JMIR Mental Health

Journal Impact Factor (JIF) (2023): 4.8
Volume 9 (2022), Issue 1 ISSN 2368-7959 Editor in Chief: John Torous, MD, MBI

Contents

Viewpoint

Original Papers

Viewpoint

A Novel Peer-to-Peer Coaching Program to Support Digital Mental Health: Design and Implementation

Benjamin M Rosenberg[1*], MA, CPhil; Tamar Kodish[1,2*], MA, CPhil; Zachary D Cohen[3], PhD; Elizabeth Gong-Guy[2], PhD; Michelle G Craske[1,3], PhD

[1]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, United States

[2]Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, United States

[3]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA, United States

[*]these authors contributed equally

Corresponding Author:
Benjamin M Rosenberg, MA, CPhil
Department of Psychology
University of California, Los Angeles
1285 Franz Hall
Los Angeles, CA, 95030
United States
Phone: 1 4083068603
Email: benrosenberg@g.ucla.edu

Abstract

Many individuals in need of mental health services do not currently receive care. Scalable programs are needed to reduce the burden of mental illness among those without access to existing providers. Digital interventions present an avenue for increasing the reach of mental health services. These interventions often rely on paraprofessionals, or coaches, to support the treatment. Although existing programs hold immense promise, providers must ensure that treatments are delivered with high fidelity and adherence to the treatment model. In this paper, we first highlight the tension between the scalability and fidelity of mental health services. We then describe the design and implementation of a peer-to-peer coach training program to support a digital mental health intervention for undergraduate students within a university setting. We specifically note strategies for emphasizing fidelity within our scalable framework, including principles of learning theory and competency-based supervision. Finally, we discuss future applications of this work, including the potential adaptability of our model for use within other contexts.

(JMIR Ment Health 2022;9(1):e32430) doi:10.2196/32430

KEYWORDS

peer support; digital mental health; university students; college students; training and supervision; scalable psychological interventions

Mental Health: A Global Crisis

Background

Mental illness is a pressing and growing global public health crisis with enormous societal costs [1]. Between 1990 and 2017, the number of cases of depression worldwide grew from 172 to 258 million [2]. Unfortunately, the majority of people in need of treatment do not receive care, due to a multitude of factors that reduce availability and accessibility of mental health services [3]. For instance, worldwide, shortages in trained professionals and resources allocated for mental health care limit access to treatment [4]. Although evidence-based treatments (EBTs) exist for mental health disorders, there is a major lag in translation of these treatments from laboratories to the real world [5]. Projections indicate that significant shortages of mental health practitioners will continue throughout the next decade, underscoring the need for innovative and scalable solutions to deliver EBTs [6,7].

One widely studied scalable approach, used most prominently in low-resource contexts, is for paraprofessionals to provide or support the delivery of scalable mental health services [8,9]. In this paper, we use the term “paraprofessionals” to refer to nonspecialists without formal mental health credentials who are trained to provide or support low-intensity mental health services in community settings. Under this umbrella, we include individuals who have been described using a variety of terms, such as “coaches,” “lay providers,” “community health workers,” and “peer specialists” [10-12]. Although

paraprofessional support models represent a clear pathway to increasing access to care, little is known about the training, quality of care delivery, and sustainability of these models.

Digital mental health innovations via phone, computers, and other electronic devices offer another pathway for increasing access to care [13]. Digital mental health interventions hold particular promise for individuals who face obstacles to traditional, face-to-face mental health services, such as stigma, financial difficulties, time constraints, and location of services [14]. Although user uptake, engagement, and dropout have been problematic for digital mental health interventions [15], especially in routine clinical care settings [16], these problems can be addressed via human support [17-19].

Accordingly, mental health care models that combine paraprofessional workforces and digital mental health innovations have unique potential to expand the reach of and engagement with high-quality EBTs. One key consideration in efforts to design and implement paraprofessional-supported digital mental health interventions involves balancing *scalability*, to maximize intervention reach, with *fidelity*, to optimize quality and standards of treatment delivery. Scalability can be defined as "the capacity of an intervention to be applied in a way that reaches a large number of people" [6]. Fidelity encompasses both adherence (ie, Was the intervention delivered as intended?) and competence (ie, How skillfully was the intervention delivered?) [20] to ensure that patients receive efficacious treatment that leads to improved mental health outcomes [21].

Study Aim

The purpose of this paper is to demonstrate 1 way of designing a coaching program that maintains a focus on the fidelity and delivery of high-quality EBTs, while preserving key strengths of paraprofessional models of care, including scalability. Our program was developed to support the delivery of a digital mental health intervention [22] on college campuses, where rates of mental health problems are rapidly growing [23]. Given the current state of the literature, we first describe gaps in our knowledge about the fidelity of treatment delivery within existing paraprofessional programs, such as peer-to-peer support programs. Next, we highlight how pairing digital mental health innovations with paraprofessional support can increase the fidelity and scalability of mental health treatment. Third, we describe our approach to the design and implementation of a peer-to-peer training program, emphasizing potential avenues for optimizing learning processes to enhance the fidelity of treatment delivery.

Paraprofessional Mental Health Delivery Paradigms

Scalability and Fidelity

Paraprofessional models have gained widespread attention and support as scalable models of mental health service delivery with great potential to address unmet needs for care [8,24]. Evidence suggests that mental health interventions can be feasibly, acceptably, and effectively delivered by paraprofessionals in low-resource settings [13]. Paraprofessional training programs have the added benefit of increasing the clinical workforce, as these individuals often move on to receive advanced training in the clinical field after serving as paraprofessionals [25].

Fidelity-monitoring practices have the capacity to increase therapist accountability in service of promoting treatment adherence and competence [26]. Indeed, greater therapist competence has been associated with superior treatment outcomes [27]. However, numerous challenges with fidelity monitoring have been identified in the context of paraprofessional service delivery [8,28], such that existing paraprofessional care programs have focused primarily on scalability needs, with less attention given to fidelity of service delivery [29]. Given pressing demands to rapidly reach millions of underserved individuals in need, even paraprofessional interventions that are supported by research and contain evidence-based strategies often lack consistent fidelity-monitoring and quality assurance procedures. For instance, only 38% of studies in a review of community health worker–delivered interventions described procedures for fidelity monitoring, and among those that did report a monitoring procedure, the review noted significant variability in levels, methods, and assessment tools for fidelity measurement [8].

The financial and human resources needed to support fidelity monitoring in real-world contexts are often not available, limiting the external validity of many fidelity-monitoring strategies typically used in clinical trials [30]. Even when fidelity and quality assurance checks are integrated into training and supervision within paraprofessional models, sustained fidelity monitoring is often restricted due to limited supervision and insufficient resources to ensure continued quality assurance [28,30]. Paraprofessional programs delivered with less fidelity monitoring are thought to reduce intervention efficacy [27] and may discourage participants from future engagement in treatment. Randomized control trials have shown that with adequate training and ongoing supervision, paraprofessionals have the capacity to deliver interventions with similar levels of fidelity compared to mental health professionals [31,32]. However, less is known about how to design and implement high-fidelity training programs in more scalable contexts. Qualitative research suggests that lay health workers involved in mental health service delivery state a desire for more robust supervision. Yet, training and supervision best practices have not been established to date [33]. The limited research describing training and supervision procedures in paraprofessional delivery paradigms underscores the need for innovative solutions that have dual goals of sustaining potential for scalability, while also ensuring the fidelity of intervention delivery.

Pairing Technological Innovation With Paraprofessional Support to Enhance Fidelity and Scalability

Digital therapies hold significant promise for addressing problems with fidelity and bridging gaps in care access within wide-scale implementation efforts [27,30]. In particular, these approaches offer 1 way to support treatment delivery, paraprofessional training, and supervision, while minimizing human error or therapist drift, a common phenomenon in manualized treatment protocols [34]. Although humans often

play a smaller role within digital therapy models relative to traditional face-to-face therapy, human support or coaching has been shown to augment the efficacy of digital interventions [35]. This is particularly important, given the many challenges and barriers associated with implementation of digital therapies, including limited engagement, poor rates of retention, lack of personalization, and significant cognitive load [15,36]. The involvement of human support increases intervention flexibility and acceptability by calibrating the fit between digital tools and users' lived experiences, thereby boosting user engagement and retention [18,37]. Lattie et al [38] provide recommendations for the development of text-based coaching protocols (eg, [39]) to support digital mental health interventions and ensure high-fidelity treatment delivery. Thus, pairing paraprofessional coach support with digital therapies has several notable advantages that attend to the need for scalable innovations, while simultaneously emphasizing fidelity.

Peer-to-Peer Support

One consideration in designing paraprofessional models is who should be trained to provide, or support the delivery of, mental health interventions. A prominent model focuses on training of peer-to-peer specialists, or peer coaches [40]. Peer coaching models have been used to provide services or support to individuals with whom coaches share communities, identities, or lived experiences, with the goal of enhancing accessibility, engagement, and scalability of interventions [41]. In doing so, these models have the potential to overcome obstacles to care, such as lack of trust, stigma, and cultural and linguistic barriers (although the significance of peers' own lived experiences is yet to be determined). One common example is peer recovery and support for individuals with substance use disorders [42], where a peer's own experience and personal knowledge is harnessed to support individuals in starting and maintaining the recovery process [43-45]. Key legislation is paving the way to expand peer specialist programs to address a variety of population mental health needs, such as the 2020 California Senate Bill SB-803: Mental Health Services: Peer Support Specialist Certification.

Yet, a major barrier to broader implementation of peer support is the mixed empirical support for these models [46-49]. There is some evidence to suggest more positive effects from formal, structured peer support (eg, [50-53]) than informal support (eg, online chat forums) [54,55]. Nonetheless, the findings are inconsistent even within structured peer support interventions (eg, [56]). Methodological inconsistencies may partly explain the disparate findings [42,56], and 1 major example is training and quality assurance. Standardized procedures for peer training, certification, and fidelity monitoring are not well described in the literature [47,56]. Well-defined and replicable methods for training and quality assurance procedures are sorely needed.

Design of Coach Training Programs

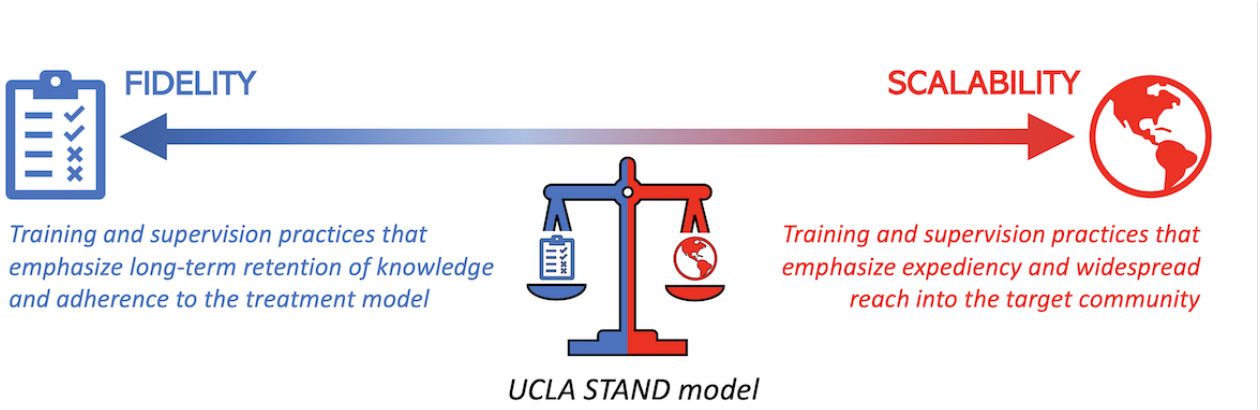
Overview

In 2015, the University of California, Los Angeles (UCLA) launched a campus-wide research initiative, the Depression Grand Challenge (DGC), with the goal of cutting the burden of depression in half by 2050. The DGC comprises a number of studies that seek to uncover mechanisms underlying depression and to develop novel treatments and innovative approaches to treatment implementation. To begin tackling this problem at UCLA, the DGC launched the Screening and Treatment for Anxiety and Depression (STAND) program for UCLA students in fall 2017 (Figure 1). The STAND program provides all UCLA students with free mental health screening and tiered care, including digital cognitive-behavioral therapy (CBT) with certified peer coach support for students experiencing mild-to-moderate symptoms of depression and mild-to-severe symptoms of anxiety, as well as in-person psychotherapy and pharmacotherapy for students experiencing severe symptoms of depression. Students who enroll in the digital CBT arm are offered coaching from certified peers, provided via 30-minute weekly coaching sessions in which they review and troubleshoot the application of module content and skills.

STAND Digital Therapy is a modular program that combines interventions for depression, sleep, panic/agoraphobia, social anxiety, worry (generalized anxiety disorder), and trauma (posttraumatic stress disorder), drawing upon existing evidence-based programs [57-66]. There are 13 available packages that cover all principal disorders and critical patterns of comorbidity (eg, depression + sleep, trauma + depression) and comprise 6-8 modules, depending on the number of disorders targeted. Individuals are assessed at baseline on an adaptive battery of disorder-specific, self-report questionnaires that guide the package selection process [22]. The personalized packages are built to maximize engagement and interactivity and with a strong focus on diversity and inclusion. The modules are transdiagnostic and skill focused, involving psychoeducation, in session exercises, and between-session practice of techniques, including behavioral activation, cognitive restructuring, self-compassion, and exposure (eg, in vivo, interoceptive, imaginal).

Fitting within this model, the initial development of our coach training program specifically targets UCLA undergraduate students as both coaches and recipients of the intervention, consistent with the peer support models described before. Enrollment as a coach trainee does not rely on any prerequisite coursework, history of service provision, or experience of personal mental health concerns or psychotherapy. Training and supervision of coaches are provided by graduate students in the clinical psychology doctoral program at UCLA for all stages of coach training. Graduate supervisors attend group supervision-of-supervision with a licensed clinical psychologist (author EGG).

Figure 1. Navigating scalability and fidelity in mental health coaching programs. STAND: Screening and Treatment for Anxiety and Depression; UCLA: University of California, Los Angeles.



Program Description

In our program, coach training occurs in weekly sessions, wherein trainees review digital CBT content, engage in didactic instruction of coaching materials, and complete role-play exercises focusing on basic interpersonal process skills. Coaches move through 4 primary phases of training: (1) beginner, (2) intermediate, (3) advanced, and (4) certified. Weekly training consists of a 2-hour training session as well as 2 hours of assignments completed between training sessions. Each level of training is completed over 1 academic quarter (10 weeks), at which point trainees are advanced to the subsequent level of training based on supervisor evaluations.

Beginner-Level Training

The goals of the beginner phase of training are to (1) introduce coaches to digital CBT content and increase knowledge of the intervention and (2) provide early practice with interpersonal process skills to initiate the process of translating declarative knowledge during coaching delivery. In service of these aims, beginner-level trainees enroll as users of the digital CBT and advance through the digital CBT content themselves, completing homework exercises associated with the program and reading foundational material on cornerstone CBT topics between didactic training sessions. In addition, beginner-level trainees are introduced to 6 core interpersonal process skills that are routinely assessed to monitor coaching effectiveness throughout the coach training program: (1) authenticity, (2) nonverbal skills, (3) open-ended questioning, (4) reflecting emotions, (5) content summaries, and (6) collaborative inquiry [67-69]. These process skills, in addition to sustained knowledge of the digital CBT content, provide the foundation for advancement throughout the coach training program.

Beginner-level trainees participate in (1) didactics regarding digital CBT content and interpersonal process skills, (2) discussions regarding other cornerstone topics (eg, mindfulness, cultural humility, trauma-informed care, ethics), and (3) role-play exercises to begin practicing application of the 6 core interpersonal process skills. Beginner trainees also attend sessions with advanced trainees, in which they serve as mock or practice participants for advanced trainees who are coaching full mock sessions (described in detail in the Advanced-Level Training section). Role-play exercises are recorded or observed live by supervisors, who provide oral and written feedback, as well as numerical ratings on each interpersonal process skill (eg, scale from 1 to 10 with behavioral anchors; see Multimedia Appendix 1). These evaluations provide benchmarks for certification and highlight areas of growth as trainees progress toward certification throughout the program.

Intermediate-Level Training

As trainees progress into the intermediate stage of the program, the primary goals are to provide trainees with intensive practice, (1) translating knowledge into coaching delivery and (2) applying interpersonal process skills to support engagement with digital CBT content. During these sessions, trainees participate in (1) brief digital CBT module content review, (2) intensive role-play exercises applying core process skills, and (3) introduction to protocols for managing advanced clinical issues (eg, suicidality, homicidality, abuse).

To continue supporting trainee development of interpersonal process skills and digital CBT content knowledge, trainees are continually rated on their process skills throughout intensive role-plays. Each week, supervisors review trainees' intensive role-play segments and provide trainees with written feedback and numerical ratings on core interpersonal process skills. In addition, group supervision sessions incorporate oral feedback from supervisors and peer coaches, including in vivo corrective feedback during role-play exercises.

Advanced-Level Training

Once trainees reach the advanced stage, the main goal is for trainees to achieve certification to serve as coaches for participants. This is accomplished by demonstrating (1) competency across all 6 core interpersonal process skills and (2) continued knowledge of digital CBT content. Advanced trainees conduct practice coaching sessions (ie, full 30 minutes) with beginner trainees as mock participants. In addition to these practice sessions, advanced trainees attend a weekly supervision group consisting of intensive role-play exercises, with role-play targets focused on digital CBT content, interpersonal process skills, and management of advanced clinical issues (eg, suicidality, homicidality, abuse, sexual assault, self-disclosure).

To support advanced coaches in progressing toward certification, advanced-level trainees receive written and numerical ratings on their full 30-minute practice coaching sessions. These ratings are used to certify trainees on competency across all process skills. Next, trainees achieve certification on digital CBT content by passing quizzes, which ensures knowledge of the intervention and promotes continued fidelity to the treatment model.

Coach Certification

Following successful advancement through the prior 3 stages of the program, trainees are certified to support the digital CBT with continued supervision. Certified trainees who are engaged in coaching continue to attend weekly supervision groups in which they discuss coaching sessions with their supervisor and peers. To ensure continued fidelity to coaching standards, supervisors review video recordings of each coaching session and rate the coaches' application of process skills according to the behavioral rating scale described before. Video review further enables supervisors to use didactics and role-play exercises in response to common challenges or to address drift from the coaching protocol. Certified trainees additionally provide feedback to the supervision team to inform potential future iterations of the coaching program.

Strategies for Monitoring and Enhancing Fidelity

Learning Theory

Increased attention to trainee learning processes within mental health provider training and supervision procedures has potential to increase fidelity to EBTs [70]. One way to enhance paraprofessional mental health service delivery, therefore, is to design training programs leveraging insights from learning theory and the use of specific pedagogical strategies (see Table 1 for examples) shown to improve knowledge building, skill acquisition, and long-term retention across domains such as learning a new language, mathematics, and sports [71-73]. Although these strategies may reduce performance in the short term (ie, during initial acquisition of skills or knowledge), research has consistently shown superior long-term retention and retrieval of learning [72,74].

Table 1. Pedagogical strategies and examples.

