation, likely because the added tool measures factors not assessed in the first.[25,42] Assessments exploring every single factor of an individual's current and past life situation may be too cumbersome to be used on every clinical visit, but exclusion of certain items, even those that have no clear link to risk, may interfere with accurate prediction of suicide risk owing to the effect this may have on the relative contribution of known risk factors. The complex relationships between different factors makes determination of the relative weight and effect of each factor very difficult. For example, an individual who is unemployed with no familial relationships may be at lower risk if they find satisfaction and purpose through a recreational interest or volunteer activity, but the presence or absence of fulfilling recreational or volunteering activities may have minimal bearing on risk in an individual who is stably employed and/or has stable family supports.

The dynamic nature of individual risk factors, especially interpersonal, socioeconomic and environmental factors, is another hurdle in the accurate assessment of suicide risk. Occupational, relationship and health status have all been identified as factors affecting an individual's suicide risk.[49] As these factors can change unpredictably over short periods of time, the use of clinical instruments to predict suicide risk after a period as short as 3 months may have limited practical use. Furthermore, intervention through medical and psychotherapeutic treatment may affect an individual's risk, and their engagement and benefit may also be affected in part by the clinical services to which they are referred.

Ultimately, suicidality may be better conceptualized as a *dynamic* state of being that is influenced by internal and external factors, both of which may be the subject of evaluation. Therefore, improvement in suicidality risk assessment may require the implementation of continuous evaluation of risk factors over time.

2.5.4 Reliance on self-report

The complex interplay between different factors is further complicated by the reliance of clinicians on patient self-report. Individuals brought to medical attention by external parties may minimize their symptoms due to cognitive distortions or in order to attempt suicide without interference. While studies have attempted to measure implicit thoughts in order to account for this factor[50,51], it is nonetheless possible for an individual to respond disingenuously to all questions in order to avoid further monitoring by medical services. The self-report component is also complicated by individuals who overendorse symptoms, such as in cases of malingering or personality disorders.

2.5.5 Application of new technology

Given the number of complicating factors as described above, it may be concluded that instruments administered through conventional means are limited in their ability to accurately assess an individual's suicide risk. New technologies may address some of these factors, however. Repeated assessment over time may ensure that dynamic factors are accurately incorporated in the determination of an individual's risk. The trajectory of an individual's symptoms may also have predictive applications, as evidenced by the results of Ayer et al. (2008).[32] Repeated assessment of dynamic factors in combination with data on static factors collected only once may ensure that changes are appropriately accounted for while minimizing the length of any given assessment.

The use of electronic biomarkers and monitoring of physiologic states is increasingly of interest for the assessment of mental illness.[52] These tools may allow clinicians to overcome the reliance on self-report in suicide risk assessment, as well facilitate continuous measurement of dynamic factors over time.

Machine learning and specifically deep learning, a subset of machine learning with its foundation in neural networks, has played an evolving role in health informatics.[53] These tools rely on large volumes of data for the developement of accurate predictive models. The use of data-driven models to predict short term behavior may allow for a substantially more accurate assessment of suicide risk. As computational power continues to improve, it may be possible to incorporate the numerous dynamic factors, including those monitored using the tools described above, to develop an evolving model of suicide prediction.

2.6 Conclusion

Existing evidence indicates limited clinical utility in the use of currently used rating scales as tools for suicide prediction. Many barriers exist in the development of an accurate, practical tool for suicide prediction, including the low prevalence of suicide even in populations identified as high-risk, the complex and dynamic interactions between risk factors, and the reliance on self-reported symptoms. New technologies may allow us to overcome some these barriers through the use of repeated assessments, physiologic monitoring and data-driven models to create improved tools of suicide prediction.

2.7 Acknowledgments

ECC received two Rudelle Hall Graduate Scholarships (2018 and 2019/20) which supported his work on the content of this chapter.

CHAPTER 3: ASSESSMENT AND MONITORING OF BIPOLAR DISORDER USING MOBILE APP-BASED SELF-REPORT QUESTIONNAIRES: A SYSTEMATIC REVIEW

Authors

Eric C Chan[1], Sudhakar Sivapalan[1]

Institutions

[1]Department of Psychiatry, University of Alberta, Canada

3.1 Abstract

Background: Bipolar disorder is a chronic, progressive illness characterized by recurrent episodes of mania and depression. Self-report scales have historically played a significant role in the monitoring of bipolar symptoms. However, these tools are reliant on episodic memory, which can be unreliable, and do not allow the clinician to monitor brief episodic symptoms or the course of symptoms over shorter periods of time. Mobile app-based questionnaires have been suggested as a tool to improve monitoring of patients with bipolar disorder.

Objective: To determine the feasibility and validity of mobile app-based self-report questionnaires.

Methods: Pubmed, Ovid Medline and Embase databases were searched for papers published in English assessing the validity of mobile app-based self-report questionnaires. Relevant studies were identified and results examining the validity and rates of adherence using app-based self-report questions were compared.

Results: Six records were identified for inclusion in this review. Of these, only two compared app-based self-report questionnaires with standardized assessment tools. Self-report questionnaires collected via mobile device appear to be valid tools in detecting the depressive phase of bipolar disorder, but the ability of these instruments to detect the elevated mood phase of the illness is still unclear. One study suggests app-based assessment tools may differentiate patients with bipolar disorder from healthy controls, but data are limited.

Conclusions: Limited data suggest that app-based assessment tools for monitoring bipolar disorder may have validity in monitoring symptoms of the depressive phase of bipolar illness. It is still unclear whether these tools are able to accurately assess symptoms of mania or monitor symptoms of patients in an acute mood episode. One study suggests app-based assessment tools may differentiate patients with bipolar disorder from healthy controls, may be more accurate than paper-and-pencil based interventions, and may contribute to improvement in depressive symptoms. Given the very limited number of studies included in this review and the increasing use of digital health tools, further studies exploring the potential utility of mobile apps in bipolar disorder are indicated.

3.2 Introduction

Bipolar disorder is a chronic, progressive illness characterized by recurrent episodes of mania and depression. The international 12-month prevalence of bipolar I disorder is 0.0% to 0.6% and international 12-month prevalence of bipolar II disorder is 0.3%.[54] Both manic and depressive episodes are associated with impairments in social and occupational functioning and the World Health Organization's World Mental Health surveys identified the disorder as having the second strongest effect on days out of role when compared to other common physical and mental

illnesses.[55-58] In addition, bipolar disorder is associated with a high risk of suicide, with one third to one half of patients attempting suicide at least once in their lifetime and 15-20% of attempts completed.[59] Given such adverse consequences of mania and depression, timely detection of relapse is an important aspect in the psychiatric care of the disease.

No biomarker has been approved for the diagnosis or assessment of bipolar disorder, and so medical practitioners must rely on clinical assessment and reports from the patient and collateral sources in order to monitor the disease. Detection of mood episodes can be delayed, however, with previous data indicating that the interval between illness onset and hospitalization is often 3 weeks or more.[60] One challenge for the detection of mood episodes is the lack of insight that can occur in patients with bipolar disorder, especially during episodes of pure mania.[61] Previous data suggests, however, that some patients in acute mania may retain awareness of their diagnosis and its potential consequences despite having impaired insight into their current symptoms.[62] Given the preserved awareness of their diagnosis even in the context of active symptoms, the use of *self-report* questionnaires has the potential to facilitate symptom monitoring, including changes over time.

Self-report scales such as the Mood Disorder Questionnaire (MDQ) and the Altman Self-Rating Mania Scale (ASRM) have previously been developed for use in the monitoring of bipolar symptoms. These scales have been validated in inpatient populations with bipolar disorder, with respective sensitivities and specificities of 86% and 71% for the MDQ and 93% and 33% for the ASRM.[63-65]

Traditionally, self-report scales have been administered via pen-and-paper; however, some limitations exist with this form of data collection. When administered in the context of visits with

a health care provider, these tools rely on retrospective reporting of symptoms, which can be unreliable, and do not allow the clinician to monitor symptoms associated with brief mood episodes or the course of symptoms over shorter periods of time.[65-67] In a study asking participants to complete paper diaries on a daily basis, participants were found to record entries outside of the requested timeframe, and to inaccurately report the date of these entries, reducing the accuracy of data collected.[68] In addition, the frequency that the clinician is able to review responses obtained via pen-and-paper is limited by the frequency in which the responses are forwarded to the provider. This often occurs on clinic visits, which limits the ability of the health care provider to respond in a timely fashion if the patient deteriorates between scheduled appointments.

The administration of self-report scales using mobile applications has the potential to circumvent some of these issues. Automatic transmission of data using a mobile device could allow clinicians to monitor symptoms in real-time, improving their ability to proactively detect and engage the patient when symptoms relapse. In addition, scale administration using a mobile application may be less disruptive for the patient, increasing the frequency that the patient is willing to complete the scale. For example, one study described a mobile application for monitoring non-affective psychosis that yielded more data points and took less time compared to their SMS text-only equivalent.[69] The increased data collection afforded by the use of mobile applications may also have uses in research settings. Frequent administration of scales may allow researchers to better characterize the course of illness over time and to identify warning signs that mark early deterioration.

Given the variability in the course of symptoms in bipolar disorder, the use of mobile applications in this population has been of considerable recent interest, with 35 apps identified using the Google Play and iOS stores in a recent systematic review.[70] Studies have shown that 60 - 70% of patients with mental illness would be interested in using a mobile application to monitor their mental health condition.[66,71,72] A 2015 review has shown, however, that 60% of symptom monitoring apps available did not use validated screening measures.[70] Furthermore, it is possible that for a given validated screening tool, data collected via a mobile app may differ from that collected via a pen-and-paper version.

The validity of a scale is defined as "the extent to which an instrument indeed measures the latent dimension or construct it was developed to evaluate."[73] It differs from reliability, which evaluates whether data collected is consistent when the measure is repeated under comparable conditions.[73] The major forms of validity are content validity, criterion validity and construct validity. Content validity refers to whether the measure adequately assesses the domain of interest and is primarily assessed through evaluation by experts and the target population. Criterion validity refers to whether the results of a measure relate to another measure of relevance. It includes predictive validity (the ability of the measure to predict a future result or answer a future question) and concurrent validity (the strength of the relationship between the new measure and a gold standard measurement made at a similar time). Construct validity refers to the degree that the measure assesses the construct of concern. Construct validity can be evaluated through convergent validity, discriminant/divergent validity, differentiation or comparison between known groups, or correlational analysis.[73]

The aim of this systematic review was to assess the feasibility and validity of mobile applications as tools for bipolar symptom monitoring through a systematic review of the literature. We identified studies in which patients with bipolar disorder were monitored using self-report scales administered by mobile application, with or without comparison to a traditional form of symptom monitoring such as pen-and-paper based rating scale or standardized clinician interview. The outcomes of interest in this review were adherence rates and the comparison of mobile application-based assessment to other methods of data collection, where available.

3.3 Methods

In order to identify data describing the feasibility and validity of mobile applications in the assessment of bipolar disorder, we conducted searches of the Pubmed, Ovid Medline and Embase databases. One researcher (EC) searched these databases using the keywords ("mental disorders", "psychiatry" or "mental health") and ("mobile application", "cell phone" or "smartphone"), excluding the term "substance-related disorders". All records published in English listed from database creation to April 11, 2018 were identified. In addition, the list of citations of records identified for inclusion were reviewed in order to identify other potential candidates for inclusion.

EC and SS independently screened the records to identify articles suitable for inclusion in this review. In the case of disagreement between the two authors, records were forwarded to the next step of screening for further evaluation of suitability. There was no disagreement between authors following review of full papers.

Titles of records resulting from the database search were screened and excluded using the following criteria: Titles and abstracts of records were screened using the following exclusion

criteria: the study did not refer to symptom assessment by mobile application, smartphone, mobile phone/technology as the primary intervention of interest or the intervention of interest was solely text-message based; bipolar disorder was not the primary condition of interest; the primary intervention was focused on the reduction of symptoms, including administration of psychotherapy; the study referred to an intervention based on physiological or behavioral monitoring; the study did not present data from an applied intervention (such as a protocol paper, review paper, or response/correction to another article).

The full text of the remaining studies were evaluated and studies were excluded if they met one of the following criteria: The study did not present data from an applied intervention (such as a protocol or review paper); the primary intervention was focused on the reduction of symptoms, including administration of psychotherapy; the study referred to an intervention based on physiological or behavioral monitoring; the study did not refer to symptom assessment by mobile application, smartphone, mobile phone/technology as the primary intervention of interest or the intervention of interest was solely text-message based; bipolar disorder was not the primary disease of interest.

Of note, we excluded studies in which the mobile application involved a significant psychotherapeutic component as the greater degree of participant involvement required could affect adherence rates and reported data.

Studies identified for inclusion in this review were then evaluated for data on adherence rates and validity of mobile application-based symptom monitoring tools with or without comparison to standardized pen-and-paper or clinical interview-based measures. One researcher (EC) assessed each of the identified studies for bias using the “Cochrane Risk of Bias 2” tool or

the “Cochrane Risk of Bias in Non-randomized Studies – of Interventions” assessment tool. These tools have been developed for the assessment of bias in randomized and non-randomized studies respectively.[74,75] These assessments were reviewed by the other author (SS) and are available as supplementary information.

3.4 Results

3.4.1 Identified Records

The flow diagram of the search method is depicted in Figure 2. Initial searches produced 1061 unique records following removal of duplicates. Thirty-one records were identified following screening of abstracts and their references were also searched for further relevant studies. Following the search procedure described above, 6 records were identified for inclusion in this review; study characteristics are listed in Table 3. Findings of each study are listed separately (Table 4).

3.4.2 Data on Validity

Two papers identified for inclusion provided data on the validity of mobile assessment tools, both compared against clinical assessments.[76,77] Measures of mood collected via the mobile tool negatively correlated with depression rating scales collected using conventional methodology; Montgomery Åsberg Depression Rating Scale (MADRS) (Pearson’s correlation coefficient, $r = -0.567$), Hamilton Depression Rating Scale-17 (HDRS-17) (regression coefficient, $\beta = -0.058$).

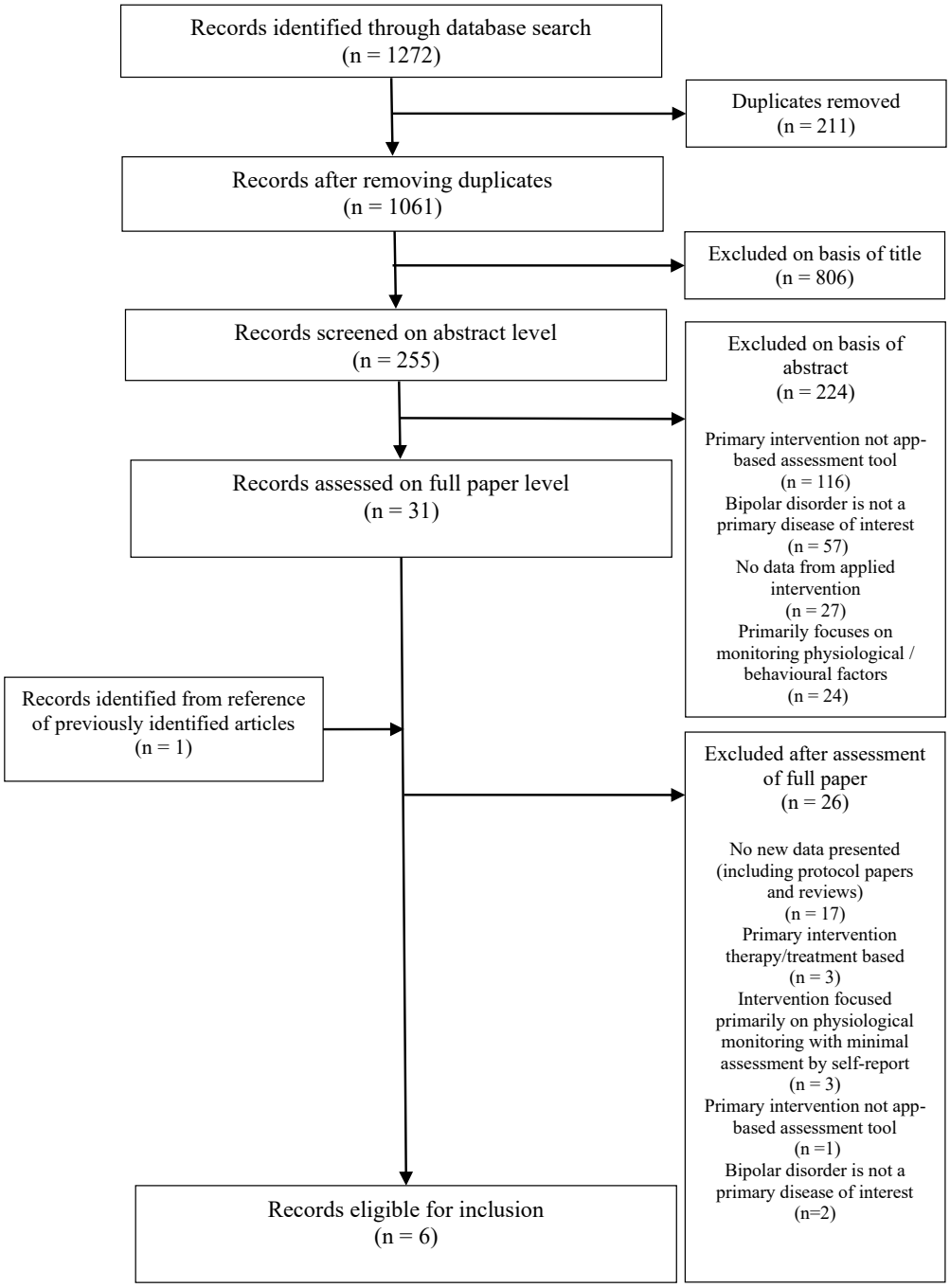

Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram

Table 3. Characteristics of included studies

Reference; Location	**Number of Subjects**	**Mobile Application-based Intervention**	**Comparison (if applicable)**	**Duration**
Depp et al. (2012)[76] San Diego, USA	18 (int) 22 (comp)	9-point bipolar anchored scale completed twice per day. Could only be completed within 2 hours after alert to complete measure.	Daily paper-and-pencil mood charts MADRS and YMRS completed at baseline, 6 weeks and 12 weeks post-baseline.	12 weeks
Faurholt-Jepsen et al. (2015)[77] Copenhagen, Denmark	30	MONARCA: Self-monitoring of 11 symptoms completed daily. Allowed for retrospective data entry up to two days later.	Monthly clinical assessment via HDRS-17 and YMRS. Scores compared to those obtained via app from day of assessment and three previous days.	6 months

Depp et al. (2015)[78] San Diego, USA	51 (int) (41 analyzed) 53 (comp) (41 analyzed)	PRISM: 10 questions followed by rating of current mood state on a 9-point bipolar anchored scale completed twice per day.	Daily pencil and paper mood charts.	10 weeks
Faurholt-Jepsen et al. (2015)[79] Copenhagen Denmark	39 (int) 39 (comp)	MONARCA: Self-monitoring of 11 symptoms completed daily. Allowed for retrospective data entry up to two days later.	Participants provided a smartphone without the MONARCA system.	6 months

Hidalgo-Mazzei (2016)[80] Barcelona, Spain	51	Short 5-item screening tests completed daily. Weekly Yes/No questions for DSM-5 criteria of manic and depressive episodes.	None	3 months
Schwartz (2016)[81] Pennsylvania, USA	10 (bipolar I or II) 10 (healthy controls)	4 items on visual analog scale and 1 item on Likert scale completed twice per day.	None	2 weeks

Int: intervention group; comp: comparison group

Table 4. Summary of findings on mobile app use in bipolar symptom monitoring

Reference; Location	Completion rates	Correlation between data obtained via mobile application and comparator	Risk of Bias
Depp et al. (2012)[76] San Diego, USA	Int: 42.1% Comp: 82.9% $t(35) = 5.8$, $P < .001$	Mood ratings: Int: MADRS: $r = -0.567$, $P = .014$ YMRS: $r = 0.294$, $P = .236$ Comp: MADRS: $r = -.243$, $P = .346$ YMRS: $r = 0.452$, $P = .069$	Patients with recent substance use disorder, hospitalization, MADRS >32 or YMRS >20 were excluded. Completion rates: The frequency of scale completion differed between groups. The mobile phone condition was time-limited but the paper-and-pencil condition was not, which could also lead to retrospective data entry. The authors report on both compliance rates including and excluding the missing data from 3 participants in the comparison group. Reported symptoms: Mean mood ratings over study duration and over the first 6 weeks were used in comparison to MADRS and YMRS, however the MADRS and YMRS were conducted at discrete points in time, weakening potential correlation of individual measurements in both mobile phone and paper-and-pencil conditions. These factors were felt to have a potential substantial impact on the results.

Faurholt-Jepsen et al. (2015)[77] Copenhagen, Denmark	No data on completion rates	Mood: HDRS-17: β = -0.058, P < .001 YMRS: β = 0.039, P < .001 Sleep: HDRS-17: β = 0.02, P = .21 YMRS: β = -0.047, P = .026 Activity: HDRS-17: β = -0.042, P < .001 YMRS: β = 0.048, P < .001 Stress: HDRS-17: β = 0.046, P < .001 YMRS: β = 0.012, P = .35	Participants lacking the technical knowledge to use a smartphone and with HDRS or YMRS score >17 were excluded. In addition, users completing the measure shortly prior to clinical interview with HDRS/YMRS may provide similar responses to show consistency even if their symptoms have changed in the interim. Overall, these limitations were felt to potentially have moderate impact on the results.
Depp et al. (2015)[78] San Diego, USA	Int: 65% Comp: 83%	No data on correlation between measures	Patients with recent substance use disorder, hospitalization, MADRS >32 or YMRS >20 were excluded. The frequency of scale completion differed between groups. The mobile phone condition was time-limited but the paper-and-pencil condition was not, which could also lead to retrospective data entry. These factors might have a significant impact on relative completion rates.

Faurholt-Jepsen et al. (2015)[79] Copenhagen Denmark	Int: 93.03% (7.15% done retrospectively)	No data on correlation between measures	Participants lacking the technical knowledge to use a smartphone and with HDRS or YMRS score >17 were excluded. These factors might have a moderate impact on completion rates; however if completion rates were to decrease in depressed and manic episodes, this might also assist practitioners in intervening.
Hidalgo-Mazzei (2016)[80] Barcelona, Spain	88% completion rate 74% users actively using application after 3 months	No data on correlation between measures	Participants with IQ <90, HDRS ⩾8, YMRS ⩾6, or without the requisite technical skills were excluded. A specific brand of smartphone was required for study inclusion, potentially causing socioeconomic status to be a confounder. The authors report on demographic data suggesting this may not be the case, but do not stratify participants by income. Users also received a brief psychoeducational message after measure completion; this positive feedback may encourage increased compliance. These factors were felt to have a moderate impact on rates of scale completion.

Schwartz (2016)[81] Pennsylvania, USA	Bipolar: 95% Controls: 88% P = .68	14-day mean of mood: Bipolar median: 48.6 Control median: 53.2 P = .043 14-day mean of energy: Bipolar median: 44.7 Control median: 52.1 P = .007 14-day range of mood: Bipolar median: 48.0 Control median: 32.5 P = .043 14-day range of thoughts: Bipolar median: 59.5 Control median: 26.5 P = .002 14-day range of impulsivity: Bipolar median: 76 Control median: 28.5 P = .005	Bipolar participants could be in any mood state at study entry. While healthy controls did not have a personal or family history of psychiatric illness, it is unclear how they initially entered the research program. It was also unclear how bipolar patients and controls were selected from the research pool and how the number of participants was chosen. Discomfort using smartphone technology was an exclusion criterion. Impact of above factors difficult to predict as potential confounders from the study population was unclear.

Int: intervention group; comp: comparison group

Depp et al. did not find a significant correlation between self-reported mood and the Young Mania Rating Scale (YMRS) ($r = 0.294$)[76], while Faurholt-Jepsen et al. did find a significant positive correlation ($\beta = 0.039$, $P < .001$).[77] In addition, self-reported mood appears to correlate specifically with the "Elevated Mood" item in the YMRS but does not appear sufficient to accurately detect manic episodes on its own.[77] Mood ratings collected using paper-and-pencil tools showed no correlation with either the MADRS ($r = -0.243$) or YMRS ($r = 0.452$).[76]

Sleep, activity and stress levels were also compared with the HDRS-17 and YMRS.[77] Sleep had a non-significant correlation with the HDRS-17 ($\beta = 0.02$) and a weak, but statistically significant, correlation with the YMRS ($\beta = -0.047$, $P < .026$). Activity level was negatively correlated with the HDRS-17 ($\beta = -0.042$, $P < .001$) and positively correlated with the YMRS ($\beta = 0.048$, $P < .001$). Stress level had a statistically significant positive correlation with the HDRS-17 ($\beta = 0.046$, $P < .001$) and a non-significant correlation with the YMRS ($\beta = 0.012$).[77]

One group compared reported symptoms between patients with bipolar disorder and healthy controls.[81] Statistically significant differences were noted in the 14-day mean of mood and energy. The authors also noted a statistically significant difference in range of mood, thoughts and impulsivity.[81]

3.4.3 Data on Adherence

Varying levels of adherence to reporting protocol were reported among different studies, ranging from 42% to 95%.[76,78-81] Compliance rates were substantially higher for the paper-and-pencil conditions in the two studies reported by Depp et al. However, the frequency of measure completion was not the same between the two groups and the paper-and-pencil condition could complete the measure at any time, whereas the phone condition was time-limited.[76,78] These

differences may have contributed substantially to the differences in completion rates between conditions.

Between studies, the variability in adherence levels may also be due in part to the differences in applications administered (such as length and frequency of assessment or presence of reminders). The lowest adherence occurred in two studies that included twice daily assessments over periods of 10 and 12 weeks.[76,78] The only other study with twice daily assessments occurred over only fourteen-days and all other studies involved daily questionnaires.[81]

3.5 Discussion

In this review, we found two studies comparing data on symptoms collected via smartphone-based symptom monitoring tools to conventional measures (scale administered by a clinician in both cases). Both studies suggest that mobile assessment tools may be accurate in detecting depressive symptoms, but disagree on the ability of mobile assessment tools to detect manic states. As patients in manic/hypomanic states are less likely to retain insight into their clinical state, it is possible that data obtained via self-report during these episodes will be less accurate. Of note, the authors of the study in which no statistically significant correlation was found had compared YMRS scores to data collected over the entire study duration and to that collected during the first 6 weeks of the study.[76] As the YMRS assesses symptoms over the preceding 48 hours, the poor correlation may be at least partly attributable to the difference in time periods observed. The other study, which did find a statistically significant correlation between measures, compared YMRS scores to data collected over the preceding three days.[77] This may be

a more appropriate comparison, especially as one goal of application-based self-report scales is the detection of acute mood states and changes in symptoms over time.

Furthermore, data collected via the paper-and-pencil condition did not have a statistically significant correlation with either the MADRS or YMRS.[76] This suggests that application-based self-report scales may more accurately collect data on depressive symptoms when compared to the paper-based counterpart. While there are few data comparing mobile assessments with validated rating scales, it has been suggested elsewhere that participants may be more forthcoming when reporting symptoms through mobile assessments.[16] In addition, it has been shown that participants completing measures via paper-and-pencil may complete the entries retrospectively and hence, outside the specified timeframe being assessed.[68] This may explain the seemingly increased accuracy of symptoms reported via application-based measure when compared to paper-and-pencil.

It is unclear if any patients developed an acute mood episode during the course of these studies, especially since a manic or depressive episode at study onset was an exclusion criterion for the majority of the studies identified. The one study in which participants could be in any mood state at inclusion did observe differences in reported symptoms between patients with bipolar disorder and healthy controls.[81] It is not clear, however, if this is due to the presence of acute mood episodes or differences in baseline symptoms between the two groups. As such, while these data suggest mobile application self-report tools can detect depressive symptoms in euthymic patients, it is unclear if these findings can be extended to patients experiencing acute mood episodes.

The relationship between protocol adherence and assessment frequency suggests that users may have difficulty completing multiple assessments per day but are able to manage assessments provided they occur once daily. The time and attention required to complete assessments and the timing of their administration may also play a role in adherence rates as longer or more complex assessments may be perceived as more intrusive, especially if administered within a restricted time-frame. As noted above, previous data have suggested that study participants completing a paper-and-pencil based measure may complete entries outside the specified timeframe. As such, it is possible that studies evaluating adherence to paper-and-pencil based measures may overestimate rates in this population.

As discussed above, removal of retrospective bias may provide a more accurate picture of the patient at a specific moment, allowing the clinician to better observe changes in symptoms over time. One study identified in this review observed significantly greater reductions in depressive symptoms over time in the mobile intervention when compared with paper-and-pencil.[78] While the reason for this difference is uncertain, a previous study on measurement-based care versus standard care for major depression had shown improvements in response and remission times in the population administered measurement-based care compared with treatment as usual.[82] This may contribute to the difference observed by Depp, et al.,[76] as the mobile intervention was monitoring their symptoms twice as frequently as the paper-and-pencil condition.

We also note that the two papers comparing mobile application-based assessment to standardized measures were assessing the criterion validity of the new tool.[76,77] One limitation of this approach to validity assessment is the appropriateness of the criterion variable.[73] It has

previously been posited that current syndrome-based psychiatric diagnoses do not adequately correspond to specific underlying pathophysiological mechanisms.[83] The criterion measures used in the relevant studies aim to detect psychiatric illness in accordance with these criteria and consequently bear similar limitations. One study evaluated the ability of mobile assessment-based tools to differentiate between known groups (a form of construct validity) and identified differences in the mean and range of symptoms related to bipolar disorder.[81] These variables could not be assessed with the same degree of resolution using clinician-driven measures, due to the limitations in assessment frequency. These variables may, however, improve our ability to detect and understand bipolar illness.

3.6 Limitations

In this review, only English studies from peer-reviewed journals were considered. As very few (n=23) non-English papers were identified prior to screening, this was felt to have minimal impact on overall results. As there were large numbers of protocol papers identified for which it is not possible to exclude unpublished data, it is also possible that publication bias may have resulted in many missed negative findings. A major limitation is that only two of the studies identified report data comparing mobile application-based self-report to conventional methods. In addition, of the six studies included in this review, two groups (one using the PRISM application and the other using the MONARCA system) were both represented twice. This may contribute to bias in the overall findings and future reviews including studies from a greater number of different groups using different applications are indicated. Almost all studies included discomfort with technology and elevated scores on clinical interviews as exclusion criteria. This may impact the generalizability of these findings to patients with more severe illness or cognitive

impairment. In addition, it is unclear to what extent these findings apply to patients in the midst of an acute manic or depressive episode.

3.7 Future Research

Further studies on the validity of mobile application-based assessment tools, especially those evaluating the ability of these tools to detect acute mood states, will better inform us about the potential utility of these tools in clinical settings. Further refinements in the administration of applications for monitoring bipolar disorder may also allow for a greater degree of adherence and accuracy in data collection. Future research into the optimal frequency and length of assessments, as well as into factors such as timing of administration and ease of use, could contribute substantially to the use of applications as symptom-monitoring tools. The use of repeated self-report questionnaires when combined with physiological and behavioral monitoring as well as other biomarkers also bears further investigation and may contribute to our understanding of bipolar disorder.

3.8 Conclusions

While it is difficult to draw firm conclusions due to the limited amount of data available, app-based assessment tools for monitoring bipolar disorder appear to have validity in monitoring symptoms of the depressive phase of illness when compared with conventional measures. It is unclear whether these tools are able to accurately detect symptoms of mania or hypomania, and further study is necessary to determine the suitability of these tools for this purpose. In addition, further research into the accuracy of data collected when the patient is experiencing an acute mood episode is needed. Data from one study suggest mobile applications have utility in

differentiating individuals with bipolar disorder from healthy controls. Further studies exploring these findings are indicated.