Principle	Definition	Example
Varying context of learning	Incorporating contextual variability (eg, physical location, types of teaching strategies) into teaching and learning	Compared with individuals who repeatedly study in 1 setting, individuals who study in a variety of physical settings have been shown to perform better on subsequent examinations in a new setting [75].
Spaced instruction	Spacing out instruction of a single topic over a period, as opposed to solely providing instruction about a topic in 1 learning event	Although cramming for an exam may be a useful strategy for performing well in the short term (eg, on a quiz), spacing the presentation of materials over a longer period has been shown to support performance in the long term (eg, on a final examination).
Interleaved instruction	Interleaving instruction of different topics within a common learning event (eg, covering multiple concepts within a single class)	Interleaving questions that assess knowledge of multiple concepts (eg, geometric equations for angles and lines intermixed) has been shown to improve student learning compared with blocking of concepts (eg, equations for angles, then lines) [76].
Retrieval practices/ examinations	Formal assessment of knowledge (eg, tests, assessments, exams)	Individuals who make incorrect guesses have been shown to benefit from these early mistakes during learning compared with individuals who are provided with the correct answers from the beginning of training [77].

Learning Theory: Applied

From the outset of coach training, we have applied core principles of learning theory to guide the instruction of digital CBT content and process skills. For example, *variability of learning contexts* is applied through (1) independent trainee review of digital CBT content (outside of sessions), (2) didactic training (during sessions), (3) role-play exercises (conducted in small groups), and (4) participation in mock sessions (observed by the entire supervision group). Likewise, applying the principle of *spaced instruction*, digital CBT content and interpersonal process skills are introduced and revisited at multiple timepoints within and across training levels. *Interleaved instruction* is similarly used to promote initial learning of digital CBT content and process skills simultaneously (eg, a single training session alternates between CBT and process skill content, and likewise combines the 2 domains, rather than blocking 1 instruction topic at a time). Furthermore, *retrieval practices* assess digital CBT knowledge throughout all stages of trainee development to support long-term retention of learning (eg, during the advanced stage of coach training, the process of obtaining certification involves trainees repeatedly completing mock coaching sessions with corrective feedback).

Following certification, ongoing fidelity-monitoring practices include (1) completion of a self-evaluation coaching checklist following all coaching sessions, (2) discussion of coach adherence to the digital CBT module during supervision, and (3) continued completion of mock coaching sessions during supervision with peer-to-peer and supervisor feedback.

Competency-Based Supervision

Following the acquisition of new knowledge and skills, competency-based supervision techniques can provide trainees with a pathway for transforming declarative knowledge into procedural knowledge [78-81]. Prior studies support the notion that competency-based supervision can increase effective CBT knowledge and acquisition [82]. Accordingly, the present coach training program integrates experiential learning and competency-based supervision strategies to support sustained fidelity to the treatment. For example, our program uses supervision practices that integrate a variety of experiential learning techniques (eg, skill modeling, role-plays, and

corrective feedback), which have been shown to increase provider fidelity to EBTs [70]. Likewise, the program continuously assesses and monitors trainee development with clearly articulated, behaviorally anchored feedback [81].

Discussion

Principal Findings

In this paper, we outlined 1 example of a scalable peer-to-peer mental health paraprofessional training and supervision program. Although many models of paraprofessional support have been described and tested previously, high demand and minimal resources have often corresponded with a reduced focus on fidelity monitoring and quality assurance [8]. Lack of standardized methods for paraprofessional training and supervision may have contributed to the disparate empirical support for paraprofessional, and specifically peer paraprofessional, models. Here we described a standardized and replicable model of training and supervision suitable for evaluation.

Strengths

We believe this model has several notable strengths. Of note, our program focuses explicitly on fidelity, while also attending to the need for scalable care. As illustrated, the focus on fidelity is integrated into the program in 2 primary ways: digital technology as the primary agent for CBT content delivery [83] and continuous, standardized procedures for fidelity monitoring of coaches who support digital CBT provision. In addition, our training and supervision program is grounded in key findings from the learning theory literature, aligned with data suggesting that optimized learning can serve as a pathway to higher fidelity of treatment delivery [70,78]. The integration of learning theory as a mechanism for enhancing fidelity is aligned with existing lay health worker training frameworks that focus on augmenting initial one-off training with on-the-job direct supervision, coaching, and feedback systems [28]. We believe that paraprofessional models anchored in learning theory principles have the greatest potential to improve quality of care.

Another strength is that our program is designed to be malleable and can be adapted in various ways based on implementation context factors. Along with fidelity, program flexibility is well established as a key ingredient to successful implementation of interventions in numerous settings [84,85]. Implementation science frameworks have frequently cited the importance of balancing both fidelity and flexibility in delivery of EBTs, and this concept has also been established as essential in lay health worker models [28]. Our program was designed with flexibility within fidelity as a key guiding principle. It contains both *core components*, defined within the Consolidated Framework for Implementation Research (CFIR) as the "essential and indispensable elements" of the program, and the *adaptable periphery*, defined as the aspects of the program that can be modified and varied from site to site [86,87]. Included in our program's core components are (1) anchoring in principles of learning theory described before, (2) training on 6 core clinical process skills, and (3) training on digital CBT content. The adaptable periphery, however, depends on the structures, systems, and contexts involved with program implementation. In the process of designing adaptations, community stakeholder partnership and input are essential [88]. Although many adaptation frameworks have focused on adaptations to the intervention itself, stakeholders can also be used to consider adaptations to the implementation context.

In our program, we have identified several components of the adaptable periphery that have been tailored for various implementation contexts, with community partnership. For instance, although this paper describes implementation at 1 university, we are currently piloting coach training and supervision for the launch of STAND digital CBT in numerous other types of community settings, including local community colleges and health care systems. In partnership with community stakeholders, 1 example of a component in the adaptable periphery that we have modified to meet the needs of a new implementation site is the length of training time, which has been shortened to accommodate local resources. This has been accomplished by combining components of beginner and intermediate levels of training and including additional review and feedback of recorded role-plays outside of sessions to accelerate learning and growth. In another example of adaptation, we have worked with various sites to situate and design our coaching risk protocols (eg, suicide risk, abuse) within the contexts of existing resources, infrastructure, and referrals. Another example of adaptation has been to integrate specific training on trauma-informed care strategies to support implementation of this program in communities with higher trauma prevalence rates. Cultural considerations are also essential, particularly in planning implementation of coach training programs in diverse settings such as ours. Working in partnership with community stakeholders to co-design cultural adaptations can lead to improved program acceptability and community engagement. Although we have made and discussed modifications within the adaptable periphery based on the unique implementation and contextual factors within various environments, the same guiding principles described in this paper serve as the foundational core components across settings.

A final strength of our program is that it is intended not only to train students to serve as coaches to their peers but also to provide critical CBT skills to trainees themselves. Many coaches in our program anecdotally report that their experience throughout training has taught them invaluable interpersonal and cognitive-behavioral skills. In the broader literature, paraprofessionals describe feeling that their training experiences were associated with personal development and growth and increases in knowledge, self-confidence, and skill use [33]. In the context of our program, formal measurement of mental health benefits conferred by coaches in our program is needed.

Limitations

Several key limitations of our program should also be noted. First, because this program is situated within the scope of a large research initiative, ongoing funding has been available to sustain coach training and supervision. Beyond the realm of research, efforts to provide continuous funding for paraprofessional support programs in routine care settings are critical. In the initial iteration of our program, coaches have served as volunteers, engaged in all program elements as an

additional responsibility outside of their other obligations. Data suggest that among volunteer staff supporting digital interventions, administrative issues, such as time constraints, may contribute to barriers to training completion and attrition [89]. Additional funding that encompasses financial payment or other incentives for peer coaches may represent 1 solution to address this obstacle. One model that is currently being tested as a component of our program's adaptable periphery is paying coaches as university employees. Alternative methods of expanding and sustaining funding and resources are worthy of exploration.

Second, although we maintain a focus on fidelity in our program, the primary objective of our peer-to-peer program is to serve as a scalable model of care in real practice settings. Thus, given the resource constraints of real-world implementation contexts, we have designed our fidelity-monitoring procedures to minimize supervisor and trainee burden. However, in doing so, we recognize limitations in our capacity to optimally monitor fidelity, and acknowledge that fidelity is not monitored to the same degree in our program compared to standard clinical trials (eg, [90]).

Third, to maximize scalability of the program, coaching is provided virtually using videoconferencing. Prior research has raised the possibility that compared with self-administered or fully automatized options, digital mental health interventions may be most effective for adolescents and young adults when incorporating in-person elements [91]. However, the extent to which virtual interactions with a human coach may provide a similar degree of benefit is unknown. Additional research may clarify the effectiveness of fully remote coaching and guide potential adaptations to this program.

Last, our program was initially designed for use in a specific setting (ie, a peer-to-peer program supporting college students). Additional efforts and reliance on existing implementation science and human-centered design frameworks, such as the CFIR, are needed to determine how this program and similar ones may be adapted and augmented for use in other types of settings and with new populations. A number of conceptual frameworks to adapt interventions in new contexts have been proposed, and these can be used to guide adaptation of paraprofessional support programs for new settings (eg, [92]).

Conclusion and Future Directions

Finally, we consider future directions for this work, falling within the scope of the paraprofessional field at large. First, to meet rising rates of mental illness worldwide, expansion of paraprofessional mental health programs into new settings is critically needed. Second, funding for these programs must also encompass sufficient resources to support quality assurance in training, supervision, and treatment delivery [93], as has been the case throughout the development of the coach training program presented here. However, fidelity assurance strategies must be integrated with careful awareness of their scalability, enabling paraprofessional programs to continue expanding in reach. Third, adaptations should be designed in collaboration with community stakeholders to reduce drift from EBT protocols, while also addressing the implementation factors that drive adaptation needs [92]. Lastly, research protocols (eg, [94]) should be developed to enable empirical testing of our model, along with potential model adaptations to determine effectiveness and inform modifications to future iterations of the coach training program.

Acknowledgments

BMR and TK were responsible for conceptualization and writing of this paper. ZDC developed the digital intervention used by this program and provided crucial edits to the paper. EGG created the training program described in this paper, conducted supervision-of-supervision, and provided crucial edits to the paper. MGC oversaw the creation and implementation of this program and provided crucial edits to the paper.

This work would not have been possible without the immense contributions of the following individuals, who were central to the development, implementation, and supervision of the coaching program described in this paper: Amanda Loerinc, PhD; Allyson Pimentel, EdD; Bita Mesri, PhD; Blanche Wright, MA, CPhil; Brittany Drake, MA, CPhil; Dana Saifan, MA, CPhil; Jennifer Gamarra, PhD; Julia Hammett, PhD; Julia Yarrington, MA; Meghan Vinograd, PhD; Meredith Boyd, MA, CPhil; Sophie Arkin, MA, CPhil; and Stassja Sichko, MA.

Conflicts of Interest

ZDC received consultancy fees from Joyable for his work on cognitive-behavioral therapy during 2016-2017.

Multimedia Appendix 1
Rating form to evaluate interpersonal process skills.
[PDF File (Adobe PDF File), 101 KB - mental_v9i1e32430_app1.pdf]

References

1. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016 Feb;3(2):171-178 [FREE Full text] [doi: 10.1016/s2215-0366(15)00505-2]

2. Liu Q, He H, Yang J, Feng X, Zhao F, Lyu J. Changes in the global burden of depression from 1990 to 2017: findings from the Global Burden of Disease study. J Psychiatr Res 2020 Jul;126:134-140 [FREE Full text] [doi: 10.1016/j.jpsychires.2019.08.002] [Medline: 31439359]
3. Betancourt T, Chambers DA. Optimizing an era of global mental health implementation science. JAMA Psychiatry 2016 Feb;73(2):99-100 [FREE Full text] [doi: 10.1001/jamapsychiatry.2015.2705] [Medline: 26720304]
4. Butryn T, Bryant L, Marchionni C, Sholevar F. The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. Int J Acad Med 2017;3(1):5. [doi: 10.4103/IJAM.IJAM_49_17]
5. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med 2011 Dec;104(12):510-520 [FREE Full text] [doi: 10.1258/jrsm.2011.110180] [Medline: 22179294]
6. Kazdin AE. Annual research review: expanding mental health services through novel models of intervention delivery. J Child Psychol Psychiatry 2019 Apr;60(4):455-472 [FREE Full text] [doi: 10.1111/jcpp.12937] [Medline: 29900543]
7. Olfson M. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. Health Aff (Millwood) 2016 Jun 01;35(6):983-990 [FREE Full text] [doi: 10.1377/hlthaff.2015.1619] [Medline: 27269013]
8. Barnett ML, Gonzalez A, Miranda J, Chavira DA, Lau AS. Mobilizing community health workers to address mental health disparities for underserved populations: a systematic review. Adm Policy Ment Health 2018 Mar;45(2):195-211 [FREE Full text] [doi: 10.1007/s10488-017-0815-0] [Medline: 28730278]
9. Singla DR, Kohrt BA, Murray LK, Anand A, Chorpita BF, Patel V. Psychological treatments for the world: lessons from low- and middle-income countries. Annu Rev Clin Psychol 2017 May 08;13:149-181 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-045217] [Medline: 28482687]
10. Lewin S, Dick J, Pond P, Zwarenstein M, Aja GN, van Wyk BE, et al. Lay health workers in primary and community health care. Cochrane Database Syst Rev 2005 Jan 25(1):CD004015. [doi: 10.1002/14651858.CD004015.pub2] [Medline: 15674924]
11. Chinman M, McInnes DK, Eisen S, Ellison M, Farkas M, Armstrong M, et al. Establishing a research agenda for understanding the role and impact of mental health peer specialists. Psychiatr Serv 2017 Sep 01;68(9):955-957 [FREE Full text] [doi: 10.1176/appi.ps.201700054] [Medline: 28617205]
12. Rosenthal EL, Brownstein JN, Rush CH, Hirsch GR, Willaert AM, Scott JR, et al. Community health workers: part of the solution. Health Aff (Millwood) 2010 Jul;29(7):1338-1342 [FREE Full text] [doi: 10.1377/hlthaff.2010.0081] [Medline: 20606185]
13. Naslund JA, Aschbrenner KA, Araya R, Marsch LA, Unützer J, Patel V, et al. Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature. Lancet Psychiatry 2017 Jun;4(6):486-500 [FREE Full text] [doi: 10.1016/s2215-0366(17)30096-2]
14. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255 [FREE Full text] [doi: 10.1007/s40501-019-00181-z]
15. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
16. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, REEACT Team. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ 2015 Nov 11;351:h5627 [FREE Full text] [doi: 10.1136/bmj.h5627] [Medline: 26559241]
17. Benton SA, Heesacker M, Snowden SJ, Lee G. Therapist-assisted, online (TAO) intervention for anxiety in college students: TAO outperformed treatment as usual. Prof Psychol: Res Pract 2016 Oct;47(5):363-371 [FREE Full text] [doi: 10.1037/pro0000097]
18. Schueller SM, Tomasino KN, Mohr DC. Integrating human support into behavioral intervention technologies: the efficiency model of support. Clin Psychol: Sci Pract 2017 Mar;24(1):27-45 [FREE Full text] [doi: 10.1037/h0101740]
19. Conley CS, Durlak JA, Shapiro JB, Kirsch AC, Zahniser E. A meta-analysis of the impact of universal and indicated preventive technology-delivered interventions for higher education students. Prev Sci 2016 Aug;17(6):659-678 [FREE Full text] [doi: 10.1007/s11121-016-0662-3] [Medline: 27225631]
20. Cross WF, West JC. Examining implementer fidelity: conceptualising and measuring adherence and competence. J Child Serv 2011 Mar 18;6(1):18-33 [FREE Full text] [doi: 10.5042/jcs.2011.0123] [Medline: 21922026]
21. Schoenwald SK, Sheidow AJ, Letourneau EJ. Toward effective quality assurance in evidence-based practice: links between expert consultation, therapist fidelity, and child outcomes. J Clin Child Adolesc Psychol 2004 Feb;33(1):94-104 [FREE Full text] [doi: 10.1207/s15374424jccp3301_10]
22. Cohen ZD, Craske MG. The development and pilot implementation of a modular, transdiagnostic, personalized digital therapy during a global pandemic. 2021 Presented at: European Association of Behavioral and Cognitive Therapies; 2021; Belfast, Northern Ireland.
23. Duffy ME, Twenge JM, Joiner TE. Trends in mood and anxiety symptoms and suicide-related outcomes among U.S. undergraduates, 2007-2018: evidence from two national surveys. J Adolesc Health 2019 Nov;65(5):590-598 [FREE Full text] [doi: 10.1016/j.jadohealth.2019.04.033] [Medline: 31279724]

24. Padmanathan P, De Silva MJ. The acceptability and feasibility of task-sharing for mental healthcare in low and middle income countries: a systematic review. Soc Sci Med 2013 Nov;97:82-86 [FREE Full text] [doi: 10.1016/j.socscimed.2013.08.004] [Medline: 24161092]
25. Bellerose M, Awoonor-Williams K, Alva S, Magalona S, Sacks E. 'Let me move to another level': career advancement desires and opportunities for community health nurses in Ghana. Glob Health Promot 2021 Jul 16:17579759211027426 [FREE Full text] [doi: 10.1177/17579759211027426] [Medline: 34269105]
26. Schoenwald SK, Garland AF, Chapman JE, Frazier SL, Sheidow AJ, Southam-Gerow MA. Toward the effective and efficient measurement of implementation fidelity. Adm Policy Ment Health 2011 Jan;38(1):32-43 [FREE Full text] [doi: 10.1007/s10488-010-0321-0] [Medline: 20957425]
27. Brown LA, Craske MG, Glenn DE, Stein MB, Sullivan G, Sherbourne C, et al. CBT competence in novice therapists improves anxiety outcomes. Depress Anxiety 2013 Feb;30(2):97-115 [FREE Full text] [doi: 10.1002/da.22027] [Medline: 23225338]
28. Murray LK, Dorsey S, Bolton P, Jordans MJ, Rahman A, Bass J, et al. Building capacity in mental health interventions in low resource countries: an apprenticeship model for training local providers. Int J Ment Health Syst 2011 Nov 18;5(1):30-12 [FREE Full text] [doi: 10.1186/1752-4458-5-30] [Medline: 22099582]
29. van Ginneken N, Tharyan P, Lewin S, Rao GN, Meera SM, Pian J, et al. Non-specialist health worker interventions for the care of mental, neurological and substance-abuse disorders in low- and middle-income countries. Cochrane Database Syst Rev 2013 Nov 19(11):CD009149. [doi: 10.1002/14651858.CD009149.pub2] [Medline: 24249541]
30. Kemp CG, Petersen I, Bhana A, Rao D. Supervision of task-shared mental health care in low-resource settings: a commentary on programmatic experience. Glob Health Sci Pract 2019 Jun 27;7(2):150-159 [FREE Full text] [doi: 10.9745/ghsp-d-18-00337]
31. Montgomery EC, Kunik ME, Wilson N, Stanley MA, Weiss B. Can paraprofessionals deliver cognitive-behavioral therapy to treat anxiety and depressive symptoms? Bull Menninger Clin 2010;74(1):45-62 [FREE Full text] [doi: 10.1521/bumc.2010.74.1.45] [Medline: 20235623]
32. Diebold A, Ciolino JD, Johnson JK, Yeh C, Gollan JK, Tandon SD. Comparing fidelity outcomes of paraprofessional and professional delivery of a perinatal depression preventive intervention. Adm Policy Ment Health 2020 Jul;47(4):597-605 [FREE Full text] [doi: 10.1007/s10488-020-01022-5] [Medline: 32086657]
33. Shahmalak U, Blakemore A, Waheed MW, Waheed W. The experiences of lay health workers trained in task-shifting psychological interventions: a qualitative systematic review. Int J Ment Health Syst 2019;13:64-15 [FREE Full text] [doi: 10.1186/s13033-019-0320-9] [Medline: 31636699]
34. Waller G, Turner H. Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. Behav Res Ther 2016 Feb;77:129-137 [FREE Full text] [doi: 10.1016/j.brat.2015.12.005] [Medline: 26752326]
35. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
36. Scholten H, Granic I. Use of the principles of design thinking to address limitations of digital mental health interventions for youth: viewpoint. J Med Internet Res 2019 Jan 14;21(1):e11528 [FREE Full text] [doi: 10.2196/11528] [Medline: 31344671]
37. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35(4):332-338 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.03.008] [Medline: 23664503]
38. Lattie EG, Graham AK, Hadjistavropoulos HD, Dear BF, Titov N, Mohr DC. Guidance on defining the scope and development of text-based coaching protocols for digital mental health interventions. Digit Health 2019;5:2055207619896145 [FREE Full text] [doi: 10.1177/2055207619896145] [Medline: 31897306]
39. Mohr D, Duffecy J, Ho J, Kwasny M, Cai X, Burns MN, et al. A randomized controlled trial evaluating a manualized TeleCoaching protocol for improving adherence to a web-based intervention for the treatment of depression. PLoS One 2013;8(8):e70086 [FREE Full text] [doi: 10.1371/journal.pone.0070086] [Medline: 23990896]
40. Myrick K, Del Vecchio P. Peer support services in the behavioral healthcare workforce: state of the field. Psychiatr Rehabil J 2016 Sep;39(3):197-203 [FREE Full text] [doi: 10.1037/prj0000188] [Medline: 27183186]
41. Gagne CA, Finch WL, Myrick KJ, Davis LM. Peer workers in the behavioral and integrated health workforce: opportunities and future directions. Am J Prev Med 2018 Jun;54(6 Suppl 3):S258-S266 [FREE Full text] [doi: 10.1016/j.amepre.2018.03.010] [Medline: 29779550]
42. Bassuk EL, Hanson J, Greene RN, Richard M, Laudet A. Peer-delivered recovery support services for addictions in the United States: a systematic review. J Subst Abuse Treat 2016 Apr;63:1-9 [FREE Full text] [doi: 10.1016/j.jsat.2016.01.003] [Medline: 26882891]
43. Watson E. The mechanisms underpinning peer support: a literature review. J Ment Health 2019 Dec;28(6):677-688 [FREE Full text] [doi: 10.1080/09638237.2017.1417559] [Medline: 29260930]

44. Gillard S, Foster R, Gibson S, Goldsmith L, Marks J, White S. Describing a principles-based approach to developing and evaluating peer worker roles as peer support moves into mainstream mental health services. MHSI 2017 Jun 12;21(3):133-143 [FREE Full text] [doi: 10.1108/mhsi-03-2017-0016]
45. Basset T, Faulkner A, Repper J, Stamou E. Lived Experience Leading the Way: Peer Support in Mental Health. London, UK: Together for Mental Wellbeing; 2010.
46. Silver J, Nemec PB. The role of the peer specialists: unanswered questions. Psychiatr Rehabil J 2016 Sep;39(3):289-291 [FREE Full text] [doi: 10.1037/prj0000216] [Medline: 27618464]
47. Lloyd-Evans B, Mayo-Wilson E, Harrison B, Istead H, Brown E, Pilling S, et al. A systematic review and meta-analysis of randomised controlled trials of peer support for people with severe mental illness. BMC Psychiatry 2014 Feb 14;14(1):1-12 [FREE Full text] [doi: 10.1186/1471-244x-14-39]
48. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020 Apr 03;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
49. Ali K, Farrer L, Gulliver A, Griffiths KM. Online peer-to-peer support for young people with mental health problems: a systematic review. JMIR Ment Health 2015;2(2):e19 [FREE Full text] [doi: 10.2196/mental.4418] [Medline: 26543923]
50. van der Zanden R, Kramer J, Gerrits R, Cuijpers P. Effectiveness of an online group course for depression in adolescents and young adults: a randomized trial. J Med Internet Res 2012 Jun 07;14(3):e86 [FREE Full text] [doi: 10.2196/jmir.2033] [Medline: 22677437]
51. Day V, McGrath PJ, Wojtowicz M. Internet-based guided self-help for university students with anxiety, depression and stress: a randomized controlled clinical trial. Behav Res Ther 2013 Jul;51(7):344-351 [FREE Full text] [doi: 10.1016/j.brat.2013.03.003] [Medline: 23639300]
52. Klatt C, Berg CJ, Thomas JL, Ehlinger E, Ahluwalia JS, An LC. The role of peer e-mail support as part of a college smoking-cessation website. Am J Prev Med 2008 Dec;35(6 Suppl):S471-S478 [FREE Full text] [doi: 10.1016/j.amepre.2008.09.001] [Medline: 19012841]
53. Conley C, Hundert CG, Charles JL, Huguenel BM, Al-khouja M, Qin S, et al. Honest, open, proud–college: effectiveness of a peer-led small-group intervention for reducing the stigma of mental illness. Stigma Health 2020 May;5(2):168-178 [FREE Full text] [doi: 10.1037/sah0000185]
54. Freeman E, Barker C, Pistrang N. Outcome of an online mutual support group for college students with psychological problems. Cyberpsychol Behav 2008 Oct;11(5):591-593 [FREE Full text] [doi: 10.1089/cpb.2007.0133] [Medline: 18817485]
55. Horgan A, McCarthy G, Sweeney J. An evaluation of an online peer support forum for university students with depressive symptoms. Arch Psychiatr Nurs 2013 Apr;27(2):84-89 [FREE Full text] [doi: 10.1016/j.apnu.2012.12.005] [Medline: 23540518]
56. Eddie D, Hoffman L, Vilsaint C, Abry A, Bergman B, Hoeppner B, et al. Lived experience in new models of care for substance use disorder: a systematic review of peer recovery support services and recovery coaching. Front Psychol 2019;10:1052 [FREE Full text] [doi: 10.3389/fpsyg.2019.01052] [Medline: 31263434]
57. Craske MG, Rose RD, Lang A, Welch SS, Campbell-Sills L, Sullivan G, et al. Computer-assisted delivery of cognitive behavioral therapy for anxiety disorders in primary-care settings. Depress Anxiety 2009;26(3):235-242 [FREE Full text] [doi: 10.1002/da.20542] [Medline: 19212970]
58. Craske MG, Stein MB, Sullivan G, Sherbourne C, Bystritsky A, Rose RD, et al. Disorder-specific impact of coordinated anxiety learning and management treatment for anxiety disorders in primary care. Arch Gen Psychiatry 2011 Apr;68(4):378-388 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2011.25] [Medline: 21464362]
59. Craske MG, Meuret AE, Ritz T, Treanor M, Dour HJ. Treatment for anhedonia: a neuroscience driven approach. Depress Anxiety 2016 Oct;33(10):927-938 [FREE Full text] [doi: 10.1002/da.22490] [Medline: 27699943]
60. Craske MG, Meuret AE, Ritz T, Treanor M, Dour HJ, Rosenfield D. Positive affect treatment for depression and anxiety: a randomized clinical trial for a core feature of anhedonia. J Consult Clin Psychol 2019 May;87(5):457-471 [FREE Full text] [doi: 10.1037/ccp0000396] [Medline: 30998048]
61. Roy-Byrne P, Craske MG, Sullivan G, Rose RD, Edlund MJ, Lang AJ, et al. Delivery of evidence-based treatment for multiple anxiety disorders in primary care: a randomized controlled trial. JAMA 2010 May 19;303(19):1921-1928 [FREE Full text] [doi: 10.1001/jama.2010.608] [Medline: 20483968]
62. Watkins ER, Mullan E, Wingrove J, Rimes K, Steiner H, Bathurst N, et al. Rumination-focused cognitive-behavioural therapy for residual depression: phase II randomised controlled trial. Br J Psychiatry 2011 Oct;199(4):317-322 [FREE Full text] [doi: 10.1192/bjp.bp.110.090282] [Medline: 21778171]
63. Watkins E, Newbold A, Tester-Jones M, Javaid M, Cadman J, Collins LM, et al. Implementing multifactorial psychotherapy research in online virtual environments (IMPROVE-2): study protocol for a phase III trial of the MOST randomized component selection method for internet cognitive-behavioural therapy for depression. BMC Psychiatry 2016 Oct 06;16(1):345 [FREE Full text] [doi: 10.1186/s12888-016-1054-8] [Medline: 27716200]
64. Harvey AG. A transdiagnostic intervention for youth sleep and circadian problems. Cogn Behav Pract 2016 Aug;23(3):341-355 [FREE Full text] [doi: 10.1016/j.cbpra.2015.06.001]

65. Harvey AG, Hein K, Dolsen MR, Dong L, Rabe-Hesketh S, Gumport NB, et al. Modifying the impact of eveningness chronotype ("night-owls") in youth: a randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2018 Oct;57(10):742-754 [FREE Full text] [doi: 10.1016/j.jaac.2018.04.020] [Medline: 30274649]
66. Harvey AG, Dong L, Hein K, Yu SH, Martinez AJ, Gumport NB, et al. A randomized controlled trial of the Transdiagnostic Intervention for Sleep and Circadian Dysfunction (TranS-C) to improve serious mental illness outcomes in a community setting. J Consult Clin Psychol 2021 Jun;89(6):537-550 [FREE Full text] [doi: 10.1037/ccp0000650] [Medline: 34264701]
67. Hettema J, Steele J, Miller WR. Motivational interviewing. Annu Rev Clin Psychol 2005;1:91-111 [FREE Full text] [doi: 10.1146/annurev.clinpsy.1.102803.143833] [Medline: 17716083]
68. Rollnick S, Miller WR. What is motivational interviewing? Behav Cogn Psychother 2009 Jun 16;23(4):325-334 [FREE Full text] [doi: 10.1017/s135246580001643x]
69. Robertson K. Active listening: more than just paying attention. Aust Fam Physician 2005 Dec;34(12):1053-1055 [FREE Full text] [Medline: 16333490]
70. Bearman SK, Schneiderman RL, Zoloth E. Building an evidence base for effective supervision practices: an analogue experiment of supervision to increase EBT fidelity. Adm Policy Ment Health 2017 Mar;44(2):293-307 [FREE Full text] [doi: 10.1007/s10488-016-0723-8] [Medline: 26867545]
71. Bjork RA. Memory and metamemory considerations in the training of human beings. In: Metacognition: Knowing about Knowing. Cambridge, MA: MIT Press; 1994.
72. Bjork EL, Bjork RA. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In: Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society. New York, NY: Worth Publishers; 2011:56-64.
73. Schmidt RA, Bjork RA. New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. Psychol Sci 2017 Apr 25;3(4):207-218 [FREE Full text] [doi: 10.1111/j.1467-9280.1992.tb00029.x]
74. Soderstrom NC, Bjork RA. Learning versus performance: an integrative review. Perspect Psychol Sci 2015 Mar;10(2):176-199 [FREE Full text] [doi: 10.1177/1745691615569000] [Medline: 25910388]
75. Smith SM. A comparison of two techniques for reducing context-dependent forgetting. Mem Cognit 1984 Sep;12(5):477-482 [FREE Full text] [doi: 10.3758/bf03198309] [Medline: 6521649]
76. Rohrer D, Dedrick RF, Burgess K. The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. Psychon Bull Rev 2014 Oct;21(5):1323-1330 [FREE Full text] [doi: 10.3758/s13423-014-0588-3] [Medline: 24578089]
77. Kornell N, Hays MJ, Bjork RA. Unsuccessful retrieval attempts enhance subsequent learning. J Exp Psychol Learn Mem Cogn 2009 Jul;35(4):989-998 [FREE Full text] [doi: 10.1037/a0015729] [Medline: 19586265]
78. Bennett-Levy J, McManus F, Westling BE, Fennell M. Acquiring and refining CBT skills and competencies: which training methods are perceived to be most effective? Behav Cogn Psychother 2009 Aug 25;37(5):571-583 [FREE Full text] [doi: 10.1017/s1352465809990270]
79. Kolb DA. Experience as the Source of Learning and Development. Hoboken, NJ: Prentice Hall; 1984.
80. Milne D, Aylott H, Fitzpatrick H, Ellis MV. How does clinical supervision work? Using a “best evidence synthesis” approach to construct a basic model of supervision. WCSU 2008 Nov 21;27(2):170-190 [FREE Full text] [doi: 10.1080/07325220802487915]
81. Falender CA. Clinical supervision in a competency-based era. S Afr J Psychol 2014 Jan 07;44(1):6-17 [FREE Full text] [doi: 10.1177/0081246313516260]
82. Bennett-Levy J. Therapist skills: a cognitive model of their acquisition and refinement. Behav Cogn Psychother 2005 Oct 20;34(1):57-78 [FREE Full text] [doi: 10.1017/s1352465805002420]
83. Enock PM, McNally RJ. How mobile apps and other web-based interventions can transform psychological treatment and the treatment development cycle. Behav Ther 2013;36(3):56-66.
84. Kendall PC, Beidas RS. Smoothing the trail for dissemination of evidence-based practices for youth: flexibility within fidelity. Prof Psychol: Res Pract 2007;38(1):13-20 [FREE Full text] [doi: 10.1037/0735-7028.38.1.13]
85. Kendall PC, Frank HE. Implementing evidence-based treatment protocols: flexibility within fidelity. Clin Psychol (New York) 2018 Dec;25(4):e12271 [FREE Full text] [doi: 10.1111/cpsp.12271] [Medline: 30643355]
86. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4(1):50-15 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
87. Kirk MA, Kelley C, Yankey N, Birken SA, Abadie B, Damschroder L. A systematic review of the use of the Consolidated Framework for Implementation Research. Implement Sci 2016 May 17;11(1):72-13 [FREE Full text] [doi: 10.1186/s13012-016-0437-z] [Medline: 27189233]
88. Wiltsey Stirman S, Baumann AA, Miller CJ. The FRAME: an expanded framework for reporting adaptations and modifications to evidence-based interventions. Implement Sci 2019 Jun 06;14(1):58-10 [FREE Full text] [doi: 10.1186/s13012-019-0898-y] [Medline: 31171014]

89. O'Dea B, King C, Subotic-Kerry M, Achilles MR, Cockayne N, Christensen H. Smooth sailing: a pilot study of an online, school-based, mental health service for depression and anxiety. Front Psychiatry 2019;10:574 [FREE Full text] [doi: 10.3389/fpsyt.2019.00574] [Medline: 31481904]
90. Wiltsey Stirman S, Gutner CA, Crits-Christoph P, Edmunds J, Evans AC, Beidas RS. Relationships between clinician-level attributes and fidelity-consistent and fidelity-inconsistent modifications to an evidence-based psychotherapy. Implement Sci 2015 Aug 13;10(1):115-110 [FREE Full text] [doi: 10.1186/s13012-015-0308-z] [Medline: 26268633]
91. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847 [FREE Full text] [doi: 10.2196/25847] [Medline: 33913817]
92. Allen JD, Linnan LA, Emmons KM, Brownson R, Colditz G, Proctor E. Fidelity and its relationship to implementation effectiveness, adaptation, and dissemination. In: Dissemination and Implementation Research in Health: Translating Science to Practice. Oxford, UK: Oxford University Press; 2012:281-304.
93. Borrelli B. The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. J Public Health Dent 2011;71(s1):S52-S63 [FREE Full text] [doi: 10.1111/j.1752-7325.2011.00233.x] [Medline: 21499543]
94. Dohnt HC, Dowling MJ, Davenport TA, Lee G, Cross SP, Scott EM, et al. Supporting clinicians to use technology to deliver highly personalized and measurement-based mental health care to young people: protocol for an evaluation study. JMIR Res Protoc 2021 Jun 14;10(6):e24697 [FREE Full text] [doi: 10.2196/24697] [Medline: 34125074]

Abbreviations

CBT: cognitive-behavioral therapy
CFIR: Consolidated Framework for Implementation Research
DGC: Depression Grand Challenge
EBT: evidence-based treatment
STAND: Screening and Treatment for Anxiety and Depression
UCLA: University of California, Los Angeles

Edited by J Torous; submitted 02.08.21; peer-reviewed by D Frank, R Pine, L Balcombe; comments to author 27.09.21; revised version received 21.11.21; accepted 22.11.21; published 26.01.22.

Please cite as:
Rosenberg BM, Kodish T, Cohen ZD, Gong-Guy E, Craske MG
A Novel Peer-to-Peer Coaching Program to Support Digital Mental Health: Design and Implementation
JMIR Ment Health 2022;9(1):e32430
URL: https://mental.jmir.org/2022/1/e32430
doi:10.2196/32430
PMID:35080504

Original Paper

A New Digital Assessment of Mental Health and Well-being in the Workplace: Development and Validation of the Unmind Index

Anika Sierk[1*], BSc, MSc, PhD; Eoin Travers[1*], BSc, PhD; Marcos Economides[1], BSc, PhD; Bao Sheng Loe[2], MA, PhD; Luning Sun[2], BSc, MSc, PhD; Heather Bolton[1], BSc, DClinPsy

[1]Unmind Ltd, London, United Kingdom

[2]The Psychometrics Centre, Judge Business School, University of Cambridge, Cambridge, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Eoin Travers, BSc, PhD
Unmind Ltd
180 Borough High Street
London, SE1 1LB
United Kingdom
Email: eoin.travers@unmind.com

Abstract

Background: Unmind is a workplace, digital, mental health platform with tools to help users track, maintain, and improve their mental health and well-being (MHWB). Psychological measurement plays a key role on this platform, providing users with insights on their current MHWB, the ability to track it over time, and personalized recommendations, while providing employers with aggregate information about the MHWB of their workforce.

Objective: Due to the limitations of existing measures for this purpose, we aimed to develop and validate a novel well-being index for digital use, to capture symptoms of common mental health problems and key aspects of positive well-being.

Methods: In Study 1A, questionnaire items were generated by clinicians and screened for face validity. In Study 1B, these items were presented to a large sample (n=1104) of UK adults, and exploratory factor analysis was used to reduce the item pool and identify coherent subscales. In Study 2, the final measure was presented to a new nationally representative UK sample (n=976), along with a battery of existing measures, with 238 participants retaking the Umind Index after 1 week. The factor structure and measurement invariance of the Unmind Index was evaluated using confirmatory factor analysis, convergent and discriminant validity by estimating correlations with existing measures, and reliability by examining internal consistency and test-retest intraclass correlations.

Results: Studies 1A and 1B yielded a 26-item measure with 7 subscales: *Calmness*, *Connection*, *Coping*, *Happiness*, *Health*, *Fulfilment*, and *Sleep*. Study 2 showed that the Unmind Index is fitted well by a second-order factor structure, where the 7 subscales all load onto an overall MHWB factor, and established measurement invariance by age and gender. Subscale and total scores correlate well with existing mental health measures and generally diverge from personality measures. Reliability was good or excellent across all subscales.

Conclusions: The Unmind Index is a robust measure of MHWB that can help to identify target areas for intervention in nonclinical users of a mental health app. We argue that there is value in measuring mental ill health and mental well-being together, rather than treating them as separate constructs.

(JMIR Ment Health 2022;9(1):e34103) doi:10.2196/34103

KEYWORDS

mental health; well-being; mHealth; measurement

Introduction

Background

Poor mental health affects hundreds of millions of people worldwide, impacting individual quality of life and creating a significant economic burden for employers [1-3]. With evidence that many mental health problems are preventable or treatable [4-6], there is a strong business case for employers to invest in preventative mental health solutions for their workforces [7,8]. In recent years, desktop and mobile health (mHealth) apps have begun to fulfill this preventative remit. Digital technologies might be particularly useful in a workplace setting, where traditional reactive approaches tend to have low uptake [9].