3.9 Acknowledgments

ECC received two Rudelle Hall Graduate Scholarships (2018 and 2019/20) which supported his work on the content of this chapter.

CHAPTER 4: INTERNAL CONSISTENCY AND CONCURRENT VALIDITY OF A NEW TOOL FOR THE ASSESSMENT OF SUICIDALITY, THE SUICIDE IDEATION AND BEHAVIOR ASSESSMENT TOOL (SIBAT)

Authors

Eric C Chan[1], Keanna Wallace[2], Esther Yang[2], Leslie Roper[2], Garima Aryal[2], Andrius Baskys[1,3,4], Richard Isenberg[5], Patrick J Carnes[1], Bradley Green[6], Katherine J Aitchison[1,2,7]

Institutions

[1]Department of Psychiatry, University of Alberta, Canada; [2]Department of Medical Genetics, University of Alberta, Canada; [3]Graduate College of Biomedical Sciences and University Medical Center, Western University of Health Sciences, USA; [4]Memory Disorders and Genomic Medicine Clinic, Riverside, USA; [5]Psychological Counseling Services, Ltd., Scottsdale, USA; [6]Department of Psychology, University of Texas at Tyler, USA; [7]Neuroscience and Mental Health Institute, University of Alberta, Canada.

4.1 Abstract

Objectives: The aim of this study was to assess the internal consistency of self-report components of the Suicide Ideation and Behavior Assessment Tool (SIBAT) and validate it with relevant elements of the Mini International Neuropsychiatric Interview (MINI).

Methods: The SIBAT is a newly developed instrument for the evaluation of suicidality. It is a comprehensive tool made up of both clinician-assessed and self-report components. In this study, we invited university students and trainees participating in a study of addictions to complete the

self-report component of the SIBAT as an add-on study. We then evaluated the internal consistency of the self-report component of the SIBAT and validated it against the suicidality component of the MINI. Data were analysed using both complete case analysis and multiple imputation.

Results: SIBAT data were collected for 394 participants, 314 of whom had also completed the MINI at the time of data analysis. The internal consistency of module 5 of the SIBAT was high (Cronbach's $\alpha = 0.87$, n = 4 items). Item-total correlations ranged from 0.22 – 0.79 for module 2 and -0.20 – 0.79 for module 3, with the majority of items in both modules having item-total correlations > 0.4. Each item from module 5 had a statistically significant association with the corresponding item from the MINI. The sum of scores from modules 2 and 3 had a moderate correlation with the assessment of suicide risk determined by the MINI (Spearman's rho = 0.44, $P < 0.001$), which assesses suicidality in the preceding month, and a strong correlation with the total score of SIBAT module 5 (Spearman's rho = 0.62, $P < 0.001$), which assesses suicidality in the preceding week. The median time to completion of modules 2, 3 and 5 was 14.3 minutes.

Conclusions: Our findings suggest that the self-report component of the SIBAT has good internal consistency overall and that items in module 5 and the total score of modules 2 and 3 have good concurrent validity with the MINI.

Keywords: e-mental health, Psychometric, Rating Scale, Risk assessment, Self harm, Suicide, Validation, Mobile applications

4.2 Introduction

The evaluation of suicide risk is an important aspect of psychiatric care. Data from the Public Health Agency of Canada indicate that 11.8% of Canadians report thoughts of suicide in their lifetime and 2.5% of Canadians report having thoughts of suicide in the past year.[84] Many completed suicides are preceded by contact with health care services. A 2011 study of suicide in Alberta reports that 58% of suicides were preceded by an emergency department visit and 28% were preceded by an inpatient hospital separation (i.e. discharge) in the preceding year, with the most common diagnoses being related to mental and behavioural disturbance or to injury and poisoning.[85] Given the frequency of contact with health services prior to suicide, accurate identification and risk stratification is an important step in ensuring that appropriate interventions are provided to those at elevated risk.

Suicide risk evaluation can be divided into two components: screening and risk assessment.[6] Screening is the process in which individuals at greater than negligible risk are identified for further assessment and possible intervention.[6] A screening process should optimally be easily administered, so that it can be applied to large populations, and have high sensitivity, in order to minimize the number of at-risk individuals being inappropriately ruled out (false negatives). Risk assessment involves a more in-depth evaluation to identify those at true risk who require further intervention.[6] Optimally, risk assessment will allow the care provider to stratify risk so that interventions appropriate to the nature and degree of risk can be offered.

Numerous instruments have been developed for the prediction of suicide risk. The National Action Alliance for Suicide Prevention in the United States has recommended that all patients identified as being at risk of suicide be assessed using a standardized instrument or

scale.[86] There is no consensus, however, on the best screening and assessment tools for use in clinical settings.[86] Multiple recent studies evaluating the utility of currently available instruments indicate that none predicted suicide or suicidal behaviour with sufficient accuracy to be relied on in clinical settings.[1-4] Similarly, the third edition of the American Psychiatric Association Practice Guidelines note that, while suicide assessment instruments may have clinical utility in assisting the clinician to develop a thorough line of questioning, no scale has been shown to produce a clinically useful score for suicide prediction.[5]

One of the major factors limiting the predictive validity of tools for suicide risk assessment is the rarity of suicide as an event. As Pokorny and Rosen have previously discussed, even a tool with high sensitivity and specificity can have a low positive predictive value given the infrequency of suicide in the general population.[45,46] Estimation of suicide risk is further complicated by the multifactorial nature of suicidality. Suicide is influenced by a wide variety of biological, psychological and social factors.[49,87] Some of these factors, such as workplace-related factors, physical health and family connectedness, can change drastically over short periods of time. These factors, their changes over time, and the interplay between them, may be important considerations in suicide risk stratification. Furthermore, recent data suggest that suicidality itself can fluctuate over the course of days and possibly even hours,[88,89] complicating assessment considerably. Given the limitations of currently available instruments for suicide risk prediction, the National Action Alliance for Suicide Prevention in the United States identified the "development of validated procedures that can determine the degree of suicide risk (e.g., imminent, near-term, long-term)" as an aspirational goal in the prevention of suicide.[6]

The Suicide Ideation and Behavior Assessment Tool (SIBAT) is a new, comprehensive rating scale developed for the assessment and monitoring of suicidality by a group of clinical trial and academic experts in scale development. It is a comprehensive tool made up of both clinician-assessed and self-report components and previous data indicate that it is sensitive to changes over time.[7-12] Our goal in this study was to assess the internal consistency of the SIBAT and cross-validate it with the Mini International Neuropsychiatric Interview (MINI), in order to evaluate the concurrent validity of responses obtained using this new measure.

4.3 Methods

4.3.1 Assessment Tools

4.3.1.1 Suicide Ideation and Behavior Assessment Tool (SIBAT)

The self-report component of the SIBAT is comprised of five modules. The first module is only completed on the first administration of the scale as it assesses static risk factors such as demographic information and history prior to the initial assessment (information on previous suicide attempts/behaviours, family history of suicide, history of abuse/neglect, etc.). Module 4 assesses recent suicidal behaviours using the same questions that are in module 1 and is only completed on repeat administrations of the SIBAT. As such, module 4 was not administered due to the single administration of the SIBAT in this study. Module 2 assesses risk/protective factors such as mood, anxiety, hopelessness and substance use. Module 3 assesses thoughts related to factors associated with suicide risk including thoughts of dying, reasons for living, and hope for the future. Both modules 2 and 3 are rated on 6-point Likert scales. Module 5 is made up of 5 items that directly ask the individual to rank their suicidal desire, intent, thoughts and the

likelihood they would commit suicide on 5-point Likert scales. Examples of question stems from the SIBAT are included in Appendix 1. As items in modules 1 and 4 are categorical responses and not scored, they were excluded in the comparison of SIBAT and MINI responses.

The SIBAT was administered using either an application (app) developed using the AppSheets platform (Figure 3) or the Qualtrics interface.

4.3.1.2 Suicidality Component of the Mini International Neuropsychiatric Interview

The suicidality component of the MINI consists of 19 Yes/No items and categorizes the participant into low, medium and high-risk categories based on their responses. It has been shown to be a significant predictor of suicidal behavior.[24]This scale was administered to participants enrolled in an addictions study and it was completed separately from the SIBAT. Depending on the timing of the participant's entry into the study, the suicidality component of the MINI may have been completed before or after the SIBAT.

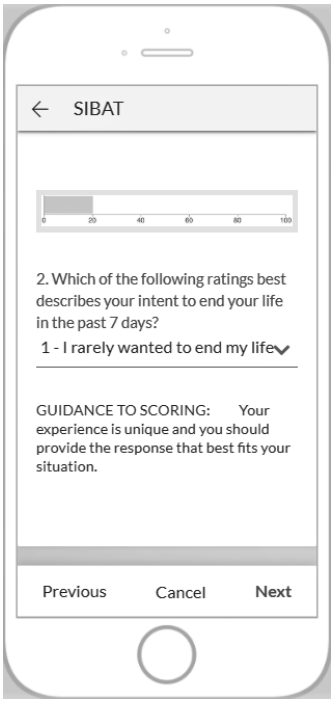

Figure 3. Screenshot of SIBAT application

4.3.2 Participant Recruitment

This study was approved by the Health Research Ethics Board at the University of Alberta. Participants were recruited from University of Alberta students and trainees participating in a study of addictions which includes a first phase of data entry via the Qualtrics platform, and a second phase, in which participants were invited to complete the Mini International Neuropsychiatric Interview (MINI) 7.0.2 through a link to the questionnaire provided online. The SIBAT was added part-way into control recruitment for the study of addictions as an optional additional measure for participants to complete. Participants were aware that the SIBAT was optional and no remuneration was offered for this.

A recruitment flowchart is shown in Figure 4. At the time of data analysis, the University of Alberta Office of the Registrar had sent 60 000 invitations to participate in the study of

addictions via email. Students who expressed interest (4214 by the time of data analysis) were asked to provide their first name or a pseudonym (to maximize anonymity and protect participant privacy) and a contact email address of their choice. A second invitation email was then sent to the email provided, specifying that, in order to participate in the study, the student must have the ability to adequately understand and respond to all questionnaires in English and have no previous or current diagnosis of addiction or other psychiatric or serious mental illnesses such as schizophrenia, dissociative identity disorder, cognitive impairment, dementia, or intellectual disability. This email also included a link to participate in the study and a form for participants to provide written informed consent to the study. Of the participants who expressed interest in the study, 2967 completed data entry of a variety of measures using the Qualtrics platform.

Invitations to complete the SIBAT were sent to 913 participants of the addictions study who completed the Qualtrics phase of the study. As this study was added-on midway through the addictions study, participants who had completed phase 1 of the addictions study and had not yet completed the MINI were invited to complete the SIBAT using the app developed on the AppSheets program. These participants were invited to complete the MINI at approximately the same time that they were invited to complete the SIBAT. Therefore, some participants completed the SIBAT prior to the MINI 7.0.2 and others completed the SIBAT after the MINI 7.0.2. Participants entering the study after addition of this add-on were invited to complete the SIBAT using the Qualtrics platform. The majority of participants who entered this add-on study completed the MINI within one or two months of the SIBAT. Approximately 30 additional participants who had already completed the MINI as part of the addictions study were invited to complete the SIBAT if they had scored moderate risk or higher on the suicidality component of the MINI.

Of those invited, 411 began the SIBAT; however, 17 of these participants did not complete each module and were excluded from the study as a result. Complete SIBAT data were collected for 394 participants, 204 of whom completed the SIBAT using a program developed on the AppSheets platform and the other 180 completed the SIBAT using the Qualtrics interface. These data were used to evaluate the internal consistency of the SIBAT. At the time of data analysis, 314 of these participants had also completed the MINI and these data were used to compare responses to the SIBAT with the MINI.

4.3.3 Statistical Analyses

Statistical analysis was performed using IBM SPSS Statistics Version 25. Data were analysed using complete case analysis as the primary method of data analysis and multiple imputation as the secondary method of data analysis. Little's Missing Completely at Random (MCAR) test was conducted in order to test whether the data were missing completely at random. Multiple imputation of responses from modules 2, 3 and 5 (73 variables total) was performed by fully conditional specification with 8 imputations. It has previously been suggested that the number of imputations should be similar to the percentage of incomplete cases.[90] As 32/394 (8.1%) cases were missing responses, we used eight imputations for this analysis.

The internal consistency for module 5 was calculated using Cronbach's alpha coefficient. The number of items in module 2 and in module 3 were 21 and 48 respectively. As scales with more than 14 items have been demonstrated to have high Cronbach's alpha irrespective of the internal consistency of the scale,[91] we calculated the internal consistency of modules 2 and 3 using item-total correlations for each item in these modules.

In order to assess the criterion validity of the SIBAT, individual items from module 5 of the SIBAT were compared to similarly worded items from the MINI. We then compared the total score of modules 2 and 3 combined to the severity rating of the suicidality component of the MINI and the total score of SIBAT module 5 (as questions in module 5 ask about suicidality directly).

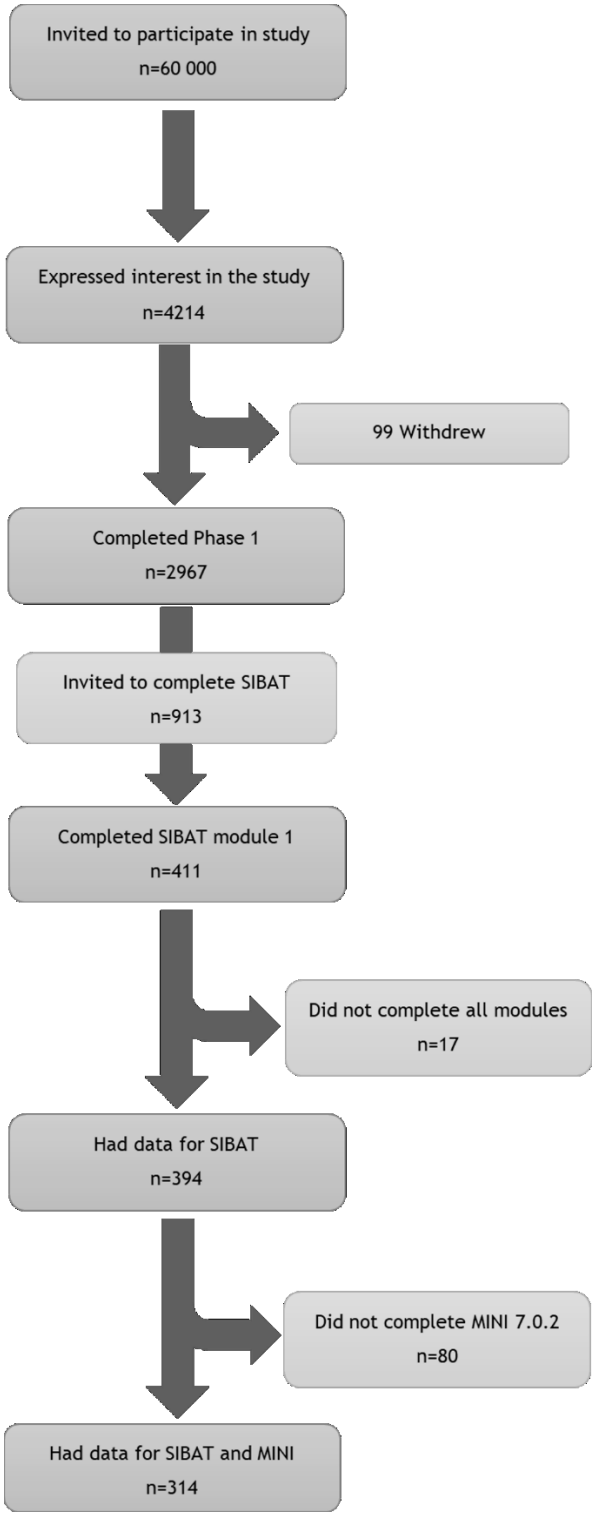

Figure 4. Recruitment flowchart of participants in study of the SIBAT rating scale

An estimate of the time required to complete modules 2, 3 and 5 was determined using the timestamps of data collected by the app developed on the AppSheets platform (n=204). The Qualtrics interface did not collect data on start or completion times of the measure and so participants who completed the SIBAT using the Qualtrics interface were not included in this analysis. The timestamp of module 1 submission was used as the starting time of module 2 and the timestamp of module 5 submission was used as the completion time (it was not possible to determine the start time of module 1 as this was not recorded). The difference between these two times was then used as an estimate of the time to complete the modules studied. As users did not necessarily need to complete the measure all in one sitting, some results greatly overestimated the duration of the measure. Given the resulting positive skew of the data, the median time for scale completion is reported.