Unmind is a workplace, digital, mental health platform providing employees with tools to help them track, maintain, and improve their mental health and well-being (MHWB) and allowing employers to gain insight into the overall well-being of their employees through anonymized, aggregated data. Consistent with the contemporary understanding of mental health as a complete state of physical, mental, and social well-being [10], the Unmind approach encourages users to take a holistic approach to understanding and managing their MHWB. This holistic approach may be particularly relevant for promoting regular, proactive use of the platform in working adults.

Measurement plays a key role on the Unmind platform. First, given the broad range of content available on the platform, it is important to guide users toward the materials best suited to their particular needs. Second, allowing users to monitor and reflect on their own mental health has been shown to improve engagement with mHealth apps [11,12]. Finally, there is some evidence that measurement tools may directly improve users' mental health, perhaps by encouraging them to reflect upon their own mental states [13,14]. The Insights section of the Unmind platform consists of 2 tools: a brief Check-In (mood tracker) and the more in-depth Unmind Index. In this article, we describe the development and validation of the Unmind Index.

The Case for a Novel Measure

There is a distinction between mental health (the absence of mental illness) and mental well-being. Existing self-report scales are typically intended to measure one or the other factor. On the one hand, diagnostic mental health measures are used in clinical practice to help diagnose patients with specific mental health disorders (as described in the Diagnostic and Statistical Manual of Mental Disorders [DSM]-V or International Classification of Diseases [ICD]-11). On the other hand, positive mental well-being scales are intended to measure broader well-being and quality of life and are typically based on principles from positive psychology. Although distinct, these 2 factors are strongly correlated [15]. Ideally, the self-monitoring features of an mHealth app should capture both factors.

As they are, existing diagnostic and positive mental well-being scales have strengths and weaknesses for use in mHealth apps. Diagnostic scales provide sensitive, well-validated measures of specific aspects of mental ill-health, such as the Patient Health Questionnaire 9 (PHQ-9; depression) [16], General Anxiety Disorder 7 (GAD-7; anxiety disorders) [17], or the Insomnia Severity Index (ISI) [18]. However, these scales are a poor fit for a digital mental health platform for 2 reasons.

First, by design, these scales focus on disorder-specific symptoms. For example, the GAD-7 will assess the extent to which anxiety impairs an individual's day-to-day life but will not directly assess their ability to relax or remain calm under usual circumstances. As a result, these scales typically have excellent sensitivity for users with poor mental health but inadequate sensitivity for healthier users who would not be seen in a clinical setting. This is also reflected in the language typically used in diagnostic tests, which is necessarily problem-focused. Presenting users with a large number of negatively phrased questions is likely to discourage user engagement in a digital mental health platform, and these questions may feel less relevant to healthier users.

Second, it is widely recognized that many mental health disorders are strongly interrelated, with largely overlapping symptoms. It has been shown that much of the variance across a broad range of mental health scales is explained by a single latent factor capturing participants' overall state of mental health or well-being [19]. Individual diagnostic scales are not designed to measure this higher-order MHWB factor, and although it could be approximated by averaging scores across diagnostic scales for different disorders, this approach has not been validated.

Holistic scales intended to assess overall mental well-being address both of these limitations. These scales are typically designed using positive psychology principles, use positive language, are calibrated to measure the range of mental health seen in the general population, and capture a broader range of mental health–related constructs than diagnostic tests can. Holistic scales include the Warwick-Edinburgh Mental Wellbeing Scales (WEMWBS) [20] and the Brief Inventory of Thriving (BIT) [21]. However, these scales do not reliably measure the various components of mental health, such as happiness, social support, or sleep quality, and so are of limited use for guiding users to appropriate content or for self-reflection.

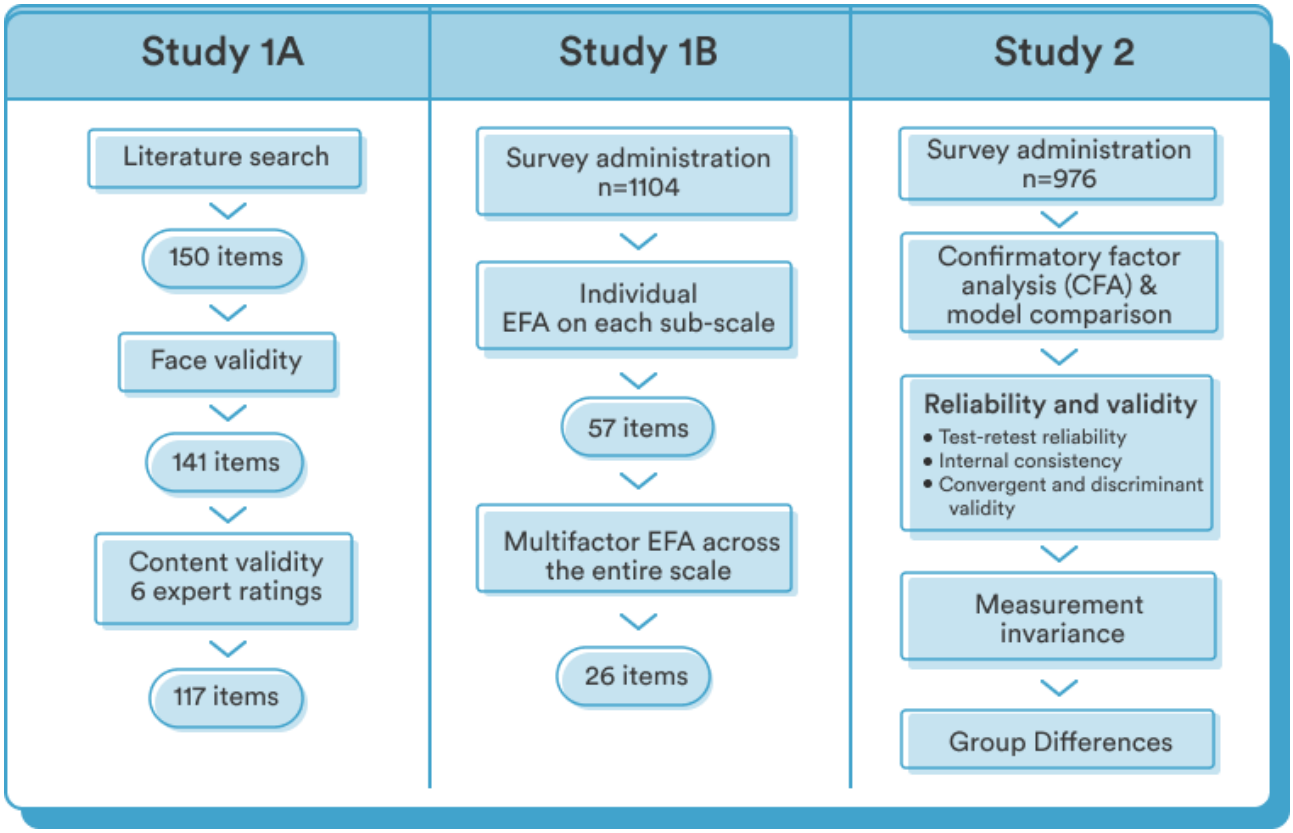
Goals for the Unmind Index

Given the limitations of existing measures for our purposes, we decided to develop a new measure for use on the Unmind platform. Five primary goals guided the development of this measure. First, we decided to combine items that measure mental health and those that measure well-being. That is, we aimed to measure MHWB as a combined construct. Second, the Unmind Index was intended to measure the different subdomains of MHWB (eg, social functioning, mood, anxiety), providing users with personalized feedback and actionable content recommendations. Third, it was also intended to provide a single overall MHWB score, combining scores from the individual subdomains in a scientifically validated way. Fourth, the Unmind Index was intended to empower users to monitor their mental health over time, spotting trends. Finally, as a workplace platform, the Unmind Index was intended to allow employers to access their employees' aggregated data to understand trends and inform their well-being strategy. Beyond these goals, we sought to create a measure that was brief enough to encourage

regular completion by casual users of the Unmind platform, easy to complete with minimal instruction, and targeted to nonclinical (workplace) populations.

This paper reports the development and validation of the Unmind Index in 3 parts. Study 1A described the generation of candidate items and the assessment of their validity. Study 1B documented the item selection process and the identification of the various facets of MHWB to be captured by the Unmind Index, using exploratory factor analysis (EFA). Finally, Study 2 described the validation of the Unmind Index, including confirmatory factor analysis (CFA) to identify the appropriate approach to calculating the overall MHWB score. It also demonstrated the psychometric properties of the Unmind Index and its convergent validity with existing diagnostic and holistic measures. It also established discriminant validity against measures of personality, documented measurement invariance, and explored gender and age differences in scores (see Figure 1 for an overview).

Figure 1. Overview of the structure of Studies 1A (scale development), 1B (exploratory factor analysis), and 2 (validation). EFA: exploratory factor analysis.



Ethics

The study received ethical approval from the University of Cambridge (Judge Business School Departmental Ethics Review Group, approval number 20-061). All participants provided informed consent prior to taking part.

Study 1A: Scale Development

Item Generation and Face Validity

An initial pool of 150 items was created by an experienced UK-trained clinical psychologist (HB) for the proposed 7 constructs underpinning our conceptualization of MHWB. The constructs were named *Happiness* (37 items), *Calmness* (20 items), *Coping* (15 items), *Health* (10 items), *Sleep* (8 items), *Energy* (7 items), and *Vitality* (44 items). All items were presented to 4 nontechnical members of staff at Unmind who were asked to assess each item for face validity [22] by providing qualitative feedback on the semantic clarity of each item. Based on this feedback, 5 items were reworded, and 9 items were discarded. The remaining pool of 141 items was reviewed and edited by a professional copywriter to improve readability and tone of voice.

Content Validity

A panel of 6 UK-trained clinical psychologists (4 female, 2 male), with a mean 14.3 (range 12-20) years of experience in adult mental health, were individually asked to rate each of the remaining items with respect to how well it assessed the defined construct it purported to measure (1=not relevant, 2=somewhat relevant, 3=quite relevant, 4=highly relevant). They also provided further qualitative feedback on content validity and suggestions for item rewording where applicable. Interrater reliability was assessed via the item content validity index (I-CVI), and items with an I-CVI <.8 were removed—a benchmark considered to present an excellent strength of agreement between raters [23]. Based on the experts' suggestions regarding item wording, we added in 9 slightly reworded items in addition to their original equivalent. The resulting final pool of 117 candidate items was then explored in an EFA study, described next.

Study 1B: Exploratory Factor Analysis

Methods

Participants

We recruited a convenience sample of UK-based adults (n=1180). The sample size was determined based on a commonly accepted item-to-variable ratio of 1:10 [24,25], with 117 items. Individuals were recruited via the online recruitment platform Prolific [26] and invited to participate in an online survey built using the Gorilla Experiment Builder [27]. Prolific has been empirically tested across key attributes such as participant response rates and data quality [28]. Upon joining the Prolific participant pool, individuals are required to complete an extensive prescreening questionnaire designed to help researchers automatically screen for eligibility criteria at the recruitment stage. Participants were eligible for the study if they were aged 18-65 years, based in the United Kingdom, proficient in English, and recently active on the Prolific platform. To increase sample representativeness, the research team stratified the study population with regard to sex and ethnicity (according to the UK census data from 2011) and recruited each strata using separate study advertisements that were identically worded. Informed consent was obtained from all participants, and they received monetary compensation for their participation. Each participant was instructed to respond to 117 candidate items and a demographics questionnaire.

Of the 1180 participants that completed the study, 76 were excluded in total, leaving 1104 participants in the final analysis. Of these, 7 completed the study faster than our minimum required time threshold of 5 minutes, 3 reported not responding honestly, and 66 answered with only 1 response option in the Unmind Index. Some of the excluded participants met more than one of these criteria. Mean age was 40.0 (SD 9.8) years, with 49.8% (550/1104) of participants identifying as female, 49.8% (550/1104) as male, and 0.4% (4/1104) as other. Regarding ethnicity, 6.9% (77/1104) participants identified as Asian/Asian-British, 3.1% (34/1104) as Black/African/Caribbean/Black British, 2.1% (23/1104) as Mixed, 0.8% (9/1104) as Other, and 87.1% (961/1104) as White.

Measures

The Unmind Index uses a reporting period of the past 2 weeks. Respondents are shown the prompt “During the past two weeks I have...”, followed by the item text (eg, “been feeling cheerful or bright in my mood”) and are asked to rate how often each item applies to them on a 6-point Likert scale from “No days” (0) to “Every day” (5). A 6-point scale was chosen as previous evidence suggests that middle response options are often misinterpreted by respondents and can encourage deviation to the mean [29,30]. To ensure the final Unmind Index would be brief enough to encourage regular completion by users of the Unmind platform, we committed to an upper limit of 29 items in total, with a minimum of 3 items per construct (based on recommendations by Hair and colleagues [31]).

Statistical Analysis

We took a 2-step data-driven approach to selecting items to include in the Unmind Index. In the first step, we performed single-factor EFA for each of the 7 subscales (*Happiness*, *Calmness*, *Coping*, *Health*, *Sleep*, *Energy*, and *Vitality*) separately and removed items with factor loadings <.7 (a stringent cut-off). This step was repeated iteratively for each subscale until a satisfactory set of items remained for each factor. All EFA analyses used the psych package for R [32].

In the second step, we combined the items identified in the first step and performed a multifactor EFA. As the various subscales were expected to be related, we used an oblimin rotation. To ensure the data were suitable for factor analysis, we assessed the Bartlett test of sphericity and the Kaiser-Meyer-Olkin test of sampling adequacy, with .5 taken as the minimal acceptance level [33]. The number of factors to retain was determined using Horn parallel analysis with 5000 iterations [34], implemented in the paran package for R [35]. Items that did not load on any factor with a loading >.4 were dropped at this stage.

Given the primary purpose of the Unmind Index is to direct users to content on the Unmind platform, it was decided that the factor structure of the Unmind Index should mirror the structure of this content wherever possible. For this reason, we made minor changes to the factor structure identified by EFA to accommodate these theoretical and practical constraints.

Finally, to test whether it was appropriate to combine the factors identified at this stage into a single overall MHWB score, we examined the proportion of variance in the final items selected that could be explained by a single-factor model.

Results

Using the iterative, single-factor EFA procedure outlined in the previous section, the item pool was reduced from 118 items to 57 items across the 7 scales. The Kaiser-Meyer-Olkin measure of sampling adequacy for the reduced item pool was high at .99, and the Bartlett test of sphericity was significant (χ^2_{56}= 62376.6, *P*<.001), indicating the items were appropriate for factor analysis. We then performed multifactor factor analysis on this pool of 57 items. Parallel analysis revealed that the eigenvalues of the randomly generated data were exceeded by the first 9 eigenvalues in our data set, and thus, 9 factors were extracted and rotated.

Of these factors, 5 corresponded to our predefined constructs of *Happiness*, *Coping*, *Health*, and *Sleep*. Items intended to assess calmness loaded onto 2 separate factors, 1 reflecting somatic feelings of tension (*Tension*) and 1 reflecting the cognitive experience of worrying (*Worry*). We combined these to form a single factor, *Calmness*. Items intended to measure the *Vitality* construct loaded onto multiple factors: 1 reflecting interpersonal relationships (*Connection*), 1 relating to meaning and purpose in life (*Purpose*), and 1 relating to a sense of achievement or accomplishment (*Achievement*). On practical grounds, we retained the *Connection* factor and combined *Purpose* and *Achievement* to create a new factor, *Fulfilment*. None of the factors identified reflected the predefined *Energy* construct, and items intended to measure this construct either did not load on any factor or loaded weakly on *Happiness*, *Health*, or *Fulfilment*. We therefore did not include *Energy* as a subscale. At this point, we excluded 31 items with factor loadings <.4.

Following these changes, 26 items remained in the Unmind Index, measuring 7 factors. These factors were *Happiness* (5 items), *Calmness* (4 items), *Coping* (3 items), *Sleep* (3 items), *Health* (3 items), *Connection* (3 items), and *Fulfilment* (5 items). Finally, there were substantial positive correlations between all factors, and we found that a single factor could explain 51.9% of the variance in these 26 items, indicating that combining factor scores to obtain a total would be appropriate.

Study 2: Scale Validation

Methods

Participants

To validate the Unmind Index developed in Study 1, a new sample of participants (n=1000) was recruited via the Prolific platform. Inclusion criteria were equivalent to Study 1. The sample composition was representative of the UK population with respect to age, sex, and ethnicity (a feature developed by Prolific but not yet available at the time of Study 1). To recruit a nationally representative sample, Prolific utilizes participants' prescreening responses to stratify their participant pool. Based on guidelines from the UK Office of National Statistics, age is stratified into 5 bands of 9 years each (18-27, 28-37, 38-47, 48-57, and ≥58 years), sex into male and female, and ethnicity into 5 categories (Asian, Black, Mixed, Other, and White), resulting in 50 subgroups. Using 2011 UK census data, Prolific automatically calculates the proportion of each subgroup in the UK national population and allocates participants accordingly.

Mean reported age was 46.1 (SD 15.7) years, with 51.2% (500/976) of participants identifying as female, 48.7% (475/976) identifying as male, and 1 identifying as Other. For ethnicity, 84.8% (828/976) identified as White, 7.1% (69/976) as Asian/Asian British, 3.8% (37/976) as Black/African/Caribbean/Black British, 2.5% (24/976) as Mixed, and 1.8% (18/976) as Other. To examine test-retest reliability, 250 participants were asked to repeat the new measure 1 week later, of whom 240 completed the follow-up. Mean age of the retest group was 48.1 (SD 15.5) years; 49.2% (118/240) of participants identified as female, and 50.8% (122/240) identified as male. For ethnicity, 86.7% (208/240) identified as White, 5.8% (14/240) as Asian/Asian British, 3.3% (8/240) as Black/African/Caribbean/Black British, 2.9% (7/240) as Mixed, and 1.3% (3/240) as Other.

Measures

Participants responded to the 26-item Unmind Index developed in Study 1, with items presented in randomized order. They also completed a demographics questionnaire matching the one that was used in Study 1B and a battery of existing self-report measures to allow for testing of convergent and discriminant validity for each well-being subconstruct. Each existing measure was expected to correlate positively or negatively with 1 Unmind Index subscale or with the overall Unmind Index score. The external measures used are summarized in Table 1.

Table 1. Convergent and discriminant validity measures used in Study 2.

Measure	Label/abbreviation	Domain	Items	Subscales	Response options	Score range	Reliability (α)	Unmind Index subscale
Patient Health Questionnaire 9 [16]	PHQ-9	Depression	9	_[a]	4	0-27	.90	Happiness
General Anxiety Disorder 7 [17]	GAD-7	Anxiety	7	-	4	0-21	.93	Calmness
Hospital Anxiety and Depression Scale [36]	HADS	Anxiety, depression	14	Anxiety, Depression	4	0 - 21	.90 (Anxiety), .86 (Depression)	Calmness (Anxiety), Happiness (Depression)
Perceived Stress Scale [37]	PSS	Stress	10	-	5	0-40	.92	Coping
Insomnia Severity Index [18]	ISI	Sleep disorders	7	-	4	0-28	.91	Sleep
Revised UCLA Loneliness Scale [38]	ULS-20	Loneliness and social isolation	20	-	4	20-80	.95	Connection
PROMIS[b] Global Health [39]	PROMIS-10	Mental, physical, and overall health	10	Metal health, Physical health, Combined health	5[c]	4-20 (subscales); 10-50 (combined)	.85 (Mental), .71 (Physical), .88 (Combined)	Health (PROMIS Physical)
Brief Inventory of Thriving [21]	BIT	Positive well-being	10	-	5	1-5	.93	Fulfilment
Warwick-Edinburgh Mental Well-being Scale [20]	WEMWBS	Overall well-being	14	-	5	14-70	.95	Total score
Ten-Item Personality Inventory [40]	TIPI	Big five personality traits	10	Extraversion, Agreeableness, Conscientiousness, Emotional stability, Openness	7	2-14	.77 (Extraversion), .46 (Agreeableness), .66 (Conscientiousness), .77 (Emotional stability), .42 (Openness)	None (control measure)

[a]The measure does not have subscales.

[b]PROMIS: Patient-Reported Outcomes Measurement Information System.

[c]PROMIS-10 includes a 10-point pain scale that was recoded to a 5-point scale.

Statistical Analysis: Confirmatory Factor Analysis

All statistical analyses were performed in R [41]. To assess the factor structure of the Unmind Index, we compared a variety of possible CFA models: a correlated factors model, a bifactor model, and a second-order model. Models were fit using the lavaan package for R [42] using maximum-likelihood estimation with robust Huber-White standard errors and fit statistics. In all models, each of the 26 items loads onto 1 of 7 Unmind Index subscales (*Happiness*, *Sleep*, *Coping*, *Calmness*, *Health*, *Connection*, and *Fulfilment*) in line with the results of the EFA reported in the previous section.

Models differed in how the relationship between these subscales was conceptualized. In the correlated factors model, the full covariance between each subscale is modelled explicitly. This approach can provide a flexible fit to the data but is complex to report to end users and does not provide an overall total score. We therefore also considered 2 simpler alternative models. In the bifactor model, all items load onto a general well-being factor, and each item also loads onto its specified subfactors. Subscale scores in the bifactor model reflect users' scores on these subfactors controlling for overall well-being (eg, scores on the *Happiness* subscale reflect whether a user is more or less happy than would be expected, given their overall score). As such, subscale scores from the bifactor model may be more difficult for users to interpret. In the second-order model, the 7 subscales load onto an overall general factor, and the subscales are assumed to be uncorrelated once the common effect of this general is taken into account. The second-order model is a special case of the bifactor model, with proportionality constraints on particular weights [43]. However, this model corresponded to our common-sense idea of how the Unmind Index is structured (eg, the various happiness items reflect different facets of the *Happiness* subscale, and our various subscales reflect different facets of MHWB).

Model fit was evaluated using several indices: comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean residual (SRMR). The CFI and TLI measure whether a given model fits the data better than a more restricted baseline model,

with the TLI applying a penalty to more complex models (and thus being the conservative index of the two). RMSEA is an absolute fit index, in that it assesses how far a hypothesized model is from a perfect model. SRMR outputs the average discrepancy between the model-estimated statistics and observed sample statistics. A model fit >.90 was considered acceptable for both CFI and TLI, and >.95 was considered good. For RMSEA and SRMR, a value between .06 and .08 was considered an acceptable fit, while a value <.06 was considered a good fit [44,45].

Given the large sample size, even extremely small differences in model fit are likely to be statistically significant. As a result, null hypothesis significance testing was not appropriate here, and we instead used information criteria (IC) for formal model comparison. The Akaike information criterion (AIC) is an estimate of expected out-of-sample prediction error, and the model with the lowest AIC is expected to provide the most accurate predictions on new data. The Bayesian information criterion (BIC) is proportional to an approximation of marginal likelihood of a model, and the model with the lowest BIC has the greatest posterior probability of being the true model, assuming one of the models considered is true. With large sample sizes, AIC will favor more complicated models than BIC, since an overcomplex model can still produce accurate predictions, given adequate data [46]. We therefore relied on the BIC when the criteria disagreed. Absolute IC values are not informative, so to facilitate comparisons between models, it is customary to subtract the score of the best fitting model from all models and report differences between the best model (ΔIC=0) and the competitors (ΔIC>0) [46].

Statistical Analysis: Test-Retest Reliability

One-week test-retest reliability for the Unmind Index was assessed by computing 2-way consistency intraclass correlation coefficients (ICC [C, 1]) using data collected from a subsample of the Study 2 population (n=238, after 12 dropouts). The sample size was based on a previously recommended item-respondent ratio of at least 1:5 [47].

Statistical Analysis: Internal Consistency

To determine the internal consistency of the Unmind Index, we computed the Cronbach α [48] given it is the most widely used index of the reliability of a scale to date. As the tau equivalence assumption of α is rarely met in practice [49], we also calculated coefficient omega (ω) [50] as an indicator of internal consistency. We found little difference between α and ω for each subscale.

Statistical Analysis: Convergent and Discriminant Validity

The existing measures of mental health and personality used in this study, and the Unmind Index subscales they were expected to correlate with, are summarized in Table 1. We expected the following to be negatively correlated: PHQ-9 [16] with the *Happiness* subscale, GAD-7 [17] with the *Calmness* subscale, the Hospital Anxiety and Depression Scale (HADS) [36] anxiety subscale with the *Calmness* subscale, HADS depression subscale with the *Happiness* subscale, the Perceived Stress Scale (PSS) [37] with the *Coping* subscale, and the ISI [18] with the *Sleep* subscale. We expected the following to be positively correlated: the physical health subscale of PROMIS-10 (Patient-Reported Outcomes Measurement Information System) Global Health [39] with the *Health* subscale, BIT [21] with the *Fulfilment* subscale, and WEMWBS [20] with the Unmind Index overall score.

To establish the discriminant validity of the Unmind Index, we also included the Ten-Item Personality Inventory (TIPI) [40], a brief scale that measures individual differences in the "Big Five" personality traits (extraversion, agreeableness, conscientiousness, emotional stability, and openness to experiences). These personality subscales were expected to correlate only weakly with the Unmind Index subscales, as the Unmind Index is intended to capture states of mental health, rather than static traits.

Pearson correlations were computed between the battery of convergent and discriminant validity measures and Unmind Index scores and adjusted for reliability (disattenuated) using the Cronbach α estimates for each measure:

Given the strong associations typically found between various mental health measures [19], we assessed convergent validity by checking that the pattern of correlations of Unmind Index subscale scores with the relevant existing measures (eg, *Happiness* and PHQ-9) were (1) strong and (2) stronger than the correlation with less relevant existing measures (eg, *Happiness* and GAD-7). Discriminant validity was similarly assessed by checking that correlations between Unmind Index subscales and TIPI personality subscales were weak and weaker than correlations between the Unmind Index and mental health measures.

As an additional test of the validity of the Unmind Index, we explored the degree to which scores on the various Unmind Index subscales were predictive of participants' self-reported health outcomes. These results are presented in Figure S4 in Multimedia Appendix 1.

Statistical Analysis: Measurement Invariance

It is important that the Unmind Index has the same factor structure (that is, measures the same constructs) and does not show bias across age and gender groups. To test this, we carried out measurement invariance analyses, fitting a series of additional second-order models where particular sets of parameters were allowed to vary between groups (multiple group CFA). Median participant age was 47 years, and so we classed participants as either older (>47 years, n=481), or younger (≤47 years, n=495); 475 participants identified as female, and 500 participants identified as male. One participant responded "Other/Prefer not to say" on the gender question and so was excluded from this analysis.

Measurement invariance was tested as follows [51]. We began by fitting a *configural invariance* model, where both groups have the same factor structure but all parameter values are allowed to differ between groups. If this model achieves a good fit, we can conclude that both groups show the same overall

factor structure. We then compared this model to a *weak/metric invariance* model, where first- and second-level factor loadings are constrained to be equal across groups. If this constraint does not appreciably reduce model fit, we can conclude that factor weights are the same across groups. We then fit a *strong/scalar invariance* model, where item intercepts are also constrained to be equal, but factor means are allowed to differ between groups. If this does not show a poorer fit than the weak invariance model, we can conclude that item intercepts are equivalent across groups or, in other words, that any differences in factor scores are not driven by group differences on just some items. It is only appropriate to compare factor scores across groups if this final condition is met. We considered a constrained model to show poorer fit than the unconstrained alternative if the CFI decreased by more than 0.01 points [52] or if the BIC was lower for the unconstrained model. For completeness, we also report the SRMR, RMSEA, and TLI for each model.

Statistical Analysis: Group Differences

After establishing gender and age measurement invariance, we proceeded to explore gender and age differences in Unmind Index scores. To assess these trends statistically, we fit a linear regression model to each scale, with gender and age as predictors. These analyses were conducted on z-transformed scores, with an overall mean of 0 and standard deviation of 1. The regression weight for gender reflects the standardized difference between groups. The age predictor was divided by 10, so that the weight for age reflected the expected standardized difference between participants 10 years apart.

Results

Factor Structure

Average inter-item correlation was examined, and no item displayed an average inter-item correlation above .8. Further, all items had an acceptable minimum average inter-item correlation (r>.2). No Heywood cases [53] were present.

CFA model comparison results are shown in Table 2. Parameter estimates for all models are reported in Tables S4-S8 in Multimedia Appendix 1. The correlated factors model provided a good fit to the data (SRMR=0.034, RMSEA=0.048, CFI=0.967, TLI=0.962), and was the superior model according to all model fit metrics considered. However, we considered this factor structure to be too complex to be interpretable by users. This structure also does not provide an overall MHWB score, one of our goals for the Unmind Index. We therefore decided not to use this model to score the Unmind Index. The bifactor and second-order models both provided good fits to the data. Although the bifactor model (SRMR=0.046, RMSEA=0.059, CFI=0.951, TLI=0.942, ΔAIC=306, ΔBIC=331) provided a slightly better fit than the second-order model (SRMR=0.049, RMSEA=0.062, CFI=0.943, TLI=0.936, ΔAIC=448, ΔBIC=380), the differences across fit indices were marginal. We therefore preferred the simpler second-order model to score the Unmind Index, as this model better accorded with our conceptualization of the Unmind Index and provided more easily interpretable factor scores. The second-order model is illustrated in Figure 2, and parameter estimates for this model are shown in Tables 3 and 4.

Table 2. Confirmatory factor analysis (CFA) model comparison results.

Model	LL[a]	χ^2	K[b]	df[c]	SRMR[d]	RMSEA[e]	CFI[f]	TLI[g]	ΔAIC[h]	ΔBIC[i]
Correlated factors	–37047	807	73	278	.034	.048	.967	.962	0	0
Bifactor	–37196	1070	78	273	.046	.059	.951	.942	306	331
Second Order	–37285	1209	59	292	.049	.062	.943	.936	448	380

[a]LL: log-likelihood.

[b]K: number of parameters.

[c]df: degrees of freedom.

[d]SRMR: standardized root mean square residual.

[e]RMSEA: root mean square error of approximation.

[f]CFI: comparative fit index.

[g]TLI: Tucker-Lewis index.

[h]ΔAIC: difference in the Akaike information criteria between the model and the best-fitting model.

[i]ΔBIC: difference in the Bayesian information criteria between the model and the best-fitting model.

Figure 2. The second-order factor structure used for the Unmind Index.

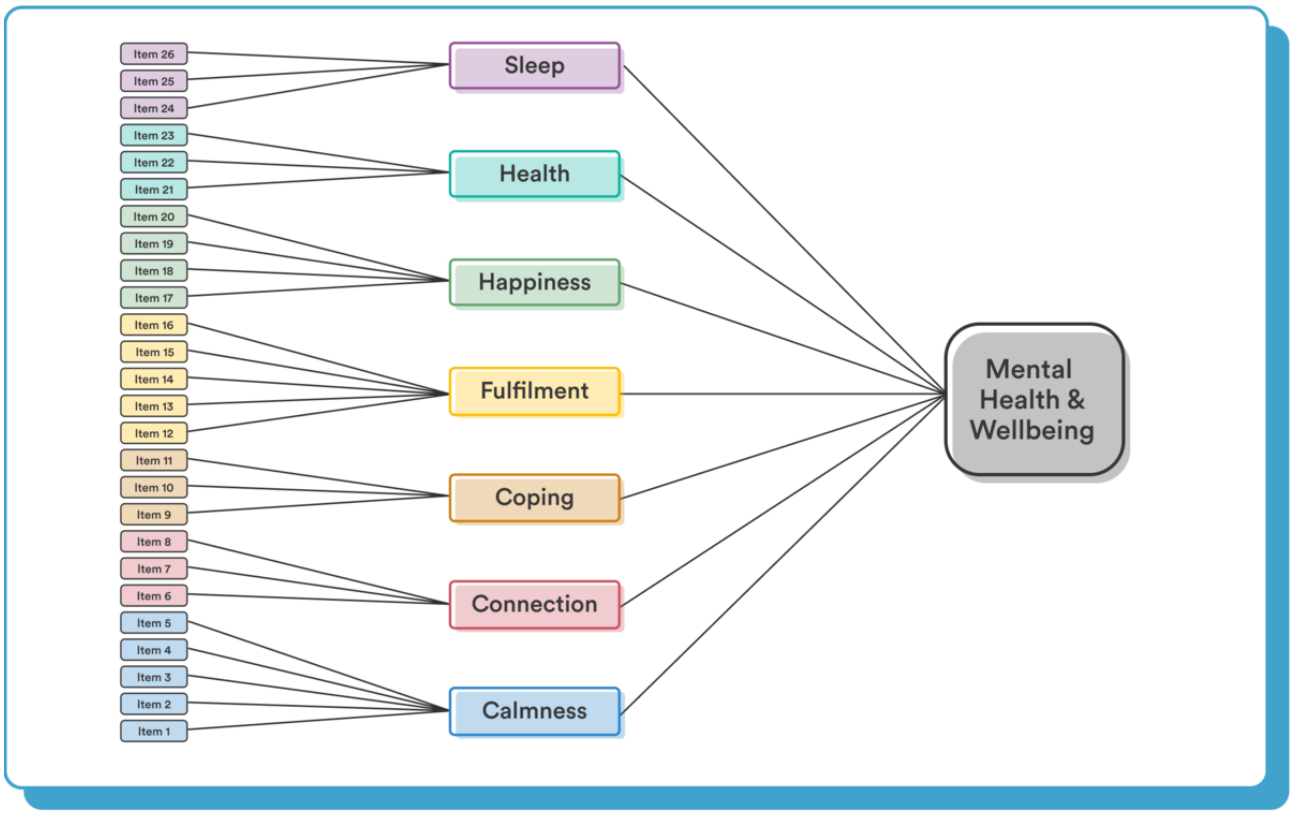


Table 3. Standardized factor loadings and residual item variances for the Unmind Index.

Factor and items	Factor loading (SE)	Residual variance (SE)	h^{2a}
Calmness			
Found it hard to stop (or control) worrying	.87 (.01)	.24 (.02)	.76
Had difficulty switching off	.76 (.02)	.42 (.03)	.58
Noticed that my body has been tense	.73 (.02)	.46 (.03)	.54
Worried that bad things might happen to me or others close to me	.67 (.02)	.56 (.03)	.44
Coping			
Felt confident that I can handle problems that come my way	.86 (.02)	.26 (.03)	.74
Been able to proactively manage my stress day to day	.74 (.02)	.45 (.03)	.55
Felt able to cope if something unexpected happens	.77 (.02)	.41 (.03)	.59
Health			
Felt like I am in a good state of health	.89 (.01)	.20 (.02)	.80
Been managing my health well	.88 (.01)	.23 (.02)	.77
Felt that my physical health is not as good as I'd like it to be (given my age/life circumstances)	.62 (.03)	.61 (.03)	.39
Sleep			
Slept well, all things considered (eg, such as caring for young children at night, snoring partner, shift work)	.90 (.01)	.19 (.02)	.81
Felt satisfied with my sleep	.91 (.01)	.18 (.02)	.82
Had trouble falling or staying asleep or waking up too early	.78 (.02)	.40 (.03)	.60
Fulfilment			
Felt a sense of accomplishment	.80 (.02)	.36 (.02)	.64
Felt that I am growing positively as a person	.77 (.02)	.41 (.03)	.59
Felt like I am leading a fulfilling life	.83 (.01)	.31 (.02)	.69
Been feeling good about myself as a person	.89 (.01)	.20 (.01)	.80
Been feeling cheerful or bright in my mood	.84 (.01)	.30 (.02)	.70
Connection			
Felt connected to people around me	.84 (.01)	.29 (.02)	.71
Felt like I have warm and trusting relationships with others	.84 (.01)	.30 (.03)	.70
Felt appreciated by others	.83 (.02)	.32 (.03)	.68
Happiness			
Had little interest in people or activities that I used to enjoy	.74 (.02)	.46 (.03)	.54
Been feeling down or sad in my mood	.86 (.01)	.25 (.02)	.75
Found it hard to motivate myself to engage with everyday tasks	.73 (.02)	.47 (.03)	.53
Felt disappointed in myself	.80 (.02)	.37 (.02)	.63
Tended to get stuck in a cycle of negativity in my head	.85 (.01)	.28 (.02)	.72

[a]h^2: item communality.

Table 4. Raw factor means, SDs, and standardized loadings onto the overall second-order factor.

Factor	Mean (SD)	Second-order factor loading (SE)
Calmness	2.92 (1.33)	.84 (.02)
Coping	2.85 (1.35)	.91 (.01)
Health	2.99 (1.18)	.79 (.02)
Sleep	2.56 (1.48)	.64 (.03)
Fulfilment	2.66 (1.31)	.94 (.01)
Connection	2.61 (1.19)	.76 (.02)
Happiness	3.03 (1.24)	.93 (.01)

Reliability and Consistency

All subscales showed excellent internal consistency, assessed by estimating Cronbach α and coefficient ω from the second-order CFA model: *Happiness*, α=.90, ω=.90; *Sleep*, α=.89, ω=.89; *Coping*, α=.83, ω=.83; *Calmness*, α=.84, ω=.85; *Health*, α=.83, ω=.83; *Connection*, α=.87, ω=.87; *Fulfilment*, α=.92, ω=.91. Internal consistency for the overall MHWB factor was also excellent: ω_H (McDonald hierarchical omega)=.92.

All subscales had excellent test-retest reliability after 1 week, based on ICCs using a 2-way mixed effects model; ICC(C, 1) scores (95% CI) for each subscale (Table 5) were as follows: *Happiness*, .84 (.79-.87); *Sleep*, .81 (.76-.85); *Coping*, .78 (.73-.83); *Calmness*, .85 (.81-.88); *Health*, .81 (.76-.85); *Connection*, .79 (.74-.83); *Fulfilment*, .85 (.81-.88); *Well-being*, .90 (.88-.92).

Table 5. Factor reliability estimates, based on internal consistency (Cronbach α and McDonald ω) and test-retest reliability (2-way consistency).

Factor	Internal consistency		Test-retest, ICC[a] (C, 1)
	Cronbach α	McDonald ω	
Total score	_[b]	.92	.90
Happiness	.90	.90	.84
Sleep	.89	.89	.81
Coping	.83	.83	.78
Calmness	.84	.85	.85
Health	.83	.83	.81
Connection	.87	.87	.79
Fulfilment	.92	.91	.85

[a]ICC: intraclass correlation coefficient.

[b]Not applicable for second-order factors.

Convergent and Discriminant Validity

Correlations between Unmind Index subscales and external measures, with correction for attenuation, are shown in Figure 3. For clarity, correlation coefficients are reversed for relationships expected to be negative, so that positive correlations indicate relationships in the expected direction. Complete correlation tables and results without disattenuation are reported in Tables S1-S2 in Multimedia Appendix 1. It is well-established that mental health measures intended to measure a variety of conditions tend to correlate strongly with each other [19]. Unmind Index subscale scores were also strongly intercorrelated (Table 6). As a result, most Unmind Index subscales correlated strongly with a range of external measures (Figure 4). Importantly, however, correlations between subscales and external measures intended to reflect similar constructs were very strong and, in almost all cases, stronger than those between subscales and the remaining external mental health measures, demonstrating convergent validity.