4.4 Results

Only 4/73 items had greater than 2/394 (0.5%) missing responses. These items were module 2, item 2, “Over the past 7 days I have felt agitated” (4/394, 1.0%), module 3, item 5, “My spiritual/religious beliefs prevent me from ending my life” (6/394, 1.5%), module 3, item 7, “My concern for others prevents me from ending my life” (10/394, 2.5%), and module 3, item 8, “If I developed a life-threatening illness, I would make every effort to overcome it” (3/394, 0.8%). No case had more than 4/73 (5.5%) missing responses. As such, we report our findings using complete case analysis. Similar findings for all outcomes were obtained when using multiple imputation.

The internal consistency of module 5 of the SIBAT was high (Cronbach's $\alpha = 0.87$, $n = 4$ items). Item-total correlations for module 2 ranged from 0.22 – 0.79, with 19/21 items having item-total correlations > 0.4 (Appendix 2). The remaining two items related to aggressive impulses and command auditory hallucinations. Item-total correlations for module 3 ranged from -0.20 – 0.79, with 41/48 items having item-total correlations > 0.4 (Appendix 3). Of the remaining 7 items, two had negative item-total correlation values. One item with a negative item-total correlation value assessed concern for others and the other item assessed fear of dying. The other items with item-total correlations < 0.4 assessed the role of spiritual/religious beliefs related to thoughts of dying, the presence of severe physical pain, the desire to improve one's life, the benefit from helping others and the desire to spend time with others.

The comparison of items from module 5 with items from the MINI is shown in Table 5. Each item from module 5 of the SIBAT had a statistically significant association with its corresponding MINI item. As three items were ordinal and 25% of cells had expected counts less than 5, they were compared using Fisher's Exact Test and validity coefficients could not be generated. The fourth item of module 5 had a moderate correlation with its corresponding item from the MINI.

The sum of scores from modules 2 and 3 had a moderate correlation with the assessment of suicide risk determined by the MINI (Spearman's rho = 0.44, $P < 0.001$), which assesses suicidality in the preceding month, and a strong correlation with the total score of SIBAT module 5 (Spearman's rho = 0.62, $P < 0.001$), which assesses suicidality in the preceding week.

The median time to completion of modules 2, 3 and 5 was 14.3 minutes. The majority of participants used in this analysis (n=175/204, 85.8%) completed the measure within 30 minutes. Some users (n=7/204, 3%) completed the modules on separate days.

Table 5. Comparison of items from SIBAT module 5 to suicidality component of MINI 7.0.2

SIBAT	**MINI (In the past month did you…)**	**Significance**
Which of the following ratings best describes your desire to die in the past 7 days?	Think (even momentarily) that you would be better off dead or wish you were dead or needed to be dead?	$p < 0.001$[a]
Which of the following ratings best describes your thinking about suicide right now?	Think (even momentarily) about harming or of hurting or of injuring yourself with at least some intent or awareness that you might die as a result or think about suicide (i.e. about killing yourself)?	$p < 0.001$[a]
Which of the following ratings best describes your intent to end your life in the past 7 days?	Intend to act on thoughts of killing yourself?	$p = 0.029 – 0.032$[a]
Given your current thinking and past experience, which of the following best describes the likelihood that you attempt to end your life in the near future?	How likely are you to try to kill yourself within the next 3 months on a scale of 0-100%?	Spearman's rho = 0.442, $p < 0.001$

[a]Compared using Fisher's Exact Test

4.5 Discussion

Overall, our findings indicate that the self-report component of the SIBAT has high internal consistency overall and items from module 5 have good concurrent validity with corresponding items from the suicidality component of the MINI. The total score of modules 2 and 3 was associated with the severity of suicide risk determined by the MINI and with the score from SIBAT module 5. Modules 2, 3 and 5 had a total of 73 items (and an additional item asking how the assessment may have changed the participant's thinking about suicide); however, most participants completed the measure within 30 minutes.

Some specific items demonstrated poor item-total correlation. Some items focused on factors related to aggressive impulses, psychotic symptoms and somatic symptoms. These factors are less likely to be present in the university population compared to other populations such as psychiatric inpatients. As such, these factors may be less likely to contribute to suicide risk in the participant population studied; however, these factors may have a greater influence on suicide risk in other populations in which this scale may be used. Two items, "My spiritual/religious beliefs prevent me from ending my life" and "My concern for others prevents me from ending my life", had both the highest rates of non-response and low item-total correlations.

The sum of scores from SIBAT modules 2 and 3 had a stronger correlation with module 5 of the SIBAT than with the suicidality component of the MINI. There are a few possible reasons for this. SIBAT module 5 would have been completed around the same time as modules 2 and 3; whereas the MINI was completed at a different time, either before or after SIBAT completion. In addition, questions from the MINI asked about symptoms occurring in the preceding month, whereas SIBAT module 5 asked about symptoms occurring in the preceding week. The

difference in timing of scale completion and in the time period assessed could both contribute to the increased correlation with SIBAT module 5. In addition, SIBAT module 5 responses were scored on a 5-point Likert scale whereas the MINI had yes/no responses. This may be due to dichotomizing data resulting in lower correlations in general.[92] Furthermore, the increased granularity afforded by the greater range of possible responses may also explain the difference in correlations, as participants may have been hesitant to provide a positive response if they felt their symptoms were quite mild.

Even though modules 2, 3 and 5 combined had 74 items, the median time to completion was less than 15 minutes. This suggests that implementation of the SIBAT may be feasible in settings where users are in a situation of waiting and can turn their attention to something else, even for relatively brief periods, such as when they are awaiting assessment in emergency departments or clinic waiting rooms. The comprehensive nature of the SIBAT may bring areas of attention to the clinician's awareness, without requiring the clinician to ask extensive screening questions in a protracted assessment.

As the SIBAT has many items, analysis of such data in a large dataset would be facilitated by approaches such as artificial intelligence/machine learning. The use of machine learning has been a topic of increasing interest in medicine. It may be particularly useful in the study of psychiatric illness as use of machine learning could lead to the development of new hypotheses around the nature of the illnesses themselves and the treatment therof.[93] This approach is in keeping with the US National Institute of Mental Health Research Domain Criteria initiative, which promotes a dimensional approach to psychiatric research given the current limitations of research centered on syndrome-based clinical diagnoses.[83] Given the

complex, multifactorial and fluctuating nature of suicidality, such strategies may be helpful in improving our understanding of the way in which contributing risk factors interact.[87]

The categorization of patients at increased risk of suicide could enhance our ability to assess risk and may improve our ability to provide the optimal treatment for specific presentations. As early as 1976, cluster analysis was described as an approach to categorize patients who attempted suicide.[94] Seven groups were recognized, varying in the severity of suicidal behavior and proximity to others, as well as in long-term prognosis and engagement with follow-up. Numerous research groups have since used machine learning to classify and characterize groups of patients with increased suicide risk.[95-100] Cluster analysis in a group of Korean patients presenting after a suicide attempt extracted two groups, one with more impulsive, low lethality attempts and another with more well-planned attempts using more lethal methods.[96] Another study conducted in Toronto identified five clusters following analysis of data from a coroner's chart review of deaths ruled as suicide.[99] These five clusters had distinct characteristics, including the proportion of males:females, marital status, presence of mental illness, previous attempts and method used. Their findings suggest that individuals who die by suicide are more likely to have certain combinations of factors that predispose them to risk. While most traditional approaches to suicide risk assessment have focused on identifying the presence of risk factors universally associated with increased risk, these data suggest that the presence of specific constellations of factors may potentially increase risk in a synergistic fashion. Further study aimed at identifying the subgroups at elevated risk of suicide may improve our ability to understand and treat this patient population.

The development of algorithms to predict future suicidal behaviours is another potential application of machine learning. In a study of US Army soldiers, an actuarial suicide risk algorithm was generated using administrative data from a population of soldiers recently admitted for treatment of a psychiatric disorder.[95] The resulting algorithm incorporated sociodemographic variables, access to firearms, crime perpetration, previous suicidal behaviour, prior treatment, characteristics of hospitalization and disorders diagnosed. Of suicides occurring in the following 12 months, 52.9% occurred within the 5% of participants predicted as having the highest suicide risk. Another study applied machine learning to data from a repository of electronic health records to develop a machine learning algorithm that predicted future suicide attempts (AUC = 0.84, precision = 0.79, recall = 0.95, Brier score = 0.14).[100] This study included a number of predictors including demographic data, diagnoses, previous health care utilization, previous suicide attempts, socioeconomic status, and medication data in model development. Of note, both of these studies used data from large data repositories, and so factors related to a patient's current state such as the presence of hopelessness, affective symptoms, psychotic symptoms, sleep disturbance, recent stressful events or social isolation were not included.

One important factor in the performance of machine learning is the quality of data obtained. Improving the quality of data collection has previously been noted to be critical in realizing the full potential of machine learning methods in suicide research.[101] The studies identified above used varying methods to collect data for analysis, with some studies using multiple rating scales.[96-98] The variability in these approaches to data collection makes comparing results between groups more challenging. The SIBAT offers a consistent and comprehensive approach to data collection. It includes a wide range of potential contributors to suicide risk including positive and negative affect, social isolation, interpersonal conflict,

anxiety, substance use, history of abuse, disturbance of sleep, psychotic symptoms, the presence of a plan and suicidal intent. As such, it may be useful as a tool to study factors associated with suicide risk. The use of a consistent measure across research groups may improve our ability to compare models developed in different settings and to assess the performance of models in different populations.

4.6 Limitations

The population examined in this study is not a group traditionally associated with being at elevated risk of suicide. As the most likely implementation of this scale would be in populations at higher risk, such as psychiatric inpatients or patients seen for suicidality in the emergency department, it is unclear whether these findings could be generalized to the populations in which the SIBAT would most likely be used. In addition, less than half of participants invited to the study had completed the SIBAT. It is possible that this relates to the fact that completion of the SIBAT was the only component of the AddGenes study for which no reimbursement was provided, which may have introduced selection bias. As such, we note that, while the SIBAT has demonstrated high internal consistency and concurrent validity in the population studied, further investigation in other settings may be necessary to establish this tool as a valid measure.

Furthermore, as the SIBAT was only administered once in this study and no follow-up data were collected, measures such as test-retest reliability, scale adherence and most notably predictive validity could not be assessed. While our initial data suggest that the SIBAT may evaluate state-related components of suicidality, repeated, longitudinal administration of the SIBAT would be necessary to establish the sensitivity of the SIBAT to changes over time.

As noted in the introduction, no current tool has been shown to perform adequately in the prediction of suicide as an event. As this study compared data obtained using the SIBAT to the suicidality component of the MINI, it is important to note that the MINI itself has limitations in its use. As noted above, most items from the suicidality component of the MINI use dichotomous responses, which may lead to lower correlations with outcomes of interest. In addition, a previous study found that the MINI has a low PPV in the prediction of future threats and acts of suicidal behavior or nonsuicidal self-injury[24], and was discussed in greater detail in Chapter 2. Given these limitations, comparison of data collected using the SIBAT to clinically relevant outcomes, such as future suicide attempts and death by suicide, may allow for greater insight into the predictive validity of this new instrument.

4.7 Future Directions

Further studies evaluating the validity of the SIBAT, especially in populations associated with higher risk of suicide, could allow us to better understand the strengths and weaknesses of this measure. Specifically, studies in which the SIBAT is administered over time and studies examining long-term clinically relevant outcomes, such as future suicide attempts or death by suicide, would help identify potential uses for the SIBAT in research and in practice. There is also room for further improvements to the SIBAT and further research involving factor analysis and item response theory may identify areas of redundancy.

As indicated in the discussion, future administration of the SIBAT could include data collection for use in machine learning. As the SIBAT covers a large number of areas associated

with suicide risk, data collected may improve our understanding of suicidality and assist us in developing treatments for different presentations.

4.8 Conclusions

The findings of this study suggest that the self-report component of the SIBAT has good internal consistency overall and items in module 5 have good concurrent validity with the suicidality component of the MINI. The total score of modules 2 and 3 combined had a moderate association with the suicidality component of the MINI and a stronger association with SIBAT module 5. The median time to complete modules 2, 3 and 5 was 14.3 minutes, with 85.8% of participants completing these components in less than 30 minutes. While these data are preliminary, they support further assessment of the validity of the SIBAT in populations at higher risk of suicide. Assessment of the SIBAT involving repeat assessment, association with long-term outcomes and refinements of the measure could provide us with insight into its potential use in research and clinical settings.

4.9 Acknowledgments

KJA was awarded a fellowship grant from Janssen Inc., Canada to support activities of a fellow (trainee). ECC was the named fellow on a grant from Janssen Inc., Canada, with the funds being used to contribute to support for research assistants involved in participant recruitment and database maintenance. ECC also received two Rudelle Hall Graduate Scholarships (2018 and 2019/20) and awards from the Department of Psychiatry Residency Project Fund at the University of Alberta. A grant from the American Foundation for Addiction Research to KJA and the Alberta Centennial Addiction and Mental Health Research Chair and transitional funding

to KJA provided support for participant recruitment, data collection, and database maintenance. PJC received a Fulbright Canada – Palix Foundation Distinguished Visiting Research Chair in Brain Science, and Child and Family Health and Wellness (host institution: the University of Alberta). The funding sources had no involvement in any aspect of the study design, data collection, data analysis, or interpretation.

CHAPTER 5: THE FEASIBILITY AND ACCEPTABILITY OF MOBILE APPLICATION-BASED ASSESSMENT OF SUICIDALITY USING A NOVEL TOOL, THE SUICIDE IDEATION AND BEHAVIOR ASSESSMENT TOOL (SIBAT)

Authors

Eric C Chan[1], Keanna Wallace[2], Esther Yang[2], Leslie Roper[2], Garima Aryal[2], Andrius Baskys[1,3,4], Richard Isenberg[5], Patrick Carnes[1], Bradley Green[6], Rohit J Lodhi[7], Katherine J Aitchison[1,2,8]

Institutions

[1]Department of Psychiatry, University of Alberta, Canada; [2]Department of Medical Genetics, University of Alberta, Canada; [3]Graduate College of Biomedical Sciences and University Medical Center, Western University of Health Sciences, USA; [4]Memory Disorders and Genomic Medicine Clinic, Riverside, USA; [5]Psychological Counseling Services, Ltd., Scottsdale, USA, [6]Department of Psychology, University of Texas at Tyler, USA; [7]Department of Psychiatry, University of Saskatchewan, Canada; [8]Neuroscience and Mental Health Institute, University of Alberta, Canada.

5.1 Abstract

Objectives: The aim of this study was to assess the validity of a mobile application-based self-report questionnaire in the assessment of suicidality.

Methods: We developed a program for the administration of self-report components of the Suicide Ideation and Behavior Assessment Tool (SIBAT). We invited university students and

trainees enrolled in a study of addictions to complete this component of the SIBAT using the program on their mobile devices or personal computer.

Results: 196 participants completed all modules of the SIBAT, with 97 using their mobile device and 99 using their personal computer. Rates of completed questionnaires between the two groups were compared, as were the responses to the items and the total scores. There was a significant difference between proportions of scale completion in both groups, with 4/97 (4.1%) participants who used a mobile device to complete the scale and 12/99 (12.1%) participants who used a personal computer to complete the scale not responding to all questions (P= 0.04). A trend toward greater disclosure of suicidality was observed in the mobile device group (P = 0.12 – 0.43).

Conclusions: Data collected via mobile device showed good concurrent validity with data collected via personal computer. Participants completing a questionnaire using a mobile device had a higher proportion of complete responses than those using their personal computer. Furthermore, a trend was observed suggesting participants may be more willing to disclose suicidal ideation using a mobile device however, replication of these findings using larger sample sizes is needed.

Keywords: Smartphone, Mobile applications, Suicide, Psychiatric status rating scales, Medical informatics applications

5.2 Introduction

Recent studies examining the predictive validity of various instruments showed that none had sufficient accuracy for routine clinical use, in part due to the rarity of suicide as an event.[1-4] Consequently, the development of validated instruments for the screening and assessment of suicide risk has been identified as a priority by the National Action Alliance for Suicide Prevention, with an emphasis on the integration of technology.[6]

The assessment of suicide risk is also complicated by the considerable variability in suicidal ideation over potentially short periods of time. Suicidality is influenced by a variety of biological, psychological, and social factors, many of which can vary over time.[49] In addition, a previous study has suggested that suicidal thoughts could be classified into different phenotypic profiles based on changes in their intensity and variability over time.[102] These findings suggest that discrete, repeated assessments of suicidality may improve our understanding of suicidal behaviour and identify those at increased risk. Historically, administration of assessments of suicidality would be complicated by the need for users to present to a health care setting; advances in smartphone technology however, may provide a way of overcoming this limitation.

The smartphone has been posited as a potentially useful tool in the assessment of suicide risk and in the development of dynamic models of suicidality.[103] Smartphones offer a unique opportunity in the field of psychiatry, making it is now possible for users to report their symptoms from anywhere, without being required to attend a clinic or to remain at home to access their computer. As a result, assessments can be conducted more frequently and in more naturalistic settings while also decreasing disruption of the user's daily routine. A recent study administered a brief three-item questionnaire on suicidal ideation administered via smartphones

four times per day, with an average of 2.20 – 2.28 responses per participant per day.[102] The authors used the data to identify five profiles of suicidal thinking based on the mean levels of suicidal ideation and variability observed, conclusions which were made possible in part by the increased temporal granularity afforded by the mode of data collection.

Many smartphone-based self-report tools are administered using mobile applications (apps). While many apps have been developed for suicide prevention, they are generally focused on education, safety planning, connection to community supports, and the development of improved coping abilities, with few apps being focused on suicide risk assessment.[17] The validity of mobile apps monitoring depression[16,104] and bipolar disorder[77,78] has been investigated in previous studies; however, to our knowledge, no study has focused on validating data collected for suicide risk assessment using a mobile app. Given the sensitive nature of questions related to suicide, there may be differences in users' willingness to disclose suicidal ideation using a mobile app when compared to other methods of data collection. As such, we developed a program for self-report components of a new measure that had been designed for serial assessment of suicide risk and compared data collected via mobile device with that collected via personal computer in order to examine the concurrent validity of suicide risk assessment using this mobile app.

5.3 Methods**Participant Recruitment**

This study was approved by the Health Research Ethics Board at the University of Alberta.

Participant recruitment was conducted through the University of Alberta Office of the Registrar using the pool of participants enrolled as controls in an addictions study as described in Chapter 4.

Controls in the aforementioned addictions study were recruited through the University of Alberta Office of the Registrar. The Office of the Registrar sent invitation emails to samples of undergraduate students, graduate students, postdoctoral students, post-graduate students, open studies students and recently convocated students at the University of Alberta who were at least 18 years of age or older, had been registered in at least one course at the University of Alberta since Fall 2016 (except those who had completely withdrawn from registration) and did not have any reasons to not return in the next academic term or not be on campus (such as suspension).

A total of 56 000 students were sent invitation emails and interested students were asked to contact the study team by email.

5.3.2 Data collection

A recruitment flowchart depicting the number of participants at each stage of the study is shown in Figure 5. In the first phase of the study of addictions, participants were provided with a link to a package of online questionnaires including the Internet Addiction Test (IAT) and the Adult ADHD (Attention Deficit Hyperactivity Disorder) Self-Report Scale (ASRS). This phase was completed by 2261 participants.

Of the participants who completed the first phase of the study of addictions, we selected 479 for inclusion in this study. Participants were selected based on the order in which they signed up for the study and completed the first phase of the study. Some additional participants were invited to complete the SIBAT if they had scored moderate risk or higher on another measure of suicidality. These participants were invited to complete the SIBAT using their choice of personal computer or mobile device, with 219 participants completing the first module of the SIBAT. Of the 219 participants, 15 did not complete all modules, including the question asking

which device they used to complete the SIBAT at the end of the measure. Consequently, their responses were not included in data analysis. In addition, eight participants completed all modules but did not have responses recorded to the question at the end of the measure asking them which device they used to complete the SIBAT, so these participants were excluded from the analysis comparing the two modes of data collection.

As participants completed the SIBAT using the device of their choice, we analysed responses pertaining to suicidality, mood, anxiety and psychotic disorders obtained via the MINI as well as data on ADHD and internet addiction (obtained via the ASRS and IAT respectively) in order to control for potential confounding, as participants with certain diagnoses could be more likely to respond using one method or the other. Furthermore, internet addiction or impairments in attention could affect the response rate.

Another 307 participants in the study of addictions who completed the first phase were invited to complete the SIBAT using the Qualtrics interface instead of the AppSheets program, with 190 complete responses at the time of data analysis. Data collected using the Qualtrics was not included in this study as participants could complete the Qualtrics using various devices and data on which device was used was not collected for all respondents. Furthermore, the differences in interfaces may have a confounding effect on our results.

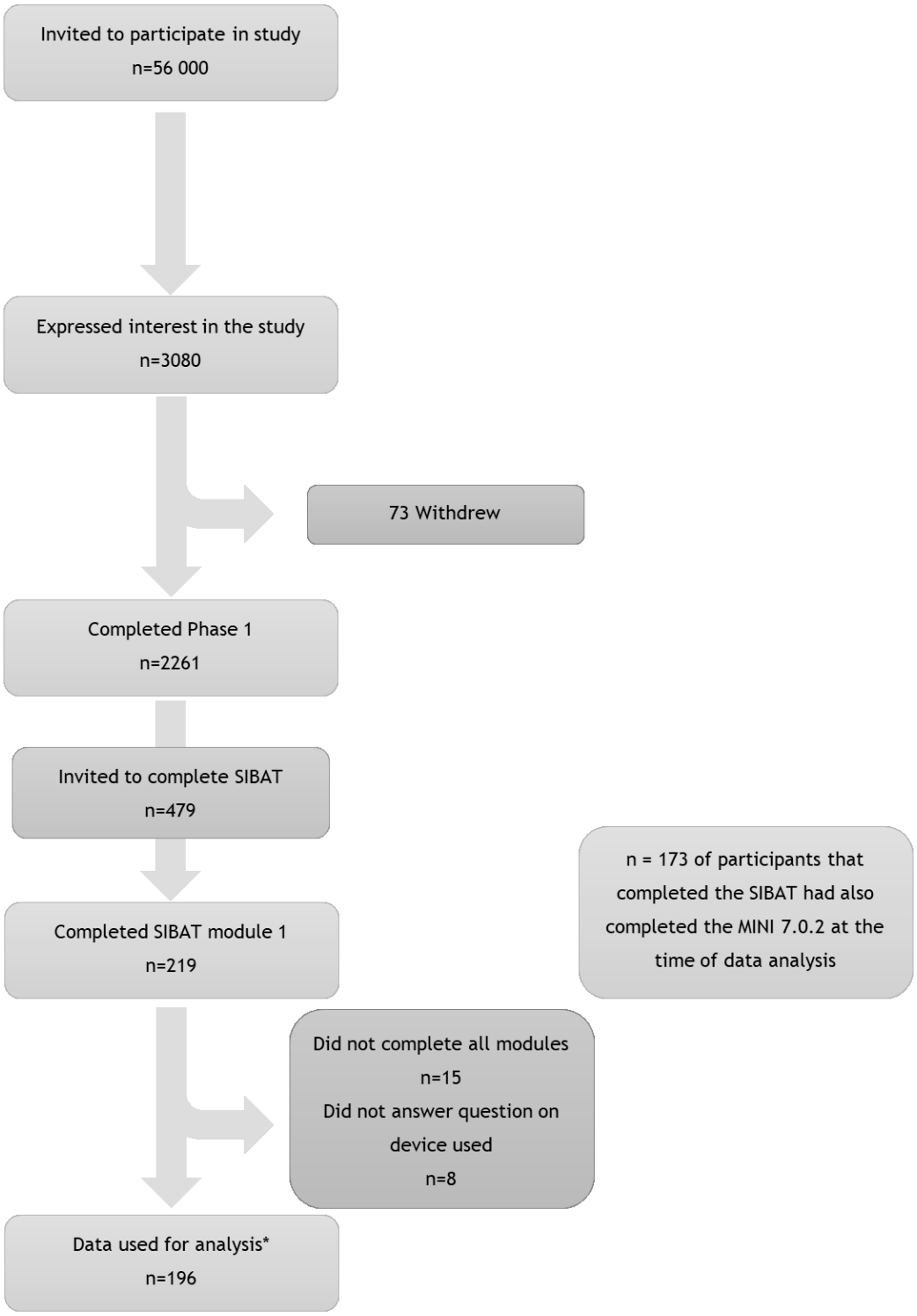

Figure 5. Recruitment flowchart of participants in study of the SIBAT mobile application

5.3.3 Assessment Tools

5.3.3.1 Suicide Ideation and Behavior Assessment Tool

The Suicide Ideation and Behavior Assessment Tool (SIBAT) is a new, comprehensive rating scale developed for the assessment and monitoring of suicidality. It was created by a group of clinical trial and academic experts in scale development. It includes both clinician-and self-report components and includes items assessing a range of factors including suicidal ideation, suicidal intent, feelings of hopelessness, feelings of worthlessness, social isolation and substance use. The content of the SIBAT is described in greater detail in Chapter 4.

A program for the self-report modules of the SIBAT was developed using the AppSheets platform (Figure 6). This program allowed users to enter data either as an app on a mobile device (phone or tablet) or via a browser on a personal computer. It could be completed from any location in which the participant had access to their device and an internet connection. Each module could be completed separately at the user's convenience.

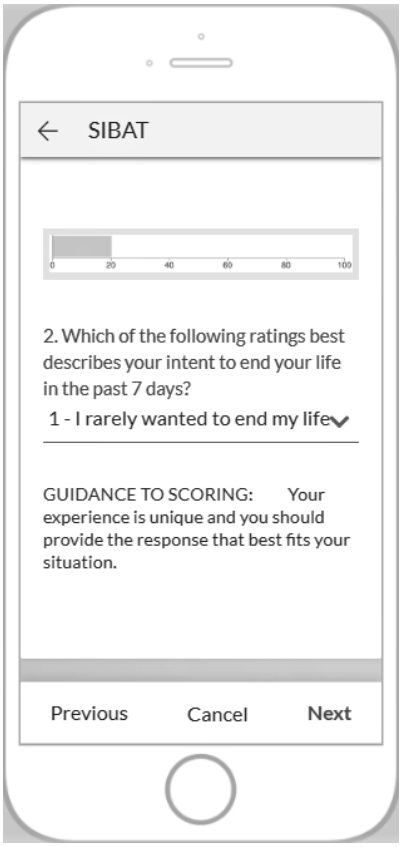

Figure 6. Screenshot of SIBAT Application

5.3.3.2 Mini International Neuropsychiatric Interview

The Mini International Neuropsychiatric Interview (MINI) is a short, structured, diagnostic interview that has been validated and used for assessment in psychiatric research settings.[105] The MINI version 7.0.2 was administered using a web-based, self-report version supplied initially by Medical Outcomes and latterly by NView. The MINI modules employed in this study assess for some of the most common psychiatric disorders in the DSM-5 including major depressive disorder, suicidality, bipolar disorder, obsessive-compulsive disorder, posttraumatic stress disorder, alcohol use disorder, substance use disorder, psychotic disorders, anorexia nervosa, bulimia nervosa, binge eating disorder, generalized anxiety disorder and antisocial personality disorder.

5.3.3.3 Adult ADHD Self-Report Scale v1.1 (ASRS V1.1)

The Adult ADHD Self-Report Scale v1.1 is a tool developed by a group of psychiatrists and researchers working in conjunction with the World Health Organization for the screening of ADHD. It consists of 18 self-report items rated on a 5-point scale that assess for symptoms of inattention and hyperactivity-impulsivity. The first 6 items (Part A) are now used as a screening measure to identify individuals with symptoms highly suggestive of ADHD (sensitivity 68.7%, specificity 99.5% for four or more marks in the gray boxes in these items).[106] The remaining 12 questions comprise Part B of the scale, which may assist a clinician's assessment regarding frequency of symptomatology.

5.3.3.4 Internet Addiction Test (IAT)

The Internet Addiction Test (IAT) is an instrument that measures the presence and severity of compulsive internet use among adults and adolescents. It consists of 20 self-report items rated on a 5-point Likert scale and was developed by expanding on a core set of items derived from DSM-IV criteria for pathological gambling.[107] It examines the effect of internet use on an individual's daily routine, social life, productivity, sleeping patterns, and feelings. The psychometric properties of the IAT have been examined across multiple studies and it has been shown to have good internal consistency and concurrent validity.[107-110] Cronbach's alpha for the IAT has been found to be 0.92 in college age students.[108-109]

5.3.4 Statistical Analyses

Statistical analysis was performed using IBM SPSS Statistics Version 25. Data were analysed using complete case analysis as the aim of the study was to evaluate the validity of data collected using mobile devices. The percentage of cases without responses while relatively low at 8.2%, was greater than 5%.

The Mann-Whitney U test was conducted to compare the responses to each item in modules 2, 3, and 5 as well as the sum of scores from each of modules 2 or 3, and the sum of module 2 and 3 scores combined. Nonparametric tests were selected as scores from these modules were not normally distributed.

The frequencies of mood, anxiety and psychotic disorders in each group were determined using participant responses to the MINI 7.0.2. Rates of ADHD were determined using the ASRS, with individuals scoring in the grey area for 4 of the first 6 items screening positive for ADHD. Participants screened positive for a moderate level of internet addiction if they scored 50 or above on the IAT.

The frequencies of bipolar I disorder, suicidality, generalized anxiety disorder and ADHD were compared between groups using chi-squared analysis. All expected statistical cell frequencies were greater than five for these disorders. The frequencies of major depressive episode, social anxiety disorder and internet addiction were compared using Fisher's exact test as the expected counts of participants screening positive for each diagnosis were less than 5 for at least one group. No participants in either group met criteria for a history of psychotic disorder. Fisher's exact test was performed to compare distributions of responses to items in module 5 as all items had two or more cells with expected count less than 5. A *P* value of less than 0.05 without adjustment for multiple testing was interpreted as significant.

5.4 Results

A total of 196 participants completed the SIBAT rating scale, with 97 completing the scale using their mobile device and 99 using their personal computer. There was a significant difference between proportions not responding to all administered SIBAT questions: 4/97 (4.1%) for participants completing using a mobile device and 12/99 (12.1%) for participants using a personal computer ($p = 0.04$). There was no significant difference however, between mode of entry on item level analysis (Appendix 4). Likewise, median total scores for module 2 and 3 combined did not differ significantly (mobile = 67, personal computer = 64, $p = 0.85$).

Responses to questions from module 5 are shown in Figure 7 below. Higher rates of positive responses (defined as scores of 2 or greater) to all items were observed in the mobile device group, though these differences did not reach statistical significance ($p = 0.12 – 0.43$).

At the time of data analysis, 83 of the 97 (86%) participants who completed the scale using a mobile device and 90 of the 99 (91%) participants who completed the scale using a personal computer had completed the MINI. The frequencies of observed suicidality, mood, anxiety and psychotic disorders by MINI criteria are shown in Table 6. The frequencies of these disorders did not differ significantly between mode of data entry groups.