Figure 3. Disattenuated Pearson correlation coefficients between external measures of mental health and personality and the following Unmind Index subscales or total score: (A) Happiness, (B) Sleep, (C) Coping, (D) Calmness, (E) Health, (F) Connection, (G) Fulfilment, (H) Total Well-being score. BIT: Brief Inventory of Thriving; GAD: General Anxiety Disorder; HADS: Hospital Anxiety and Depression Scale; PHQ: Patient Health Questionnaire; PROMIS: Patient-Reported Outcomes Measurement Information System; PSS: Perceived Stress Scale; SI: Severity Index; TIPI: Ten-Item Personality Inventory; UCLA: University of California Los Angeles; WEMWBS: Warwick-Edinburgh Mental Well-being Scale.

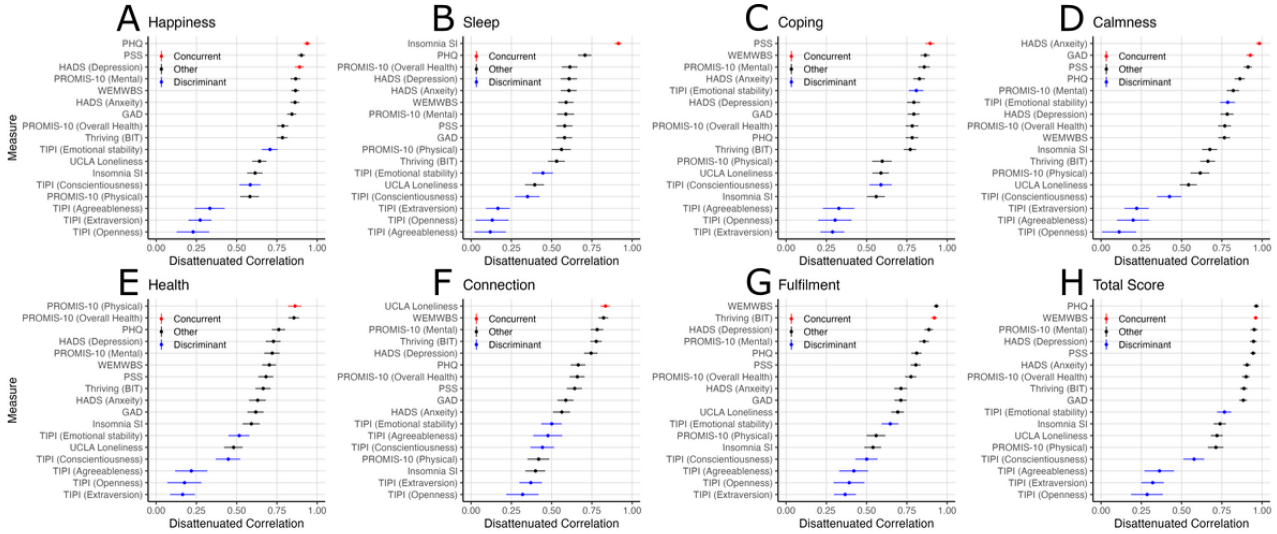


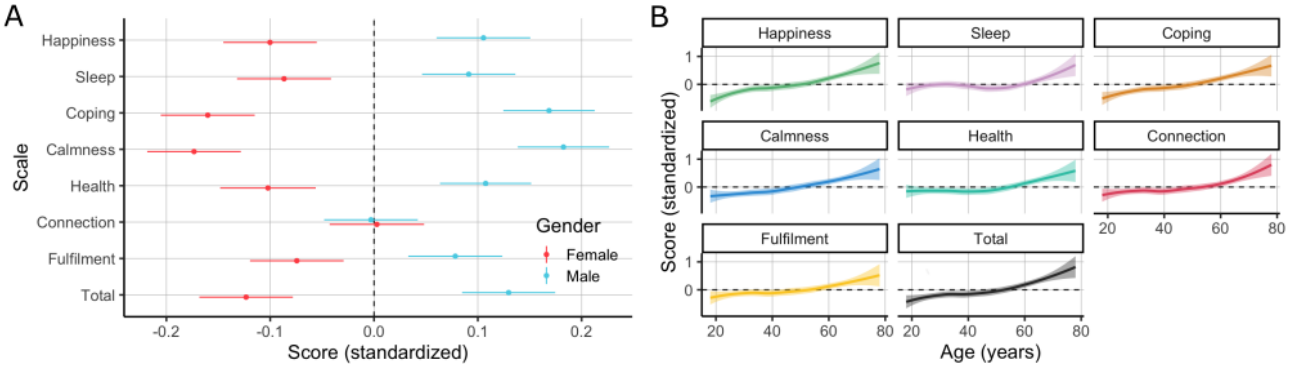
Table 6. Observed correlations between Unmind Index scales.

Variable	Calmness	Coping	Health	Sleep	Fulfilment	Connection	Happiness	Total
Calmness	-[a]	0.67 (0.02)[b]	0.55 (0.03)	0.56 (0.03)	0.60 (0.03)	0.45 (0.03)	0.79 (0.02)	0.83 (0.02)
Coping	0.67 (0.02)	-	0.57 (0.03)	0.48 (0.03)	0.75 (0.02)	0.59 (0.03)	0.72 (0.02)	0.84 (0.02)
Health	0.55 (0.03)	0.57 (0.03)	-	0.49 (0.03)	0.63 (0.02)	0.45 (0.03)	0.61 (0.03)	0.75 (0.02)
Sleep	0.56 (0.03)	0.48 (0.03)	0.49 (0.03)	-	0.52 (0.03)	0.38 (0.03)	0.52 (0.03)	0.69 (0.02)
Fulfilment	0.60 (0.03)	0.75 (0.02)	0.63 (0.02)	0.52 (0.03)	-	0.72 (0.02)	0.77 (0.02)	0.89 (0.01)
Connection	0.45 (0.03)	0.59 (0.03)	0.45 (0.03)	0.38 (0.03)	0.72 (0.02)	-	0.59 (0.03)	0.73 (0.02)
Happiness	0.79 (0.02)	0.72 (0.02)	0.61 (0.03)	0.52 (0.03)	0.77 (0.02)	0.59 (0.03)	-	0.91 (0.01)
Total	0.83 (0.02)	0.84 (0.02)	0.75 (0.02)	0.69 (0.02)	0.89 (0.01)	0.73 (0.02)	0.91 (0.01)	-

[a]Not applicable.

[b]Values in parentheses indicate standard error.

Figure 4. Standardized Unmind Index scores by (A) gender (mean and standard error of measurement within each group) and (B) age (LOWESS fit and standard error).



There were several moderate exceptions to this pattern. The Unmind Index *Happiness* subscale was strongly related to the PHQ-9 and HADS depression subscale, as expected, but was similarly related to the PSS stress measure. This suggests our *Happiness* subscale captures a broader construct than these clinical depression inventories do. This did not diminish the predicted association between the Unmind Index *Coping* subscale and the PSS. Although the Unmind Index *Fulfilment* subscale was strongly correlated with the BIT, as expected, its correlation with the WEMWBS well-being scale was slightly stronger. Finally, the Unmind Index total score was strongly associated with many measures, although this is unsurprising

given that this scale is a composite of our 7 subscales, and was most strongly correlated with WEMWBS, as expected.

Correlations between Unmind Index subscales and 4 of the 5 TIPI personality subscales (extraversion, agreeableness, conscientiousness, and openness) were generally smaller than those between the Unmind Index and any mental health measures and close to 0 in some cases, demonstrating reasonable discriminant validity. However, the TIPI emotional stability subscale ("I see myself as anxious, easily upset" [reverse-coded] and "I see myself as calm, emotionally stable") was moderately correlated with several of our subscales. It should be noted that the test-retest reliability of this TIPI subscale is estimated to be only .70 [40], suggesting that it may, in part, capture state rather than trait emotional stability.

Measurement Invariance

Gender measurement invariance results are shown in Table 7. The configural invariance model achieved good model fit across all indices. Adding metric and scalar constraints led to extremely small changes in fit and improvements in BIC, indicating that scalar invariance held across gender groups; therefore, Unmind Index scores can be directly compared between male and female users.

Table 7. Measurement invariance by gender.

Invariance model	Constraints	df[a]	χ^2	CFI[b]	BIC[c]	SRMR[d]	RMSEA[e]	TLI[f]
Configural	Factor structure	584	1796	.936	235	.051	.065	.929
Weak/metric[g]	Structure and loadings	609 (+25)	1819 (+23)	.936 (–.000)	86 (–149)	.053 (+.002)	.064 (–.001)	.932 (+.003)
Strong/scalar[g]	Structure, loadings, and item intercepts	627 (+18)	1857 (+38)	.935 (–.001)	0 (–86)	.054 (+.001)	.063 (–.000)	.933 (+.001)

[a]df: degrees of freedom.

[b]CFI: comparative fit index.

[c]BIC: Bayesian information criterion.

[d]SRMR: standardized root mean square residual.

[e]RMSEA: root mean square error of approximation.

[f]TLI: Tucker-Lewis index.

[g]Values in parentheses provide the comparisons with the less-constrained models reported in the previous row, shown as the difference between the values.

Age measurement invariance results are shown in Table 8 and reveal similar findings, indicating that scalar invariance holds across age groups; therefore, Unmind Index scores can be directly compared between older and younger users.

Table 8. Measurement invariance by age group (≥48 years vs ≤47 years).

Invariance model	Constraints	df[a]	χ^2	CFI[b]	BIC[c]	SRMR[d]	RMSEA[e]	TLI[f]
Configural	Factor structure	584	1728	.939	147	.051	.063	.932
Weak/metric[g]	Structure and loadings	609 (+25)	1778 (+50)	.937 (–.001)	25 (–122)	.059 (+.008)	.063 (–.001)	.933 (+.001)
Strong/scalar[g]	Structure, loadings, and item intercepts	627 (+18)	1877 (+99)	.933 (–.004)	0 (–25)	.060 (+.000)	.064 (+.001)	.931 (–.003)

[a]df: degrees of freedom.

[b]CFI: comparative fit index.

[c]BIC: Bayesian information criterion.

[d]SRMR: standardized root mean square residual.

[e]RMSEA: root mean square error of approximation.

[f]TLI: Tucker-Lewis index.

[g]Values in parentheses provide the comparisons with the less-constrained models reported in the previous row, shown as the difference between the values.

Group Differences

Female participants scored significantly lower than males on all scales except for Connection: total score (95% CI), b=–0.26 (–0.38 to –0.14); *Happiness*, b=–0.22 (–0.34 to –0.10); *Calmness*, b = –0.37 (–0.49 to –0.25); *Coping*, b=–0.34 (–0.46 to –0.22); *Sleep*, b=–0.18 (–0.31 to –0.06); *Health*, b=–0.22 (–0.34 to –0.09); *Fulfilment*, b=–0.16 (–0.28 to –0.04); *Connection*, b=–0.00 (–0.13 to 0.12). Older participants scored significantly higher on all scales, although the effect on *Sleep* was somewhat smaller: total score, b=0.15 (0.12 to 0.19); *Happiness*, b=0.18 (0.14 to 0.22); *Calmness*, b=0.15 (0.11 to 0.19); *Coping*, b=0.17 (0.13 to 0.20); *Sleep*, b=0.06 (0.02 to 0.10); *Health*, b=0.10 (0.06 to 0.14); *Fulfilment*, b=0.10 (0.06 to 0.14); *Connection*, b=0.11 (0.07 to 0.15).

Discussion

Summary

In Study 1A, we reported the process by which candidate items for the Unmind Index were generated, screened for validity, and initially clustered into subdomains. In Study 1B, we used an iterative data-driven approach to shorten the list of candidate items, used multifactor EFA to identify the underlying factor structure of these items, and finally integrated this data-driven factor structure with practical and theoretical considerations to establish the items and factor structure of the Unmind Index. This consists of 26 items and 7 subscales: *Happiness*, capturing positive mood or the absence of depressive symptoms; *Coping*, capturing perceived capacity to deal with stress; *Health*, capturing physical health and its impact on everyday life; *Sleep*, capturing sleep quality and its impact on functioning; *Calmness*, capturing calm or the absence of anxiety symptoms; *Connection*, capturing a sense of feeling supported and valued; and *Fulfilment*, capturing a sense of accomplishment, growth, or purpose.

These subscales differ from the 7 factors we used to guide the item generation process: *Happiness*, *Coping*, *Health*, *Sleep*, *Calmness*, *Energy*, and *Vitality*. We found that items intended to measure *Energy* did not load onto a single factor, and so, this construct was eliminated. Items intended to measure *Vitality* formed 2 factors: *Connection*, capturing the social aspects of the vitality construct, and *Fulfilment*, capturing the self-directed aspects. Although the EFA results indicated that the *Calmness* factor could be partitioned into *Worry* and *Tension*, we chose to maintain the single factor for practical reasons.

In Study 2, we validated the Unmind Index with new participants. We established that a second-order factor structure provides good fit to the data, that the scales have good internal and test-retest reliability, and that the subscales correlate as expected with existing measures of MHWB and do not correlate strongly with personality scales, with the exception of the emotional stability trait. Finally, the Unmind Index displayed measurement invariance with regard to gender and age, meaning that scores can be validly compared across these groups.

Although the second-order factor model fit the data well, it was outperformed by the correlated factors model, which directly modeled the correlations between all 7 subscales. This implies that some subscales are more closely related than others, a result that is confirmed by the information presented in Table 5. This is consistent with a growing body of work showing that the symptoms of many mental health issues largely overlap [19,54], suggesting that a smaller number of transdiagnostic features, such as cognitive inflexibility or repetitive negative thinking may underpin many mental health problems [55]. In particular, the *Calmness* and *Happiness* subscales were strongly correlated. This is unsurprising, given that these subscales are negatively associated with existing measures of anxiety and depression, respectively, and that anxiety and depression are strongly linked [56]. However, although the second-order model did not utilize this information, it provided a clear, practical structure for communicating results to users and is preferred for this reason.

Scoring

It is important that scores on the Unmind Index are easy for users to understand, can be compared across subscales, and can be compared to a meaningful reference value. For this reason, Unmind Index subscale scores reported to users are standardized to population norms estimated from this validation study, with a mean of 100 and a standard deviation of 15. This makes scores directly interpretable by users in a way that is not the case for unstandardized measures and allows for direct comparisons between subscale scores. It is also in line with recent appeals [57] that mental health measures should be reported in a way that makes scores across measures comparable.

Limitations and Future Directions

A number of limitations and directions for future work remain. The Unmind Index asks respondents to report their mental state over the previous 2 weeks. It is not yet known to what extent Unmind Index scores fluctuate over time, although our high test-retest reliability indicates that scores do not change considerably over a single week. Further work is also needed to determine to what degree the Unmind Index is sensitive to changes in mental health. To address this, we are currently including the Unmind Index as a secondary outcome measure in randomized controlled efficacy trials, with the intention of testing whether pre-post changes in existing measures such as the PHQ-8 are predictive of changes in Unmind Index scores.

We reported results from (exploratory and confirmatory) linear factor analyses in this paper. However, responses to the Unmind Index are given on a 6-point Likert scale, from “No days” to “Every day.” In future work, we will reanalyze these data using multivariate item response theory modelling [58]. Doing so will allow us to better understand how users make use of this response scale and may lead to an adaptive version of the Unmind Index, where the questions asked are calibrated to individual users’ score profiles.

Lastly, our validation is currently limited to a UK population, and we acknowledge that the subjective experience of mental health and conceptualization of well-being can vary across cultures [59]. We are planning future studies to validate the Unmind Index in other geographies and establish relevant norms and scoring bandings.

Conclusion

This work demonstrated the Unmind Index is a robust measure of MHWB that is underpinned by a general factor and 7 underlying constructs. We suggest that MHWB can usefully be measured in conjunction, challenging the false dichotomy (and associated stigma) that is perpetuated when mental ill health and mental well-being are described and measured separately. This is particularly relevant for assessment offered to working adults who are likely to encompass the full spectrums of MHWB. We would encourage other mHealth app developers to capture the broader aspects of positive well-being when aiming to measure mental health.

Acknowledgments

The authors would like to thank Juan Giraldo and Dean Ottewell for conceptual input and Steve Dineur for assistance with the design of figures.

Authors' Contributions

AS, ET, ME, and HB conceptualized the study. AS and ME collected the data. AS, ET, and BSL analyzed the data. AS, ET, and HB drafted the manuscript. All authors were involved in revising the manuscript. BSL and LS consulted on the study.

Conflicts of Interest

AS, ET, ME, and HB are employed by and own share options in Unmind Ltd. They created the Unmind Index that was developed and validated in this study. The University of Cambridge Psychometrics Centre (with which BSL and LS are affiliated) was contracted as an academic partner to provide research consulting services to Unmind Ltd for the purposes of this study and received financial compensation for this work.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 1095 KB - mental_v9i1e34103_app1.docx]

References

1. Pinheiro M, Ivandic I, Razzouk D. The Economic Impact of Mental Disorders and Mental Health Problems in the Workplace. In: Razzouk D, editor. Mental Health Economics. Cham, Switzerland: Springer International Publishing; 2017:415-430.
2. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 2013 Nov 09;382(9904):1575-1586. [doi: 10.1016/S0140-6736(13)61611-6] [Medline: 23993280]
3. Hampson E, Jacob A. Mental health and employers: Refreshing the case for investment. Deloitte. 2020 Jan. URL: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consultancy/deloitte-uk-mental-health-and-employers.pdf [accessed 2021-12-16]
4. Deady M, Glozier N, Calvo R, Johnston D, Mackinnon A, Milne D, et al. Preventing depression using a smartphone app: a randomized controlled trial. Psychol. Med 2020 Jul 06:1-10. [doi: 10.1017/s0033291720002081]
5. Furber G, Segal L, Leach M, Turnbull C, Procter N, Diamond M, et al. Preventing mental illness: closing the evidence-practice gap through workforce and services planning. BMC Health Serv Res 2015 Jul 24;15(1):283 [FREE Full text] [doi: 10.1186/s12913-015-0954-5] [Medline: 26205006]
6. Tan L, Wang M, Modini M, Joyce S, Mykletun A, Christensen H, et al. Erratum to: preventing the development of depression at work: a systematic review and meta-analysis of universal interventions in the workplace. BMC Med 2014 Nov 13;12(1):1. [doi: 10.1186/s12916-014-0212-4]
7. Chisholm D, Sweeny K, Sheehan P, Rasmussen B, Smit F, Cuijpers P, et al. Scaling-up treatment of depression and anxiety: a global return on investment analysis. The Lancet Psychiatry 2016 May;3(5):415-424. [doi: 10.1016/S2215-0366(16)30024-4]
8. Stevenson D, Farmer P. Thriving at work: a review of mental health and employers. gov.uk. 2017. URL: https://www.gov.uk/government/publications/thriving-at-work-a-review-of-mental-health-and-employers [accessed 2021-12-16]
9. Azzone V, McCann B, Merrick EL, Hiatt D, Hodgkin D, Horgan C. Workplace stress, organizational factors and EAP utilization. J Workplace Behav Health 2009;24(3):344-356 [FREE Full text] [doi: 10.1080/15555240903188380] [Medline: 24058322]
10. Galderisi S, Heinz A, Kastrup M, Beezhold J, Sartorius N. Toward a new definition of mental health. World Psychiatry 2015 Jun 04;14(2):231-233 [FREE Full text] [doi: 10.1002/wps.20231] [Medline: 26043341]
11. Dugas M, Gao G, Agarwal R. Unpacking mHealth interventions: A systematic review of behavior change techniques used in randomized controlled trials assessing mHealth effectiveness. Digit Health 2020 Feb 20;6:2055207620905411 [FREE Full text] [doi: 10.1177/2055207620905411] [Medline: 32128233]
12. Szinay D, Jones A, Chadborn T, Brown J, Naughton F. Influences on the uptake of and engagement with health and well-being smartphone apps: systematic review. J Med Internet Res 2020 May 29;22(5):e17572 [FREE Full text] [doi: 10.2196/17572] [Medline: 32348255]
13. Kauer SD, Reid SC, Crooke AHD, Khor A, Hearps SJC, Jorm AF, et al. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. J Med Internet Res 2012 Jun 25;14(3):e67 [FREE Full text] [doi: 10.2196/jmir.1858] [Medline: 22732135]
14. Wichers M, Simons CJP, Kramer IMA, Hartmann JA, Lothmann C, Myin-Germeys I, et al. Momentary assessment technology as a tool to help patients with depression help themselves. Acta Psychiatr Scand 2011 Oct;124(4):262-272. [doi: 10.1111/j.1600-0447.2011.01749.x] [Medline: 21838742]