Table 6. Frequencies of disorders determined by the MINI 7.0.2

	Smartphone or other mobile device	**Personal Computer**	**Significance (*p*)**
Diagnoses determined by MINI 7.0.2			
Current Major Depressive Episode	5/83 (6.0%)	3/90 (3.3%)	0.48[a]
Suicidality in Past Month	23/83 (27.7%)	23/90 (25.6%)	0.75[b]
History of Bipolar I Disorder	5/83 (6.0%)	11/90 (12.2%)	0.16[b]
Social Anxiety Disorder	4/83 (4.8%)	6/90 (6.7%)	0.75[a]
Any Psychotic Disorder	0	0	-
Generalized Anxiety Disorder	6/83 (7.2%)	10/90 (11.1%)	0.38[b]
Diagnoses determined using ASRS or IAT			
ADHD	19/97 (19.6%)	20/99 (20.2%)	0.91[b]
Internet addiction	5/97 (5.2%)	3/99 (3.0%)	0.50[a]

[a]Compared using Fisher's Exact Test

[b]Compared using Chi Squared Analysis

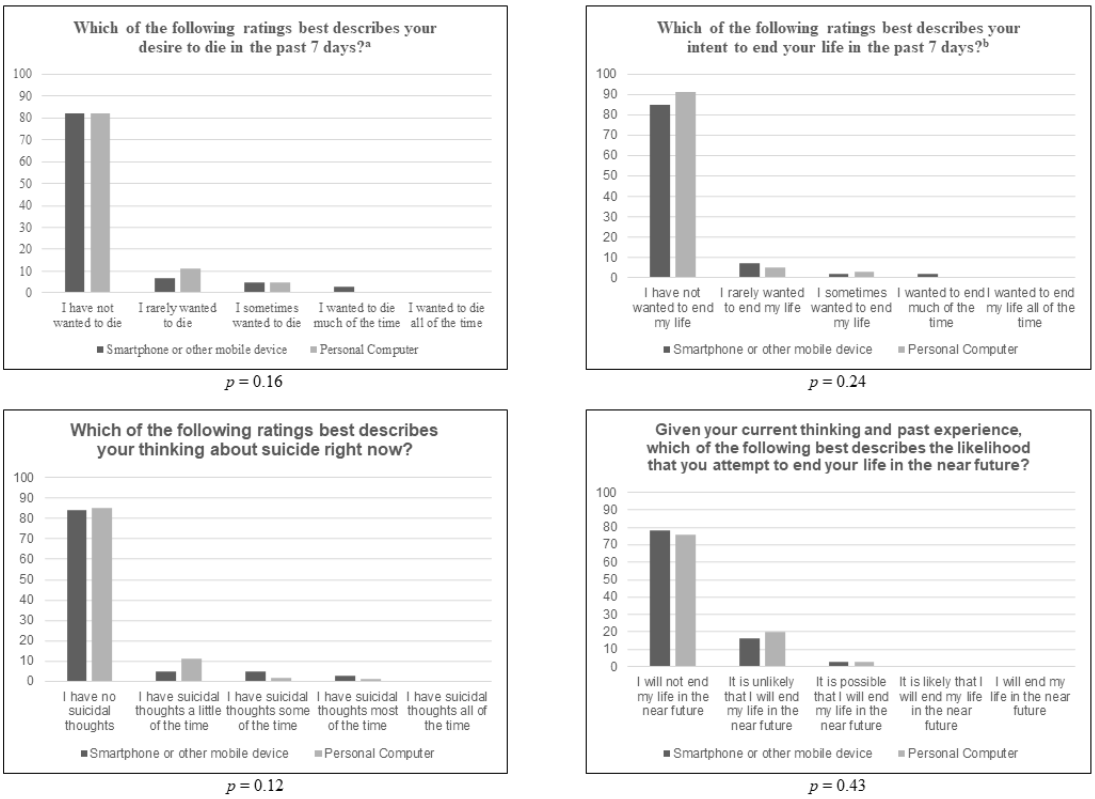

Figure 7. Distribution of responses to module 5 items

5.5 Discussion

Technological advances have the potential to substantially improve our assessment and understanding of mental illness, but validation of these tools is required. Our results indicate that users completing the SIBAT using a mobile device had a significantly higher proportion of complete responses compared to users completing the SIBAT using a personal computer. Data collected was similar between methods of data collection and a trend was observed suggesting participants may be more willing to disclose suicidality when completing the measure using a mobile device, however this did not reach statistical significance.

In terms of difference between proportions of complete data by mode of data entry group, there are various possible explanations. Certain items may have been skipped intentionally due to the sensitive nature of the questions being asked, with users entering data using a mobile device being more willing to respond due to the increased privacy afforded by the smaller screen, though this is felt to be less likely as no item had more than two missing responses. Other programs and alerts on the personal computer may have distracted the user while completing the rating scale, leading to inadvertently skipped items. It is also possible that users completing the SIBAT via mobile device may have done so while completing other tasks, such as eating or waiting for public transit, and this may have paradoxically improved their ability to focus on the measure by for example, ensuring adequate stimulation.

The trend in the mobile device group for higher scores on questions from module 5 is also notable, especially since levels of "suicidality in the past month" as determined by the MINI were similar between the two groups. There was, however, no significant difference between the MINI suicidality scores by data entry mode group. Similar findings have been previously reported in a study examining the validity of a mobile app for the PHQ-9.[16] In this study, 9 of the 13 subjects reported suicidal ideation on a PHQ-9 app while none of these subjects reported suicidality using the paper-based PHQ-9. The authors suggested that electronic monitoring may provide a more accurate method to record and understand suicidal ideation. It is possible that participants are more willing to disclose suicidality on their mobile device than they are willing to disclose using their personal computer. This may be due to the increased privacy afforded to users due to the smaller screen of mobile devices or the fact that mobile devices tend not to be shared with others. Further investigation with larger study populations could be beneficial in determining whether this trend continues.

Current predictive instruments for suicidal behaviour have been shown to have poor positive predictive value, in part due to the low prevalence of suicidal behaviour in the general population.[1] Machine learning and artificial intelligence have the potential to improve our ability to predict suicide but require high quality data on risk factors for optimal performance.[101] Mobile app-based assessments may improve our ability to identify risk factors and patterns of symptoms that have previously not been the foci of clinical attention.[103] In addition, the integration of mobile apps with passive data collection, such as GPS location tracking, physiological monitoring, usage patterns, and facial recognition software, could contribute substantially to developing improved models of suicide prediction.

5.6 Limitations

It was not possible to compare dropout rates between participants using mobile devices and those using their personal computer as determination of the device used occurred at the end of the assessment and consequently, we were unable to determine which device was used by participants who did not complete all modules.

The study population consisted primarily of university students, which limits our ability to generalize these findings to other populations in which assessment and monitoring of suicidality would typically be performed. In clinical populations, differences in level of cognitive functioning, such as in those with severe and persistent mental illness including schizophrenia and severe depression, could impact user ability to complete the measure. Physical illnesses including chronic pain and neurological disorders may interfere with user ability to interface with the app due to limitations in mobility. Many life-threatening illnesses such as cancer and

end-stage renal disease may lead to decreased attention which could also limit user ability to complete the measure.

In addition, differences in social circumstances may impact response rates compared to the study population. Certain demographics may have less familiarity with mobile technology, making scale completion more challenging and others may have limited access to a mobile device, such as psychiatric inpatients. Some populations, such as those who are incarcerated or those lacking the financial means to afford a smartphone, may not have access to a suitable mobile device at all.

The participants in this study were limited to those willing to participate in the component assessing suicidality. As only around half of the participants invited to this study had completed the measure, lack of reimbursement may have introduced selection bias or there may have been exclusion of those with suicidality who do not wish to record their symptoms.

The SIBAT self-report modules were only administered once via the app in this study. As such, data regarding the validity of the SIBAT for repeated assessment, which is one of its potential major benefits, remain to be published. Given the single administration of one measure of defined length, no conclusions can be drawn from this study about the optimal length or frequency of assessments.

5.7 Future Directions

Further research with larger populations investigating the rates of disclosed suicidal ideation in populations completing the SIBAT using the app and through more conventional means could clarify whether the trend observed in this study is due to increased willingness to report

symptoms or to other factors. Studies involving populations at greater risk of suicide, such as those with a history of serious suicide attempt or severe mental illness could also inform the rates of disclosed suicidal ideation using different methods of data collection. In addition, studies assessing adherence rates of participants completing measures at different frequencies could inform us on the optimal frequency of assessments, improving our ability to integrate these measures into research and clinical settings.

5.8 Conclusions

Mobile apps offer a novel method in suicide risk assessment. Data collected using an app for the self-report component of the SIBAT demonstrated concurrent validity with the suicidality component of the MINI. A significant difference in proportions of completed responses was observed between the two modes of data entry and a trend was observed in which participants completing the measure using a mobile device appeared more willing to disclose suicidality. Due to the small number of participants, further investigation with larger sample sizes is needed to see if these results are replicated. In addition, further research into the optimal frequency of assessments, as well as investigation into which populations are suitable for the use of mobile apps, would help optimize the use of mobile app-based questionnaires in suicide research.

5.9 Acknowledgments

KJA was awarded a fellowship grant from Janssen Inc., Canada to support activities of a fellow (trainee). ECC was the named fellow on a grant from Janssen Inc., Canada, with the funds being used to contribute to support for research assistants involved in participant recruitment and database maintenance. ECC also received two Rudelle Hall Graduate Scholarships (2018 and 2019/20) and awards from the Department of Psychiatry Residency Project Fund at the

University of Alberta. A grant from the American Foundation for Addiction Research to KJA and the Alberta Centennial Addiction and Mental Health Research Chair and transitional funding to KJA provided support for participant recruitment, data collection, and database maintenance. PJC received a Fulbright Canada – Palix Foundation Distinguished Visiting Research Chair in Brain Science, and Child and Family Health and Wellness (host institution: the University of Alberta). The funding sources had no involvement in any aspect of the study design, data collection, data analysis, or interpretation.

CHAPTER 6: CONCLUSIONS AND DISCUSSION

6.1 Overview of findings and their relevance

In chapter 2, I conducted a review of the literature to identify studies examining the predictive validity of rating scales for suicide assessment. Sixteen records were identified for inclusion. These studies evaluated a wide range of assessment tools; however, no instrument predicted suicide or suicidal behaviour with sufficient accuracy for clinical use. The PPV of all instruments studied was low, with none reporting PPV > 41%. The numerous challenges that exist in the development of instruments for suicide risk assessment were discussed. I proceeded to describe how new technologies such as repeat measurements, physiologic monitoring, and machine learning may assist us in overcoming these challenges.

In chapter 3, another reviewer (SS) and I conducted a systematic review of the literature to determine the feasibility and validity of mobile app-based self-report questionnaires. We searched digital databases for studies assessing the validity of mobile app-based self-report questionnaires and identified six records for inclusion, in which two groups were each represented twice. Only two studies were identified that compared mobile app-based self-report questionnaires to standardized assessment tools (clinician-administered rating scales in both cases). These data suggested that self-report questionnaires administered via a mobile device may detect depressive symptoms in patients with bipolar disorder, but, interestingly, the correlation between self-report and clinician ratings for symptoms of mania or hypomania was more variable. One study suggested app-based assessment tools may differentiate patients with bipolar disorder from healthy controls. Several limitations in this review were noted: first and foremost, the low number of studies limits the ability to draw conclusions. Furthermore, as

patients needed to be euthymic at the time of recruitment in the studies, it is unclear whether these findings can be extended to the detection of symptoms associated with an acute mood episode.

In chapter 4, data collected by the SIBAT application using a mobile device was compared to data collected by the SIBAT application using a personal computer. In total, data from 196 participants was analysed, 97 of whom completed the scale using a mobile device and 99 using a personal computer. A significant difference was observed in proportions of scale completion between the two groups: 4.1% of users using a mobile device did not answer all items versus 12.1% of users using a personal computer ($p = 0.04$). Median total scores for modules 2 and 3 combined did not differ significantly between groups (mobile = 67, personal computer = 64, $p = 0.85$). A trend toward higher rates of positive responses to module 5 items was observed in the mobile device group, though these differences did not reach statistical significance ($p = 0.12 – 0.43$). The median completion time of modules 2, 3 and 5 was 14.3 minutes.

In chapter 5, data collected using the SIBAT administered either via the application or the Qualtrics interface were analysed and compared to data collected using the MINI in order to evaluate the internal consistency and criterion validity of the SIBAT rating scale. SIBAT data from 394 participants were analyzed, 314 of which had also completed the MINI. Module 5 of the SIBAT demonstrated high internal consistency (Cronbach's alpha = 0.87, n = 4) and items from module 5 had a statistically significant association with corresponding items from the MINI. A moderate correlation was observed between the sum of module 2 and 3 scores and the

assessment of suicide risk by the MINI. A strong correlation was observed between the sum of module 2 and 3 scores and the total score of SIBAT module 5.

6.2 Discussion

In summary, current data on instruments for suicide risk assessment indicate that none has sufficient predictive validity for clinical use. While preliminary data suggest app-based assessment tools may accurately monitor symptoms of depression in patients with bipolar disorder, these findings are limited by the low number of studies identified. These findings suggest that there is a need for further development of instruments for suicide risk assessment and there is a need for further evaluation of the ability of app-based assessment tools to monitor symptoms of bipolar disorder.

When we evaluated the validity of the SIBAT rating scale and the validity of the SIBAT when administered as a mobile app, both the rating scale itself and administration of the scale as an app demonstrated good concurrent validity with comparison measures. Furthermore, each module of the SIBAT demonstrated high internal consistency. These findings support the use of the SIBAT rating scale and the administration of the SIBAT mobile application as potential tools for the assessment of suicidality. Furthermore, higher rates of scale completion were observed in the mobile device group when compared to the personal computer group and a trend toward increased disclosure of suicidality was observed in the mobile device group when compared to the personal computer group. These findings are encouraging and support further study of the validity of these tools in the evaluation of suicidality.

As noted in the introduction, both the SIBAT rating scale and administration of the SIBAT as a mobile application have some potential advantages over conventional rating scales assessing suicidality. The SIBAT rating scale has been developed for repeated administration and, as it distinguishes between static and dynamic risk factors, it could improve our ability to monitor which factors are likely to change over time. In turn, this may allow us to better monitor the trajectory of suicidality over time and observe which risk factors, or combination of factors, may contribute to an elevated risk or affect response to various treatments. As mobile application-based assessments can be administered without requiring the patient come to a clinician's office, this may further improve our ability to monitor symptoms and risk factors more frequently. Overall, both the SIBAT and mobile application-based assessments including the SIBAT and other mood assessment tools may allow us to collect a larger volume of data that may inform our understanding of suicidality.

As noted in chapter 4, the collection of large amounts of data may be especially useful when combined with new methods of data analysis, such as machine learning and other artificial intelligence approaches. Potential methodologies that could be taken include cluster analysis in order to identify subgroups of individuals at elevated risk of suicide, and factor analysis in order to identify overarching constructs affecting risk. Subgroups could then be studied separately in future studies to improve our understanding of the trajectory of their symptoms as well as their response to different treatments. This in turn may inform which treatment modalities may be most beneficial for different presentations.

6.3 Limitations

The relatively small sample sizes and low baseline risk of this population are important limitations to consider in the data-based studies presented in chapters 4 and 5. Given the rarity of suicide as an event, very large sample sizes are necessary in order to reach clear conclusions on an instrument`s ability to accurately assess suicidality. This is compounded by the population studied (university students), which is at lower risk compared to populations in which these tools may typically be administered (such as psychiatric inpatients). As such, firm conclusions on the validity of these instruments in higher risk populations cannot be made at this time.

As noted above, one of the potential advantages of both the SIBAT rating scale and application-based assessment tools in general is the ability to administer repeat assessments over time. As the SIBAT and the app version thereof were only administered once in the studies described in this thesis, we were unable to assess the potential feasibility and utility of these to detect changes over time.

Other major limitations of these studies are related to the nature of the tools examined, as well as the standards to which they are compared. In both studies reported in chapters 4 and 5, we compared two self-report measures (the self-report version of the MINI and the self-report modules of the SIBAT). As such, we do not know whether individual ability or willingness to disclose might influence willingness to disclose symptoms or risk factors using these tools. In addition, self-report scales assess the patient's subjective experience of symptoms and are unable to capture data obtained through other means, such as observation by the clinician or report from collateral sources. Consequently, some factors, such as the potential lethality of a recent suicide attempt or the presence of neurovegetative features may not be captured by this method of

assessment. As the SIBAT also consists of a clinician-rated component, further studies incorporating this component will likely be necessary in order to improve the clinical performance of this tool.