15. Franken K, Lamers SM, Ten Klooster PM, Bohlmeijer ET, Westerhof GJ. Validation of the Mental Health Continuum-Short Form and the dual continua model of well-being and psychopathology in an adult mental health setting. J Clin Psychol 2018 Dec 05;74(12):2187-2202 [FREE Full text] [doi: 10.1002/jclp.22659] [Medline: 29978482]
16. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
17. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
18. Bastien C, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med 2001 Jul;2(4):297-307. [doi: 10.1016/s1389-9457(00)00065-4] [Medline: 11438246]
19. Caspi A, Houts RM, Belsky DW, Goldman-Mellor SJ, Harrington H, Israel S, et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? Clin Psychol Sci 2014 Mar 14;2(2):119-137 [FREE Full text] [doi: 10.1177/2167702613497473] [Medline: 25360393]
20. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007 Nov 27;5:63 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
21. Su R, Tay L, Diener E. The development and validation of the Comprehensive Inventory of Thriving (CIT) and the Brief Inventory of Thriving (BIT). Appl Psychol Health Well Being 2014 Nov 12;6(3):251-279. [doi: 10.1111/aphw.12027] [Medline: 24919454]
22. Holden RR. The Corsini Encyclopedia of Psychology Face Validity. In: Weiner IB, Craighead WE, editors. The Corsini Encyclopedia of Psychology. Hoboken, NJ: John Wiley & Sons, Inc; 2010.
23. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. West J Nurs Res 2003 Aug 01;25(5):508-518. [doi: 10.1177/0193945903252998] [Medline: 12955968]
24. Kyriazos TA. Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. PSYCH 2018;09(08):2207-2230. [doi: 10.4236/psych.2018.98126]
25. Wang J, Wang X, editors. Structural Equation Modeling: Applications Using Mplus. Hoboken, NJ: John Wiley & Sons, Inc; 2012.
26. Prolific. URL: https://prolific.co/ [accessed 2021-12-16]
27. Gorilla. URL: https://gorilla.sc/ [accessed 2021-12-16]
28. Peer E, Brandimarte L, Samat S, Acquisti A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology 2017 May;70:153-163. [doi: 10.1016/j.jesp.2017.01.006]
29. Nadler JT, Weston R, Voyles EC. Stuck in the middle: the use and interpretation of mid-points in items on questionnaires. J Gen Psychol 2015;142(2):71-89. [doi: 10.1080/00221309.2014.994590] [Medline: 25832738]
30. Kulas JT, Stachowski AA, Haynes BA. Middle response functioning in Likert-responses to personality items. J Bus Psychol 2008 Jan 24;22(3):251-259. [doi: 10.1007/s10869-008-9064-2]
31. Hair JF, Black B, Black WC, Babin RJ, Anderson RE. Multivariate Data Analysis: Global Edition, 7th Edition. New York City, NY: Pearson Education; 2010.
32. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. The Comprehensive R Archive Network. 2015. URL: https://CRAN.R-project.org/package=psych [accessed 2021-12-16]
33. Kaiser HF. An index of factorial simplicity. Psychometrika 1974 Mar;39(1):31-36. [doi: 10.1007/BF02291575]
34. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika 1965 Jun;30(2):179-185. [doi: 10.1007/bf02289447]
35. Dinno A. paran: Horn's Test of Principal Components/Factors. The Comprehensive R Archive Network. 2018. URL: https://CRAN.R-project.org/package=paran [accessed 2021-12-16]
36. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
37. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
38. Russell D, Peplau LA, Cutrona CE. The revised UCLA Loneliness Scale: Concurrent and discriminant validity evidence. Journal of Personality and Social Psychology 1980 Sep;39(3):472-480. [doi: 10.1037/0022-3514.39.3.472]
39. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009 Sep 19;18(7):873-880 [FREE Full text] [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
40. Gosling SD, Rentfrow PJ, Swann WB. A very brief measure of the Big-Five personality domains. Journal of Research in Personality 2003 Dec;37(6):504-528. [doi: 10.1016/S0092-6566(03)00046-1]
41. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. URL: https://www.R-project.org/ [accessed 2021-12-16]
42. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. J. Stat. Soft 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
43. Yung Y, Thissen D, McLeod LD. On the relationship between the higher-order factor model and the hierarchical factor model. Psychometrika 1999 Jun;64(2):113-128. [doi: 10.1007/bf02294531]

44. Cangur S, Ercan I. Comparison of model fit indices used in structural equation modeling under multivariate normality. J. Mod. App. Stat. Meth 2015 May 01;14(1):152-167. [doi: 10.22237/jmasm/1430453580]
45. Hooper D, Coughlan J, Mullen M. Structural equation modelling: guidelines for determining model fit. Electronic Journal of Business Research Methods 2008;6(1):53-60 [FREE Full text]
46. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York, NY: Springer; 2007.
47. Park MS, Kang KJ, Jang SJ, Lee JY, Chang SJ. Evaluating test-retest reliability in patient-reported outcome measures for older people: A systematic review. Int J Nurs Stud 2018 Mar;79:58-69. [doi: 10.1016/j.ijnurstu.2017.11.003] [Medline: 29178977]
48. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334. [doi: 10.1007/BF02310555]
49. Yang Y, Green SB. Coefficient alpha: a reliability coefficient for the 21st century? Journal of Psychoeducational Assessment 2011 May 19;29(4):377-392. [doi: 10.1177/0734282911406668]
50. McDonald RP. Test theory: A unified treatment. Hove, East Sussex, United Kingdom: Psychology Press; 2013.
51. van de Schoot R, Kluytmans A, Tummers L, Lugtig P, Hox J, Muthén B. Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. Front Psychol 2013;4:770 [FREE Full text] [doi: 10.3389/fpsyg.2013.00770] [Medline: 24167495]
52. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal 2002 Apr;9(2):233-255. [doi: 10.1207/S15328007SEM0902_5]
53. Krijnen WP, Dijkstra TK, Gill RD. Conditions for factor (in)determinacy in factor analysis. Psychometrika 1998 Dec;63(4):359-367. [doi: 10.1007/bf02294860]
54. Yee CM, Javitt DC, Miller GA. Replacing DSM categorical analyses with dimensional analyses in psychiatry research: the research domain criteria initiative. JAMA Psychiatry 2015 Dec 01;72(12):1159-1160. [doi: 10.1001/jamapsychiatry.2015.1900] [Medline: 26559005]
55. Morris L, Mansell W. A systematic review of the relationship between rigidity/flexibility and transdiagnostic cognitive and behavioral processes that maintain psychopathology. Journal of Experimental Psychopathology 2018 Jul 19;9(3):204380871877943. [doi: 10.1177/2043808718779431]
56. Dobson KS. The relationship between anxiety and depression. Clinical Psychology Review 1985 Jan;5(4):307-324. [doi: 10.1016/0272-7358(85)90010-8]
57. Fried EI, Böhnke JR, de Beurs E. Common measures or common metrics? A plea to harmonize measurement results. PsyArXiv Preprints. 2021. URL: https://psyarxiv.com/m4qzb/ [accessed 2021-12-16]
58. van der Linden WJ, Hambleton RK. Handbook of Modern Item Response Theory. New York, NY: Springer; 2013.
59. Gopalkrishnan N. Cultural diversity and mental health: considerations for policy and practice. Front Public Health 2018 Jun 19;6:179 [FREE Full text] [doi: 10.3389/fpubh.2018.00179] [Medline: 29971226]

Abbreviations

AIC: Akaike information criterion
BIC: Bayesian information criterion
BIT: Brief Inventory of Thriving
CFA: confirmatory factor analysis
CFI: comparative fit index
DSM: Diagnostic and Statistical Manual of Mental Disorders
EFA: exploratory factor analysis
GAD-7: General Anxiety Disorder 7
HADS: Hospital Anxiety and Depression Scale
I-CVI: item content validity index
IC: information criteria
ICC: intraclass correlation coefficient
ICD: International Classification of Diseases
ISI: Insomnia Severity Index
mHealth: mobile health.
MHWB: mental health and well-being
PHQ-9: Patient Health Questionnaire 9
PROMIS: Patient-Reported Outcomes Measurement Information System
PSS: Perceived Stress Scale
RMSEA: root mean square error of approximation
SRMR: standardized root mean residual
TIPI: Ten-Item Personality Inventory

TLI: Tucker-Lewis index
WEMWBS: Warwick-Edinburgh Mental Well-being Scale

Edited by G Eysenbach; submitted 08.10.21; peer-reviewed by A Tannoubi; comments to author 02.11.21; accepted 21.11.21; published 17.01.22.

Please cite as:
Sierk A, Travers E, Economides M, Loe BS, Sun L, Bolton H
A New Digital Assessment of Mental Health and Well-being in the Workplace: Development and Validation of the Unmind Index
JMIR Ment Health 2022;9(1):e34103
URL: https://mental.jmir.org/2022/1/e34103
doi:10.2196/34103
PMID:35037895

Original Paper

Automatic Assessment of Emotion Dysregulation in American, French, and Tunisian Adults and New Developments in Deep Multimodal Fusion: Cross-sectional Study

Federico Parra[1], PhD; Yannick Benezeth[1], PhD; Fan Yang[1], PhD

LE2I EA 7508, Université Bourgogne Franche-Comté, Dijon, France

Corresponding Author:
Federico Parra, PhD
LE2I EA 7508
Université Bourgogne Franche-Comté
UFR Sciences et techniques, avenue Alain Savary
Dijon, 21000
France
Phone: 33 782132695
Email: federico.parra@hotmail.com

Abstract

Background: Emotion dysregulation is a key dimension of adult psychological functioning. There is an interest in developing a computer-based, multimodal, and automatic measure.

Objective: We wanted to train a deep multimodal fusion model to estimate emotion dysregulation in adults based on their responses to the Multimodal Developmental Profile, a computer-based psychometric test, using only a small training sample and without transfer learning.

Methods: Two hundred and forty-eight participants from 3 different countries took the Multimodal Developmental Profile test, which exposed them to 14 picture and music stimuli and asked them to express their feelings about them, while the software extracted the following features from the video and audio signals: facial expressions, linguistic and paralinguistic characteristics of speech, head movements, gaze direction, and heart rate variability derivatives. Participants also responded to the brief version of the Difficulties in Emotional Regulation Scale. We separated and averaged the feature signals that corresponded to the responses to each stimulus, building a structured data set. We transformed each person's per-stimulus structured data into a *multimodal codex*, a grayscale image created by projecting each feature's normalized intensity value onto a cartesian space, deriving each pixel's position by applying the Uniform Manifold Approximation and Projection method. The codex sequence was then fed to 2 network types. First, 13 convolutional neural networks dealt with the spatial aspect of the problem, estimating emotion dysregulation by analyzing each of the codified responses. These convolutional estimations were then fed to a transformer network that decoded the temporal aspect of the problem, estimating emotional dysregulation based on the *succession* of responses. We introduce a Feature Map Average Pooling layer, which computes the mean of the convolved feature maps produced by our convolution layers, dramatically reducing the number of learnable weights and increasing regularization through an ensembling effect. We implemented 8-fold cross-validation to provide a good enough estimation of the generalization ability to unseen samples. Most of the experiments mentioned in this paper are easily replicable using the associated Google Colab system.

Results: We found an average Pearson correlation (r) of 0.55 (with an average P value of <.001) between ground truth emotion dysregulation and our system's estimation of emotion dysregulation. An average mean absolute error of 0.16 and a mean concordance correlation coefficient of 0.54 were also found.

Conclusions: In psychometry, our results represent excellent evidence of convergence validity, suggesting that the Multimodal Developmental Profile could be used in conjunction with this methodology to provide a valid measure of emotion dysregulation in adults. Future studies should replicate our findings using a hold-out test sample. Our methodology could be implemented more generally to train deep neural networks where only small training samples are available.

(JMIR Ment Health 2022;9(1):e34333) doi:10.2196/34333

KEYWORDS

emotion dysregulation; deep multimodal fusion; small data; psychometrics

Introduction

Emotion regulation is currently conceptualized as involving the following 5 distinct abilities: (1) having awareness and an understanding of one's emotions, (2) being able to accept them, (3) being able to control impulsive behaviors related to them, (4) having the capacity to behave according to our desired goals in the midst of negative emotions, and (5) having the capacity to implement emotion regulation strategies as required to meet individual goals and situational demands. The absence of these abilities indicates the presence of *emotion dysregulation* [1]. Psychopathology is characterized by intense or protracted maladaptive negative emotional experiences. Emotion dysregulation is a core vulnerability to the development of both internalizing and externalizing mental disorders [2]. For example, high emotion dysregulation is a key component of substance abuse [3], generalized anxiety disorder [4], complex posttraumatic stress disorder [5], and borderline personality disorder [6].

Emotion dysregulation is typically assessed through a self-report questionnaire, the Difficulties in Emotional Regulation Scale (DERS) [1], or one of its shorter forms (eg, Difficulties in Emotion Regulation Scale, brief version [DERS-16]) [7]. It can also be assessed physiologically by measuring heart rate variability (HRV) in a controlled experiment, with the advantage that this requires no insight from the participant and represents an objective measure. However, traditionally, this form of assessment represented serious costs of collection, and varying baselines among people posed a problem [8]. Since at least one study has shown that the DERS and the HRV-based assessment of emotion dysregulation are correlated [8], the DERS has become the de-facto "gold standard."

Attempts to measure psychological dimensions "in the wild" (ie, a naturalistic approach) using machine learning and unimodal sensing approaches, such as measuring heart rate throughout the day with a smartwatch or measuring the patterns of social media interactions by a user, have not yet produced good enough results leading to major changes in the way the mental health industry practices psychometrics. It still relies almost entirely on self-assessment questionnaires or professional interviews [9]. In our view, this absence of disruption comes down to 2 issues. First, the problem of relying on a single modality. In the field of affective computing, multimodal fusion has shown promise by beating unimodal approaches in several benchmarks [10]. This is because multimodality provides cross-validation of hypotheses, where one sense modality can reaffirm or negate what was perceived by another, reducing error and increasing reliability. This is how we, humans, perceive. Second, measuring psychological dimensions "in the wild" might be a bad idea due to the unknown number of confounding factors surrounding daily life. In particular, many authors underline the need for considering the specific demands of the situation at hand, as well as the specific goals of the individual in that context, when evaluating emotion dysregulation [1].

To overcome these limitations, in 2017, we introduced the Biometric Attachment Test (BAT) in the Journal of Medical Internet Research [11]. It was and continues to be the first automated computer test to measure adult attachment in a multimodal fashion, including physiology measures (HRV) as well as behavioral ones. The BAT uses picture and music stimuli to evoke situations and feelings related to adult attachment, such as loss, fear, parent-children relationships, or romantic relationships. It sits well within the psychometric tradition of projective tests, such as the Thematic Apperception Test [12]. In 2019, we presented a machine learning methodology to automatically score the BAT using a small training data set, and we validated the use of a remote photoplethysmography (RPPG) algorithm to measure HRV in a contactless fashion as part of the BAT software [13]. We have now renamed our test to the Multimodal Developmental Profile (MDP), because we hypothesize its stimuli and design can work for measuring not only adult attachment, but also several other dimensions of psychological functioning that are developmental in nature and crucial to the forming of psychopathology [14]. In particular, we hypothesize that the MDP can measure emotion dysregulation in adults.

Developing deep multimodal fusion models to combine the MDP obtained features in order to predict actual psychological dimensions, such as emotion dysregulation, is a challenge due in part to the small nature of samples in psychology research [13].

In this work, we propose a series of methods that we hypothesize will allow us to train a scoring model for the MDP to estimate emotion dysregulation in adults. We hypothesize that such an estimation of emotion dysregulation will have psychometric convergence with the "gold standard" measure, the DERS. Our approach of choice is particularly important for the machine learning field. We hypothesize that our methodology will unleash training deep neural networks for multimodal fusion with a very small training sample.

The organization of the rest of this paper is as follows. First, we will introduce the multimodal codex, which is the heart of our approach, and the techniques required to build it and fill its missing values. Second, we will present our convolutional neural network (CNN)-transformer network architecture, including our new layer, the Feature Map Average Pooling (FMAP) layer. Third, we will discuss our training methodology. Fourth, we will present our results, including the quality of our estimation of emotion dysregulation in adults. Lastly, we will discuss these results.

Methods

Recruitment

American Subsample

This subsample consisted of 69 participants (39 females and 30 males) and was recruited online using Amazon Mechanical Turk and Prolific services between January and July 2019. The mean age for this subsample was 35.05 years (SD 12.5 years, minimum 18 years, maximum 68 years). We did not intentionally recruit any clinical participants for this subsample, but we cannot guarantee the absence of clinical patients within it.

French Subsample

This subsample consisted of 146 participants (88 females and 58 males) recruited between the months of January and July 2019, and was formed from multiple sources in different regions of France. Of the 146 participants, 10 clinical patients were recruited at University Hospital Center Sainte-Etienne and 22 at the Ville-Evrard Center of Psychotherapy and Psychotrauma in Saint-Denis, 33 volunteers were enrolled in Paris and 19 in Lyon, 3 college students were enrolled at Paris Descartes University and 11 at University Bourgogne Franche-Comté (Dijon), and 43 clinical private practice patients were enrolled in Paris and 5 in Lyon. The mean age for this subsample was 39.25 years (SD 13.6 years, minimum 18 years, maximum 72 years). Clinical patients were included to examine whether the MDP was capable of rightly assessing more extreme emotion dysregulation cases.

Tunisian Subsample

This subsample consisted of 33 Tunisian participants (21 females and 12 males) recruited in July 2019 in the city of Tunis. The mean age was 37.6 years (SD 10.5 years, minimum 17 years, maximum 55 years). While there was no intention to recruit clinical participants for this subsample, we cannot guarantee the absence of clinical patients within it.

Measures

DERS-16

The original DERS [1] is a 36-item self-report questionnaire that measures an individual's typical level of emotion dysregulation. Internally, it is based on the following 6 different subscales: (1) nonacceptance of negative emotions, (2) inability to engage in goal-oriented behaviors when in distress, (3) difficulties for controlling impulsive behaviors when in distress, (4) limited or no access to emotion regulation strategies perceived as effective, (5) lack of awareness of one's emotions, and (6) lack of emotional clarity. Respondents have to rate items on a 5-point Likert-type scale from 1 (*almost never*) to 5 (*almost always*) depending on how much they believe each proposition applies to them. The shortened version of the DERS that we used in this work, called DERS-16 [7], consists of 16 items that assess the same 6 dimensions of emotion regulation difficulties. The total score on the DERS-16 ranges from 16 to 80, where higher scores reflect greater levels of emotion dysregulation. Importantly, this shortened version of the DERS retained excellent internal consistency, good test-retest reliability, and good convergent and discriminant validity, with only minimal differences when compared to the original DERS [7].