The standards against which these tools were examined also present a limitation for these studies. Research assessing the validity of any new instrument must consider the validity of the measure against which the instrument is being assessed. Optimally, new assessment tools would be compared to a gold-standard, which is felt to have the highest degree of accuracy. For many reasons, this is often not possible. As noted in the introduction, no current instrument has been shown to perform adequately for routine clinical use. As a result, no instrument currently exists as a gold-standard for the assessment of suicidality. Even prospective collection of data on future suicide attempts or death by suicide has limitations. As noted in the introduction, suicidality is a multifactorial construct that is variable over time. Treatment interventions, including the initiation of psychotropic medication or psychotherapy, admission to hospital, and safety planning, would likely affect suicide risk. A naturalistic study in which no intervention is administered would be unethical and as such our ability to demonstrate a tool's ability to assess suicidality is limited by context. Given this consideration, I suggest that the assessment tools described in this thesis may best be used to inform our understanding of suicidality and identify which treatments are most beneficial in given populations, rather than being used as predictive instruments.

6.4 Future Directions

With regard to the validity of app-based self-report assessment tools for the monitoring of symptoms of bipolar disorder, further studies evaluating these instruments is necessary in order

to inform their validity in clinical settings. Our findings in chapter 2 indicate that the development of improved instruments for the assessment of suicide risk is needed. Our research evaluating the SIBAT aims to contribute to attaining this goal.

Further research evaluating the validity of the SIBAT and mobile applications for the assessment of suicidality could include the following. Studies incorporating larger populations, and assessing populations most likely to require comprehensive suicide risk assessment, may be most informative validating the use of these tools for research and clinical purposes. Studies in which the SIBAT and/or mobile application-based assessment tools are administered multiple times would also provide insight into whether these tools are able to detect changes over time. Factor analysis to identify areas of redundancy in the SIBAT may allow for item reduction or other modifications of the SIBAT.

Combining the SIBAT application with other methods of data collection, such as passive data collection and administrative data, and with new forms of data analysis, such as machine learning, may provide us with greater insight into the nature of suicidality and the optimal treatment approach for individual patients. Future studies could look at incorporating combinations of these methods in order to allow us to categorize presentations related to suicidality. Analyzing such data in conjunction with data on administered interventions and long-term outcomes (re-hospitalization, future suicide attempts, death by suicide) could also inform the best interventions for different subgroups. Further improvements to the SIBAT rating itself could also be considered. As discussed above, factor analysis and item reduction may be one direction to consider. Development of an adaptive questionnaire form of the SIBAT may also be

worth considering if future studies indicate low adherence to scale completion when it is administered repeatedly.

In summary, there are multiple directions that future research involving the SIBAT could take. Initial work would likely involve further validation of the SIBAT using larger populations and using the measure in participants at higher risk. Studies examining the tool's ability to detect changes over time would also help establish the validity of the measure both as a tool for assessment and to monitor treatment response. Studies combining the tool with other techniques may then be used to improve the measure and provide us with greater insight into suicidality.

References

1. Carter G, Milner A, McGill K et al. 2017. Predicting suicidal behaviours using clinical instruments: Systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry. 210(6):387-395.
2. Chan MK, Bhatti H, Meader N et al. 2016. Predicting suicide following self-harm: Systematic review of risk factors and risk scales. Br J Psychiatry. 209(4):277-283.
3. Runeson B, Odeberg J, Pettersson A et al. 2017. Instruments for the assessment of suicide risk: A systematic review evaluating the certainty of the evidence. PLoS One. 12(7):e0180292.
4. Steeg S, Quinlivan L, Nowland R et al. 2018. Accuracy of risk scales for predicting repeat self-harm and suicide: A multicentre, population-level cohort study using routine clinical data. BMC Psychiatry. 18(1):113.
5. American Psychiatric Association. 2016. Practice guidelines for the psychiatric evaluation of adults, third edition. Arlington (VA): American Psychiatric Association.
6. Boudreaux ED, Horowitz LM. 2014. Suicide risk screening and assessment: Designing instruments with dissemination in mind. Am J Prev Med. 47(3, Supplement 2):S163-S169.
7. Alphs L, Canuso C, Williamson DE et al. 2015. Suicide Ideation and Behavior Assessment Tool (SIBAT): A novel measure of suicidal ideation and behavior and perceived suicide risk. Poster presented at: International Society for CNS Clinical Trials and Methodology 2015 Autumn Conference. Amsterdam, The Netherlands.
8. Alphs L, Fu DJ, Williamson DE et al. 2018. Empirical evaluation of suicidal ideation and behavior categories established by the Columbia Classification Algorithm of Suicide Assessment (C-CASA). Poster presented at: 14th Annual Scientific Meeting of the International Society for CNS Clinical Trials and Methodology (ISCTM). Washington DC.
9. Alphs L, Williamson D, Turkoz I et al. 2018. Validation of Suicidal Ideation and Behavior Assessment Tool (SIBAT): Intra- and inter-rater reliability. Poster presented at: 17th European Symposium on Suicide and Suicidal Behaviour (ESSSB). Ghent, Belgium.
10. Alphs L, Williamson D, Turkoz I et al. 2018. Validation of Suicidal Ideation and Behavior Assessment Tool (SIBAT): Intra- and inter-rater reliability. Poster presented at: American College of Neuropharmacology. Hollywood, Florida, USA.
11. Turkoz I, Fu DJ, Canuso C et al. 2018. Suicidal Ideation and Behavior Assessment Tool (SIBAT): Psychometric characterization of patient-reported modules. Poster presented at: The International Society for Clinical Trials and Methodology, 2018 Autumn Conference. California, USA.
12. Williamson D, Canuso C, Fu DJ et al. 2017. 862. Patient report with the Suicide Ideation and Behavior Assessment Tool (SIBAT): Acceptability and sensitivity to rapid change. Biol Psychiatry. 81(10):S349.
13. Posner K, Brent D, Lucas C et al. 2009. Columbia-Suicide Severity Rating Scale (C-SSRS): Lifetime recent. 2009. [accessed 2020 April 14]. http://cssrs.columbia.edu/wp-content/uploads/C-SSRS-1-14-09-m9-12-17-Lifetime-Recent.doc.
14. Beck AT, Kovacs M, Weissman A. 1979. Assessment of suicidal intention: The scale for suicide ideation. J Consult Clin Psychol. 47(2):343-352.

15. Posner K, Brown GK, Stanley B et al. 2011. The Columbia-Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry. 168(12):1266-1277.
16. Torous J, Staples P, Shanahan M et al. 2015. Utilizing a personal smartphone custom app to assess the Patient Health Questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR Ment Health. 2(1):e8.
17. Larsen ME, Nicholas J, Christensen H. 2016. A systematic assessment of smartphone tools for suicide prevention. PLoS One. 11(4):e0152285.
18. Canadian community health survery, 2015. 2017. [accessed 2018 December 1, 2018]. https://www150.statcan.gc.ca/n1/daily-quotidien/170322/dq170322a-eng.htm.
19. Claassen CA, Harvilchuck-Laurenson JD, Fawcett J. 2014. Prognostic models to detect and monitor the near-term risk of suicide: State of the science. Am J Prev Med. 47(3 Suppl 2):S181-185.
20. Parikh R, Mathai A, Parikh S et al. 2008. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 56(1):45-50.
21. Patterson WM, Dohn HH, Bird J et al. 1983. Evaluation of suicidal patients: The SAD PERSONS scale. Psychosomatics. 24(4):343-349.
22. Neufeld E, O'Rourke N, Donnelly M. 2010. Enhanced measurement sensitivity of hopeless ideation among older adults at risk of self-harm: Reliability and validity of likert-type responses to the Beck Hopelessness Scale. Aging Ment Health. 14(6):752-756.
23. Green KL, Brown GK, Jager-Hyman S et al. 2015. The predictive validity of the Beck Depression Inventory suicide item. J Clin Psychiatry. 76(12):1683-1686.
24. Roaldset JO, Linaker OM, Bjørkly S. 2012. Predictive validity of the MINI suicidal scale for self-harm in acute psychiatry: A prospective study of the first year after discharge. Arch Suicide Res. 16(4):287-302.
25. Stefansson J, Nordstrom P, Jokinen J. 2012. Suicide Intent Scale in the prediction of suicide. J Affect Disord. 136(1-2):167-171.
26. Ryan CJ, Large MM. 2013. Suicide risk assessment: Where are we now? Med J Aust. 198(9):462-463.
27. Katz C, Randall JR, Sareen J et al. 2017. Predicting suicide with the SAD PERSONS scale. Depress Anxiety. 34(9):809-816.
28. Simon GE, Coleman KJ, Rossom RC et al. 2016. Risk of suicide attempt and suicide death following completion of the Patient Health Questionnaire depression module in community practice. J Clin Psychiatry. 77(2):221-227.
29. Simon GE, Rutter CM, Peterson D et al. 2013. Does response on the PHQ-9 depression questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv. 64(12):1195-1202.
30. Lindenmayer JP, Czobor P, Alphs L et al. 2003. The InterSePT scale for suicidal thinking reliability and validity. Schizophr Res. 63(1):161-170.
31. Potkin SG, Alphs L, Hsu C et al. 2003. Predicting suicidal risk in schizophrenic and schizoaffective patients in a prospective two-year trial. Biol Psychiatry. 54(4):444-452.
32. Ayer DW, Jayathilake K, Meltzer HY. 2008. The InterSePT suicide scale for prediction of imminent suicidal behaviors. Psychiatry Res. 161(1):87-96.
33. Bolton JM, Spiwak R, Sareen J. 2012. Predicting suicide attempts with the SAD PERSONS scale: A longitudinal analysis. J Clin Psychiatry. 73(6):e735-741.

34. Saunders K, Brand F, Lascelles K et al. 2014. The sad truth about the SADPERSONS scale: An evaluation of its clinical utility in self-harm patients. Emerg Med J. 31(10):796-798.
35. Yaseen ZS, Kopeykina I, Gutkovich Z et al. 2014. Predictive validity of the Suicide Trigger Scale (STS-3) for post-discharge suicide attempt in high-risk psychiatric inpatients. PLoS one. 9(1):e86768-e86768.
36. Galynker I, Yaseen ZS, Cohen A et al. 2017. Prediction of suicidal behavior in high risk psychiatric patients using an assessment of acute suicidal state: The suicide crisis inventory. Depress Anxiety. 34(2):147-158.
37. Haglund A, Lindh AU, Lysell H et al. 2016. Interpersonal violence and the prediction of short-term risk of repeat suicide attempt. Sci Rep. 6:36892.
38. Mundt JC, Greist JH, Gelenberg AJ et al. 2010. Feasibility and validation of a computer-automated Columbia-Suicide Severity Rating Scale using interactive voice response technology. J Psychiatr Res. 44(16):1224-1228.
39. Mundt JC, Greist JH, Jefferson JW et al. 2013. Prediction of suicidal behavior in clinical research by lifetime suicidal ideation and behavior ascertained by the electronic Columbia-Suicide Severity Rating Scale. J Clin Psychiatry. 74(9):887-893.
40. Madan A, Frueh BC, Allen JG et al. 2016. Psychometric reevaluation of the Columbia-Suicide Severity Rating Scale: Findings from a prospective, inpatient cohort of severely mentally ill adults. J Clin Psychiatry. 77(7):e867-873.
41. Bryan CJ, David Rudd M, Wertenberger E et al. 2014. Improving the detection and prediction of suicidal behavior among military personnel by measuring suicidal beliefs: An evaluation of the suicide cognitions scale. J Affect Disord. 159:15-22.
42. Stefansson J, Nordstrom P, Runeson B et al. 2015. Combining the Suicide Intent Scale and the Karolinska Interpersonal Violence Scale in suicide risk assessments. BMC Psychiatry. 15:226.
43. van Veen M, van Weeghel I, Koekkoek B et al. 2015. Structured assessment of suicide risk in a psychiatric emergency service: Psychometric evaluation of the Nurses' Global Assessment of Suicide Risk scale (NGASR). Int J Soc Psychiatry. 61(3):287-296.
44. Waern M, Sjostrom N, Marlow T et al. 2010. Does the Suicide Assessment Scale predict risk of repetition? A prospective study of suicide attempters at a hospital emergency department. Eur Psychiatry. 25(7):421-426.
45. Pokorny AD. 1983. Prediction of suicide in psychiatric patients. Report of a prospective study. Arch Gen Psychiatry. 40(3):249-257.
46. Rosen A. 1954. Detection of suicidal patients: An example of some limitations in the prediction of infrequent events. J Consult Psychol. 18(6):397-403.
47. Barzilay S, Yaseen ZS, Hawes M et al. 2018. Determinants and predictive value of clinician assessment of short-term suicide risk. Suicide Life Threat Behav. 49(2):614-626.
48. Randall JR, Sareen J, Chateau D et al. 2019. Predicting future suicide: Clinician opinion versus a standardized assessment tool. Suicide Life Threat Behav. 49(4):941-951.
49. Mental Health Commission of Canada. 2018. Research on suicide and its prevention: What the current evidence reveals and topics for future research. Ottawa, ON.
50. Nock MK, Park JM, Finn CT et al. 2010. Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. Psychol Sci. 21(4):511-517.
51. Randall JR, Rowe BH, Dong KA et al. 2013. Assessment of self-harm risk using implicit thoughts. Psychol Assess. 25(3):714-721.

52. Valenza G, Lanata A, Paradiso, R et al. 2014. Advanced technology meets mental health: How smartphones, textile electronics, and signal processing can serve mental health monitoring, diagnosis, and treatment. IEEE Pulse. 5(3):56-59.
53. Ravi D, Wong C, Deligianni F et al. 2017. Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics. 21(1):4-21.
54. American Psychiatric Association. 2013. Bipolar and related disorders. In: Diagnostic and statistical manual of mental disorders, fifth edition. Arlinghton, VA, American Psychiatric Association. p. 123-154.
55. O'Donnell LA, Deldin PJ, Grogan-Kaylor A, et al. 2017. Depression and executive functioning deficits predict poor occupational functioning in a large longitudinal sample with bipolar disorder. J Affect Disord. 215:135-142.
56. Parker G, McCraw S, Hadzi-Pavlovic D. 2015. Unipolar and bipolar patient responses to a new scale measuring the consequences of depression. Psychiatry Res. 230(2):676-681.
57. Rosa AR, Reinares M, Michalak EE, et al. 2010. Functional impairment and disability across mood states in bipolar disorder. Value Health. 13(8):984-988.
58. Alonso J, Petukhova M, Vilagut G, et al. 2011. Days out of role due to common physical and mental conditions: results from the WHO World Mental Health surveys. Mol Psychiatry. 16(12):1234-1246.
59. Grande I, Berk M, Birmaher B, Vieta E. 2016. Bipolar disorder. Lancet. 387(10027):1561-1572.
60. Francis A, Gasparo P. 1994. Interval between symptom onset and hospitalization in mania. J Affect Disord. 31(3):179-185.
61. Cassidy F. 2010. Insight in bipolar disorder: relationship to episode subtypes and symptom dimensions. Neuropsychiatr Dis Treat. 6:627-631.
62. Silva Rde A, Mograbi DC, Bifano J, Santana CM, Cheniaux E. 2016. Insight in bipolar mania: evaluation of its heterogeneity and correlation with clinical symptoms. J Affect Disord. 199:95-98.
63. Hirschfeld RM. 2002. The Mood Disorder Questionnaire: A simple, patient-rated screening instrument for bipolar disorder. Prim Care Companion J Clin Psychiatry. 4(1):9-11.
64. Altman E, Hedeker D, Peterson JL et al. 2001. A comparative evaluation of three self-rating scales for acute mania. Biol psychiatry. 50(6):468-471.
65. Kung S, Palmer BA, Lapid MI et al. 2015. Screening for bipolar disorders: Clinical utilization of the mood disorders questionnaire on an inpatient mood disorders unit. J Affect Disord. 188:97-100.
66. Areàn PA, Hoa Ly K, Andersson G. 2016. Mobile technology for mental health assessment. Dialogues in clinical neuroscience. 18(2):163-169.
67. Torous J, Friedman R, Keshavan M. 2014. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. JMIR mHealth and uHealth. 2(1):e2.
68. Stone AA, Shiffman S, Schwartz JE et al. 2003. Patient compliance with paper and electronic diaries. Controlled Clinical Trials. 24(2):182-199.
69. Ainsworth J, Palmier-Claus JE, Machin M et al. 2013. A comparison of two delivery modalities of a mobile phone-based assessment for serious mental illness: Native smartphone application vs text-messaging only implementations. J Med Internet Res. 15(4):e60.

70. Nicholas J, Larsen ME, Proudfoot J et al. 2015. Mobile apps for bipolar disorder: A systematic review of features and content quality. J Med Internet Res. 17(8):e198.
71. Firth J, Cotter J, Torous J, et al. 2016. Mobile phone ownership and endorsement of "mhealth" among people with psychosis: A meta-analysis of cross-sectional studies. Schizophrenia bulletin. 42(2):448-455.
72. Torous J, Chan SR, Yee-Marie Tan S et al. 2014. Patient smartphone ownership and interest in mobile apps to monitor symptoms of mental health conditions: A survey in four geographically distinct psychiatric clinics. JMIR mental health. 1(1):e5.
73. Boateng GO, Neilands TB, Frongillo EA et al. 2018. Best practices for developing and validating scales for health, social, and behavioral research: A primer. Front Public Health. 6:149-149.
74. Sterne JAC, Savovic J, Page MJ et al. 2019. RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ (Clinical research ed). 366:l4898.
75. Sterne JA, Hernán MA, Reeves BC et al. 2016. ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 355;i4919.
76. Depp CA, Kim DH, de Dios LV et al. 2012. A pilot study of mood ratings captured by mobile phone versus paper-and-pencil mood charts in bipolar disorder. J Dual Diagn. 8(4):326-332.
77. Faurholt-Jepsen M, Vinberg M, Frost M et al. 2015. Smartphone data as an electronic biomarker of illness activity in bipolar disorder. Bipolar disorders. 17(7):715-728.
78. Depp CA, Ceglowski J, Wang VC et al. 2015. Augmenting psychoeducation with a mobile intervention for bipolar disorder: A randomized controlled trial. J Affect Disord. 174:23-30.
79. Faurholt-Jepsen M, Frost M, Ritz C et al. 2015. Daily electronic self-monitoring in bipolar disorder using smartphones - the MONARCA I trial: A randomized, placebo-controlled, single-blind, parallel group trial. Psychol Med. 45(13):2691-2704.
80. Hidalgo-Mazzei D, Mateu A, Reinares M et al. 2016. Psychoeducation in bipolar disorder with a simple smartphone application: Feasibility, acceptability and satisfaction. J Affect Disord. 200:58-66.
81. Schwartz S, Schultz S, Reider A et al. 2016. Daily mood monitoring of symptoms using smartphones in bipolar disorder: A pilot study assessing the feasibility of ecological momentary assessment. J Affect Disord. 191:88-93.
82. Guo T, Xiang YT, Xiao L et al. 2015. Measurement-based care versus standard care for major depression: A randomized controlled trial with blind raters. Am J Psychiatry. 172(10):1004-1013.
83. Sanislow CA. 2016. Updating the research domain criteria. World Psychiatry. 15(3):222-223.
84. Public Health Agency of Canada. Suicide in Canada: Key statistics (infographic). [accessed 2019 September 19]. https://www.canada.ca/en/public-health/services/publications/healthy-living/suicide-canada-key-statistics-infographic.html.
85. Morrison KB, Laing L. 2011. Adults' use of health services in the year before death by suicide in Alberta. Health Rep. 22(3):15-22.
86. National Action Alliance for Suicide Prevention: Transforming Health Systems Initiative Work Group. 2018. Recommended standard care for people with suicide risk: Making health care suicide safe. Washington, DC: Education Development Center, Inc.

87. Hawgood J, De Leo D. 2016. Suicide prediction – a shift in paradigm is needed. Crisis. 37(4):251-255.
88. Giddens JM, Sheehan DV. 2014. Is there value in asking the question "do you think you would be better off dead?" In assessing suicidality? A case study. Innov Clin Neurosci. 11(9-10):182-190.
89. Kleiman EM, Turner BJ, Fedor S et al. 2017. Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. J Abnorm Psychol. 126(6):726-738.
90. Von Hippel PT. 2009. How to impute interactions, squares, and other transformed variables. Sociol Methodol. 39(1):265-291.
91. Streiner DL. 2003. Starting at the beginning: An introduction to coefficient alpha and internal consistency. J Pers Assess. 80(1):99-103.
92. Streiner DL. 2002. Breaking up is hard to do: The heartbreak of dichotomizing continuous data. Can J Psychiatry. 47(3):262-266.
93. Oquendo MA, Baca-Garcia E, Artés-Rodríguez A et al. 2012. Machine learning and data mining: Strategies for hypothesis generation. Mol Psychiatry. 17(10):956-959.
94. Kiev A. 1976. Cluster analysis profiles of suicide attempters. Am J Psychiatry. 133(2):150-153.
95. Kessler RC, Warner CH, Ivany C et al. 2015. Predicting suicides after psychiatric hospitalization in us army soldiers: The army study to assess risk and resilience in service members (Army STARRS). JAMA Psychiatry. 72(1):49-57.
96. Kim H, Kim B, Kim SH et al. 2018. Classification of attempted suicide by cluster analysis: A study of 888 suicide attempters presenting to the emergency department. J Affect Disord. 235:184-190.
97. Morales S, Barros J, Echávarri O et al. 2017. Acute mental discomfort associated with suicide behavior in a clinical sample of patients with affective disorders: Ascertaining critical variables using artificial intelligence tools. Front Psychiatry. 8:7.
98. Oh J, Yun K, Hwang JH et al. 2017. Classification of suicide attempts through a machine learning algorithm based on multiple systemic psychiatric scales. Front Psychiatry. 8:192.
99. Sinyor M, Schaffer A, Streiner DL. 2014. Characterizing suicide in Toronto: An observational study and cluster analysis. Can J Psychiatry. 59(1):26-33.
100. Walsh CG, Ribeiro JD, Franklin JC. 2017. Predicting risk of suicide attempts over time through machine learning. Clin Psychol Sci. 5(3):457-469.
101. Torous J, Larsen ME, Depp C et al. 2018. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: A review of current progress and next steps. Curr Psychiatry Rep. 20(7):51.
102. Kleiman EM, Turner BJ, Fedor S et al. 2018. Digital phenotyping of suicidal thoughts. Depress Anxiety. 35(7):601-608.
103. Torous J, Walker R. 2019. Leveraging digital health and machine learning toward reducing suicide-from panacea to practical tool. JAMA Psychiatry. 76(10):999–1000.
104. Matthews M, Doherty G, Sharry J et al. 2008. Mobile phone mood charting for adolescents. Br J Guid Counc. 36(2):113-129.
105. Sheehan DV, Lecrubier Y, Sheehan KH et al. 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry. 59(Suppl 20):22-33.

106. Kessler RC, Adler L, Ames M et al. 2005. The World Health Organization Adult ADHD Self-Report Scale (ASRS): A short screening scale for use in the general population. Psychol Med. 35(2):245-256.
107. Chin F, Leung CH. 2018. The concurrent validity of the Internet Addiction Test (IAT) and the Mobile Phone Dependence Questionnaire (MPDQ). PLoS One. 13(6):e0197562.
108. Faraci P, Craparo G, Messina R et al. 2013. Internet Addiction Test (IAT): Which is the best factorial solution? J Med Internet Res. 15(10):e225.
109. Jelenchick LA, Becker T, Moreno MA. 2012. Assessing the psychometric properties of the Internet Addiction Test (IAT) in US college students. Psychiatry Res. 196(2-3):296-301.
110. Widyanto L, McMurran M. 2004. The psychometric properties of the Internet Addiction Test. Cyberpsychol Behav. 7(4):443-450.

Appendix 1: Sample Question Stems from SIBAT

- The number of members in my family who have died by suicide is…
- Over the past 7 days I have felt hopeful.
- Over the past 7 days my sleep has been good.
- I have drunk alcohol on ___ of the past 7 days.
- I am glad to be alive.
- Nothing in life gives me pleasure.
- Nobody will care if I am dead.
- My emotional (mental) pain is so severe that I want to end my life.
- I have been shamed and should die.
- Which of the following ratings best describes your intent to end your life in the past 7 days?

Appendix 2: Item-total Correlations of SIBAT Module 2

	Item-Total Correlation
M2Q1	.557
M2Q2	.568
M2Q3	.718
M2Q4	.724
M2Q5	.585
M2Q6	.712
M2Q7	.614
M2Q8	.728
M2Q9	.587
M2Q10	.787
M2Q11	.684
M2Q12	.734
M2Q13	.424
M2Q14	.414
M2Q15	.552
M2Q16	.522
M2Q17	.585
M2Q18	.653
M2Q19	.536
M2Q20	.224
M2Q21	.275

Appendix 3: Item-total Correlations of SIBAT Module 3

	Item-Total Correlation
M3Q1	.745
M3Q2	.753
M3Q3	.573
M3Q4	.666
M3Q5	.093
M3Q6	.658
M3Q7	-.200
M3Q8	.545
M3Q9	.636
M3Q10	.611
M3Q11	.790
M3Q12	.635
M3Q13	.610
M3Q14	.755
M3Q15	.740
M3Q16	.443
M3Q17	.401
M3Q18	.714
M3Q19	.783
M3Q20	.506
M3Q21	.667
M3Q22	.504
M3Q23	.731
M3Q24	.727

M3Q25	.510
M3Q26	.715
M3Q27	.484
M3Q28	.611
M3Q29	.282
M3Q30	.641
M3Q31	.693
M3Q32	.467
M3Q33	.528
M3Q34	.277
M3Q35	.640
M3Q36	.665
M3Q37	.341
M3Q38	.755
M3Q39	.225
M3Q40	.560
M3Q41	.584
M3Q42	.586
M3Q43	.590
M3Q44	.677
M3Q45	.669
M3Q46	-.029
M3Q47	.619
M3Q48	.556

Appendix 4: Item level comparison of SIBAT Modules 2, 3 and 5 by mode of entry

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of SibB_1 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.589	Retain the null hypothesis.
2	The distribution of SibB_2 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.492	Retain the null hypothesis.
3	The distribution of SibB_3 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.410	Retain the null hypothesis.
4	The distribution of SibB_4 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.591	Retain the null hypothesis.
5	The distribution of SibB_5 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.385	Retain the null hypothesis.
6	The distribution of SibB_6 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.852	Retain the null hypothesis.
7	The distribution of SibB_7 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.856	Retain the null hypothesis.
8	The distribution of SibB_8 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.890	Retain the null hypothesis.
9	The distribution of SibB_9 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.575	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
10	The distribution of SibB_10 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.444	Retain the null hypothesis.
11	The distribution of SibB_11 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.135	Retain the null hypothesis.
12	The distribution of SibB_12 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.407	Retain the null hypothesis.
13	The distribution of SibB_13 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.879	Retain the null hypothesis.
14	The distribution of SibB_14 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.478	Retain the null hypothesis.
15	The distribution of SibB_15 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.950	Retain the null hypothesis.
16	The distribution of SibB_16 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.649	Retain the null hypothesis.
17	The distribution of SibB_17 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.663	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
18	The distribution of SibB_18 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.707	Retain the null hypothesis.
19	The distribution of SibB_19 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.060	Retain the null hypothesis.
20	The distribution of SibB_20 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.435	Retain the null hypothesis.
21	The distribution of SibB_21 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.186	Retain the null hypothesis.
22	The distribution of SibC_1 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.645	Retain the null hypothesis.
23	The distribution of SibC_2 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.964	Retain the null hypothesis.
24	The distribution of SibC_3 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.464	Retain the null hypothesis.
25	The distribution of SibC_4 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.803	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
26	The distribution of SibC_5 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.511	Retain the null hypothesis.
27	The distribution of SibC_6 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.832	Retain the null hypothesis.
28	The distribution of SibC_7 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.932	Retain the null hypothesis.
29	The distribution of SibC_8 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.824	Retain the null hypothesis.
30	The distribution of SibC_9 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.554	Retain the null hypothesis.
31	The distribution of SibC_10 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.251	Retain the null hypothesis.
32	The distribution of SibC_11 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.382	Retain the null hypothesis.
33	The distribution of SibC_12 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.416	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
34	The distribution of SibC_13 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.975	Retain the null hypothesis.
35	The distribution of SibC_14 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.306	Retain the null hypothesis.
36	The distribution of SibC_15 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.935	Retain the null hypothesis.
37	The distribution of SibC_16 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.387	Retain the null hypothesis.
38	The distribution of SibC_17 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.536	Retain the null hypothesis.
39	The distribution of SibC_18 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.147	Retain the null hypothesis.
40	The distribution of SibC_19 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.393	Retain the null hypothesis.
41	The distribution of SibC_20 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.125	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
42	The distribution of SibC_21 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.280	Retain the null hypothesis.
43	The distribution of SibC_22 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.626	Retain the null hypothesis.
44	The distribution of SibC_23 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.962	Retain the null hypothesis.
45	The distribution of SibC_24 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.156	Retain the null hypothesis.
46	The distribution of SibC_25 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.605	Retain the null hypothesis.
47	The distribution of SibC_26 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.446	Retain the null hypothesis.
48	The distribution of SibC_27 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.632	Retain the null hypothesis.
49	The distribution of SibC_28 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.118	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
50	The distribution of SibC_29 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.455	Retain the null hypothesis.
51	The distribution of SibC_30 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.181	Retain the null hypothesis.
52	The distribution of SibC_31 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.471	Retain the null hypothesis.
53	The distribution of SibC_32 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.372	Retain the null hypothesis.
54	The distribution of SibC_33 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.562	Retain the null hypothesis.
55	The distribution of SibC_34 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.827	Retain the null hypothesis.
56	The distribution of SibC_35 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.137	Retain the null hypothesis.
57	The distribution of SibC_36 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.487	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
58	The distribution of SibC_37 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.307	Retain the null hypothesis.
59	The distribution of SibC_38 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.973	Retain the null hypothesis.
60	The distribution of SibC_39 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.746	Retain the null hypothesis.
61	The distribution of SibC_40 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.375	Retain the null hypothesis.
62	The distribution of SibC_41 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.866	Retain the null hypothesis.
63	The distribution of SibC_42 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.208	Retain the null hypothesis.
64	The distribution of SibC_43 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.756	Retain the null hypothesis.
65	The distribution of SibC_44 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.748	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
66	The distribution of SibC_45 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.296	Retain the null hypothesis.
67	The distribution of SibC_46 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.754	Retain the null hypothesis.
68	The distribution of SibC_47 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.659	Retain the null hypothesis.
69	The distribution of SibC_48 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.465	Retain the null hypothesis.
70	The distribution of SibD_1 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.978	Retain the null hypothesis.
71	The distribution of SibD_2 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.421	Retain the null hypothesis.
72	The distribution of SibD_3 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.995	Retain the null hypothesis.
73	The distribution of SibD_4 is the same across categories of Device Used.	Independent-Samples Mann-Whitney U Test	.551	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Appendix 5: Publications arising from this work

Chan EC, Sun Y, Aitchison KJ, Sivapalan S. Assessment and monitoring of bipolar disorder using mobile app-based self-report questionnaires: a review of current evidence. Manuscript submitted for publication.

Chan EC, Wallace, K, Yang E, Roper L, Aryal G, Baskys A, Isenberg R, Carnes PJ, Green BA, Lodhi RJ, Aitchison KJ. The feasibility and acceptability of mobile application-based assessment of suicidality using a novel tool, the Suicide Ideation and Behavior Assessment Tool (SIBAT). Manuscript in submission.

Chan EC, Wallace, K, Yang E, Roper L, Aryal G, Baskys A, Isenberg R, Carnes PJ, Green BA, Aitchison KJ. Internal consistency and concurrent validity of a new tool for the assessment of suicidality, the Suicide Ideation and Behavior Assessment Tool (SIBAT). Manuscript in submission.