MDP

Explored in depth in an article in the Journal of Medical Internet Research [11], the MDP as a test consists of 14 themes or narratives that depict human experiences that can be either stressing or soothing in nature (loss, grief, and solitude, as well as human connection, romantic love, and kinship). The themes are evoked using rotating stimuli from a pool of pictures and short music clips that were vetted through a standardized procedure using crowd-sourced feedback. Some themes are evoked using picture stimuli alone, some are evoked using a combination of picture and music, and some are evoked by music alone (to evoke raw emotions such as sadness and fear). During the test situation, each stimulus is shown and/or heard for 15 seconds, after which the computer asks the participant to describe aloud what they have felt. They have 20 seconds to respond, before a 5-second break and then moving to the next stimulus. The whole session takes 9 minutes and 33 seconds to be completed.

Importantly, the first stimulus is fully neutral and allows us to acquire a baseline for all our measurements, which is later subtracted from them. In theory, this allows us to work with signals that react solely to the stimuli. Whether the participants came already upset to the test situation or whether they were already fatigued, the test will measure this during the first stimulus and then subtract it from the following signals; thus, it will only take into account whether a stimulus made them more upset or more fatigued, or perhaps whether a stimulus managed to soothe or relax them. The short duration of the test assures us that any abrupt changes in the signals from which the baseline was subtracted will indeed be caused by the test situation itself and not due to time simply passing by. Furthermore, the order of the stimuli themselves is such that stress and soothing themes are alternated, allowing us to get more contrast in our measurements of what each stimulus is doing to the person.

A simple way of conceptualizing the MDP is as a series of *dependent* experiments. Each stimulus intends to evoke a certain range of reactions on its own but is also linked to the reactions that the next stimulus intends to evoke. For example, stimulus 11 will attempt to provoke fear, and stimulus 12 will attempt to evoke loss, whereas stimulus 13 will evoke a soothing comforting experience of human connection. We will be interested in the reactions to each of those stimuli separately, but we will, more importantly, be interested in the relationship between them, for example, "If the person was upset by the first 2 stimuli, were they able to calm down during the last one?"

As the participant perceives the stimuli and responds aloud to them, the software automatically collects video and audio data and automatically extracts features from them. Specifically, the MDP uses an RPPG method to extract HRV features that allow measuring the sympathetic and parasympathetic branches of the autonomic nervous system; detects facial action units, head movements, and gaze direction with respect to the stimuli being presented; and analyzes speech, extracting paralinguistic features as well as conducting a linguistic analysis [13].

An important aspect of the MDP is that it does not rely on a naturalistic approach. Rather, it is based on a tightly controlled experiment carefully conceived and validated in order to evoke specific reactions.

In addition, the MDP has *content validity* [11], because it is underpinned by a strong theoretical foundation and interpretation. This sets it apart from most machine learning attempts at measuring mental health, which typically focus on prediction and convergence with a disregard for content validity [15].

Finally, contrary to most projects, wherein a machine learning system is trained to predict a category with relation to mental

health, such as depressed vs not depressed, the MDP is *dimensional*. It measures psychological phenomena in terms of their continuum score, from which it is easy to produce categorical decisions (whereas the opposite is impossible to accomplish). These continuum scores are far more precise and nuanced, and could allow, among other things, to conduct outcome studies, measuring the degree of change of a psychological construct over time.

Machine Learning Methodology

Important Note on Data Leakage

To prevent any form of data leaking, every step described below was conducted *within* the 8-fold cross-validation loop. This loop begins by separating the available data into a validation set and a training set containing the rest of the samples.

A few participants took the test twice at intervals of a few weeks to help with a future study on test-retest reliability, and we included both of their sessions in this study, treating them as if they were different participants. To prevent data leakage, however, when one of them was randomly put into the validation set, their other session got automatically put there as well. This explains why the validation set size changes from fold to fold (with a range of 29 to 35).

Data Preparation

All data preparation was performed in MATLAB 2021b (MathWorks). The MDP outputs a set of CSV files containing the structured data for each sense modality (facial expressions, linguistic analysis, etc). In most cases, this comes in the form of a table containing the timestamps as rows and the features as columns.

We averaged each feature per stimulus (ie, an average of values for facial action unit 10 from the moment stimulus 3 was shown till the moment it disappeared). We discounted the first stimulus's results, the neutral one (see previous section), from all others so that we dealt solely with the variance produced by the test itself. Features were scaled to the −1 to 1 range, using either previous knowledge about the actual signal's minimum and maximum values, or the empirical minimum and maximum levels found within the signal in all our training samples for a given fold.

DERS-16 scores were also linearly scaled, to the 0-1 range, to allow for quicker training times and easier interpretation of results. An important step in our data preparation procedure was to uniformize our training sample with regards to the ground truth (ie, DERS-16 scores) so that all levels of the ground truth could be equally represented in terms of the number of samples being fed to our learning algorithm. Our code did this by binning the DERS-16 score, and up-sampling our data set until all bins (ie, all score levels) had the same number of cases representing them. This, of course, presented the problem of potentially overfitting these repeated cases. In the section about test time data augmentation, we present how we dealt with this problem.

Multimodal Codex Sequence

From a clinician's perspective, a typical assessment interview can be thought of as having 2 main components as follows: what is happening *at any given moment* during the interview, that is, the specific behavioral or verbal responses a patient might show to a specific question or nonverbal queue coming from the clinician, and the manner those interpreted moments *intertwine.*

Based on years of clinical experience, we argue that the psychologist or psychiatrist ends the interview with a newly acquired succession of *intuitive mental images*, representing key moments of the encounter with the patient. These mental images encode information from multiple sense modalities: a specific word that was said as well as the tone and posture in which it was said, and how that led to a long silence. They represent an utter distillation of the experience, which is the simplest representation of it.

The multimodal codex is our attempt to imitate this clinical phenomenon in a machine learning multimodal fusion context.

The multimodal codex is a grayscale computer image that encodes within it a set of meaningful multimodal features representing human responses to a controlled experiment. A multimodal codex *sequence* is the series of multimodal codexes that together encode the *flow* of the test situation.

The multimodal codex is also a practical way to encode structured tabular data in a format that can more readily be taken advantage of by CNNs. CNNs are of practical interest because (1) they ditch the need for feature engineering as they create their own features and (2) they can be trained with relatively few learnable parameters, helping to prevent overfitting.

Converting tabular data sets to images in order to use CNNs on them has been exploited by several researchers recently. Alvi et al showed that tabular data on neonatal infections could be successfully exploited using a CNN by implementing a simple transformation where features (ie, columns) are assigned, one by one, to an X-Y coordinate, with their values becoming the pixel's intensity [16]. We will describe how we implemented their method in order to perform missing data imputation for our sample a few paragraphs below.

Buturović et al designed a tabular-data-to-graphical mapping in which each feature vector is treated as a kernel, which is then applied to an arbitrary base image [17]. Sun et al experimented using pretrained production-level CNN models implementing a diametrically opposite approach consisting of projecting the literal value of the features graphically onto an image; for example, if a feature has a value of 0.2 for a given participant in the sample, the image would include the actual number 0.2 on it [18].

The approach clearly closer to ours is that of DeepInsight [19]. Theirs is the realization that we can use a visualization technique, t-distributed stochastic neighbor embedding, in a different manner to what it was intended. While typically one applies the said technique on a data set in order to reduce the dimensions of the *feature space* to foster intuitive visualization of the sample distribution, they applied the method to their *transposed* data set, such that the *sample space* was reduced to a cartesian space for an intuitive understanding of the distribution of the *features*.

The approach we used for creating the multimodal codexes is similar, yet it differs from DeepInsight's approach in that we implement a more modern and reliable dimensionality reduction method, the Uniform Manifold Approximation and Projection (UMAP) [20]. Its strength is to better preserve the global structure of the data and thus the relationship between the features. In addition, we apply this procedure to a very specific kind of tabular data (multimodal sensing data). To the best of our knowledge, this has not been proposed before.

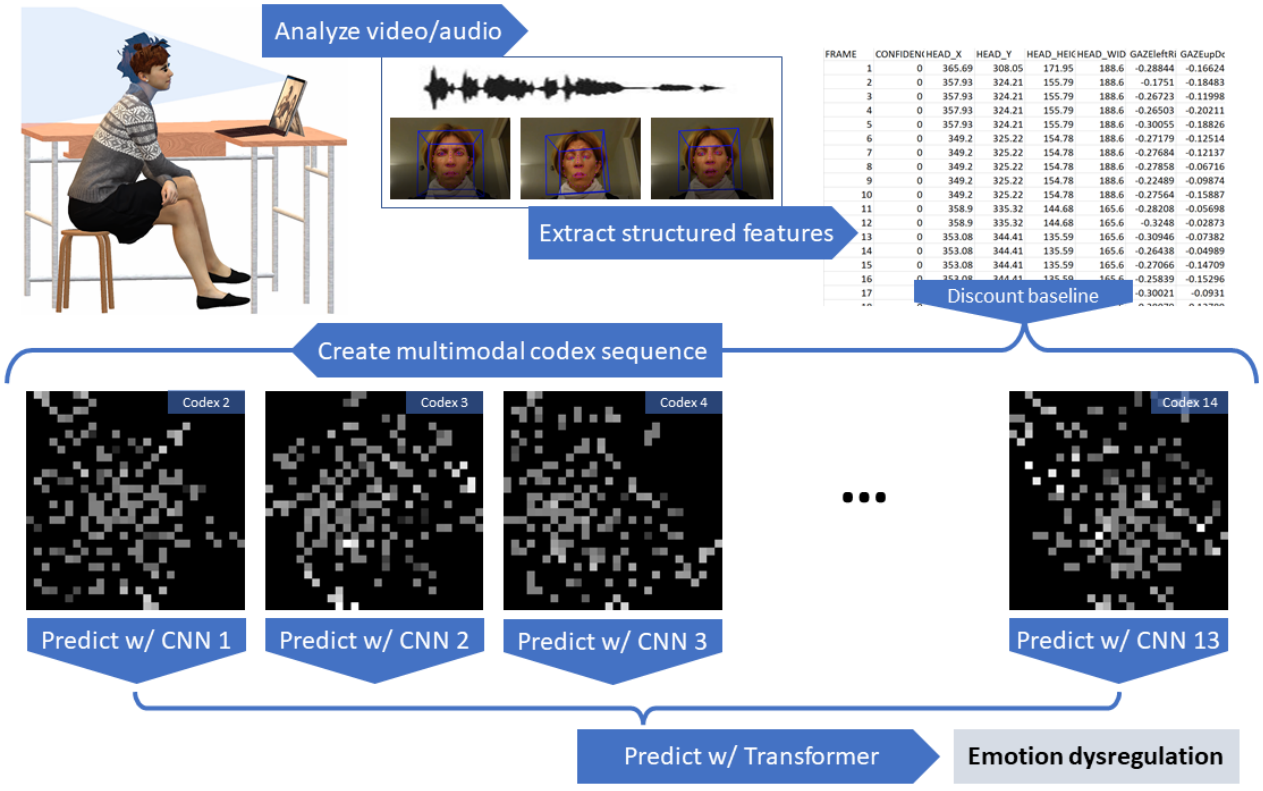
Our proposed method to missing data imputation can be described by the following pseudocode: *For each feature in the data set, (1) produce an image by disposing each feature vector in the dataset, EXCEPT the current one, as pixels in a grayscale image, with the intensity of the feature representing the pixel's intensity; (2) feed the created picture for each participant to a simple CNN consisting of 2 convolutional layers and a dense layer, the mission of which is to find visual patterns in the projected data that can predict the left-out feature; and (3) use the created model to predict the missing values corresponding to that feature.*

For each fold, we learn the missing data imputation models from the learning set and fill with it the missing values of both training and validation sets.

Our proposed process to create a multimodal codex sequence is resumed in the following pseudocode: *For each of the 13 stimuli, (1) group all features corresponding to a given stimulus in the form of a SAMPLES × FEATURES matrix; (2) use the UMAP method over the transposed matrix to obtain the X and Y coordinates for each feature; and (3) create a 28×28 pixel grayscale image per person, printing the value of each feature in their respective X and Y coordinates.*

The resultant images look like those in Figure 1.

Figure 1. From test to result. Top left: a woman taking the Multimodal Developmental Profile test. Top center: the audio wave and video frames, with the latter showing the analysis for head pose, eye gaze, and facial expressions. Top right: tabular data of some of the features extracted from the audio and video. Bottom: the 2nd, 3rd, 4th, and 14th multimodal codexes for a participant in the sample. CNN: convolutional neural network; w/: with.



This process naturally builds images with distinct clusters of features for each stimulus depending on the specific relationship between the typical responses to the said stimulus in the sample and the ground truth variable. Like a clinician's intuition described earlier, our approach could end clustering together a series of language markers, facial expressions, and HRV features, which might not initially be obvious, in the context of what is evoked by a specific stimulus and the typical response pattern in the sample.

Practically, this takes the guessing out of feature engineering, while also providing the CNNs with smaller clusters to "look at," which in turn puts less stringent requirements on the *receptive field* of the network, leading potentially to smaller kernels and fewer layers.

An important limitation of UMAP and all other visualization techniques of the sort is that the proximity of points in the projection they generate does not follow a predictable pattern. While points that are closer together typically are more related than those projected far away, this is not guaranteed for all cases, and the relationship between distance and importance is certainly not linear.

On occasion, the mapping for two or more features falls in the exact same X and Y coordinates. While this could be easily

remediated by enlarging the codex resolution, we decided to leave this as a feature. When UMAP considers 2 features to be so close, they might as well mean the exact same thing. In that case, we average the value of the features to find the value of the pixel in question.

For each fold, we learned the mapping from the learning set and created with it the multimodal codexes for the learning and validation sets.

Multimodal Fusion Network Architecture

As described in the previous section, the problem of assessing a psychological construct during an interview is both a spatial problem (ie, measuring different things that happen simultaneously) and a temporal problem (understanding the succession of events and their relationship).

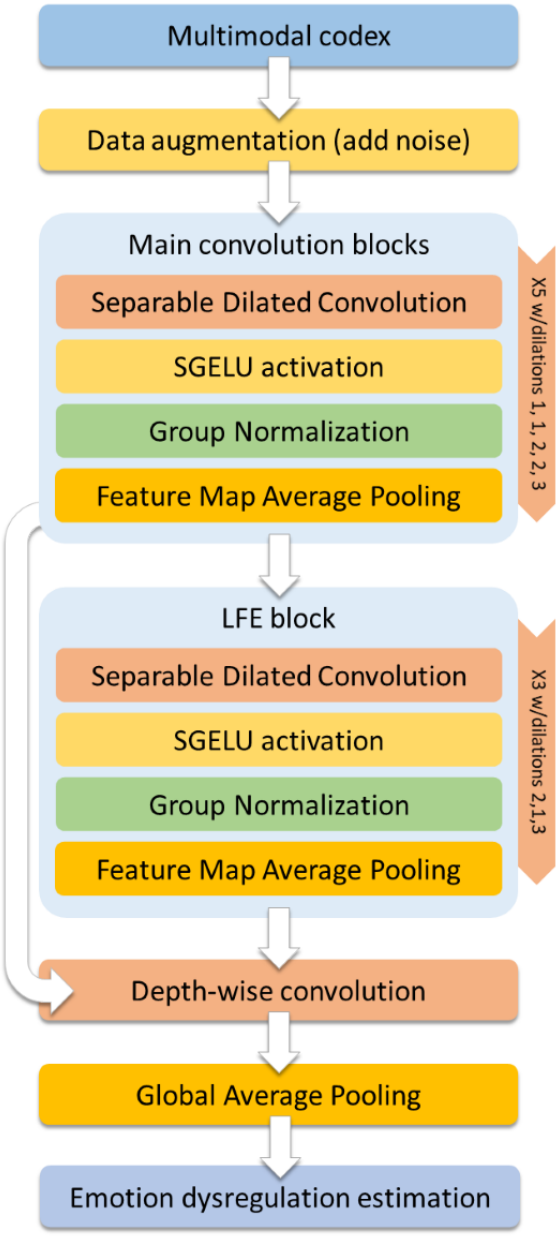
For dealing with the first part of the problem, we implemented 13 CNNs, with 1 per stimulus (minus the baseline stimulus). The reason not to rely on just 1 network for all of the stimuli is that we do not assume the features that are important to predict emotion dysregulation are the same during each stimulus response. On the contrary, a clinician will look for specific patterns in the patient's behaviors depending on the queue the therapist has sent right before during the interview. Patterns can actually reverse. A cluster of features indicative of emotion dysregulation given 1 stimulus can actually be indicative of good regulation during another.

We confronted the following challenges when designing the architecture for our CNNs: (1) How to create a deep enough network that will be able to extract complex concepts, while keeping the number of learnables (ie, weights) very lean to avoid overfitting (ie, memorizing) our small training set? (2) How to avoid downsampling/blurriness of the codex when going deeper into the network, a classic byproduct of pooling layers, so that deeper layers can still take advantage of details while simultaneously uncovering more global patterns? To overcome these challenges, we implemented cutting-edge best practices as well as some innovations.

The network begins with a multimodal codex augmentation layer that we will explore later. The rest of the network is basically constituted of 8 convolutional blocks, each containing a depth-wise separable convolution layer [21] with 8 3×3-sized kernels, with different dilation factors (more below), a stringent L1-L2 norm weight-decay regime, and a constrained range of values for the weights to take, lying between −1 and 1; a mean-shifted Symmetrical Gaussian Error Linear Units (SGELU) [22] activation layer; a group normalization layer [23]; and our new FMAP layer (details are presented in the next section). There is a residual connection that allows gradients to flow directly from the end of the network toward the output of the 5th convolutional block. After adding the residual and the upcoming connection from the last convolution block, the network ends with a depth-wise convolution layer (ie, kernel 1×1), a linear activation layer, and a Global Average Pooling (GAP) [24] layer. The whole CNN can be seen in Figure 2 (all 13 networks share identical architecture). It has only 339 weights overall.

Importantly, our proposed architecture dispenses with pooling layers entirely. They are typically used as a means to increase the effective receptive field when moving deeper into the network. They were replaced with a carefully calculated set of kernel dilation factors, which increase from the 1st block to the 5th, then decrease for blocks 6 and 7, and then increase once again in block 8 before the network ends. This decrease and increase between blocks 6 and 8 is what Hamaguchi et al have called a local feature extraction (LFE) module [25]. In their important work on satellite imagery, they have shown that in scenarios where both general patterns and details are important for prediction, reducing and then rapidly increasing the dilation factor can allow the network to take into account both detail *and* structure all the way to the deepest layers of the network. In our case, this is crucial, because although we trust the thinking behind the multimodal codex design, the UMAP method is not infallible, and a very important feature to predict emotion dysregulation might still end lying away (graphically) from the main clusters, as a single pixel somewhere in the image, that would tend to disappear when down-sampled. Different from the approach by Hamaguchi et al, though, we included a residual connection going from block 5 (right before entering the LFE module) directly into the last block, basically short-circuiting the LFE module. This allows our network to decide during training if the module is needed or not, depending on the actual data correlations it finds, and even to find the right balance of detail and structure automatically. The dilation factor of each convolutional layer was carefully calculated so that the *effective* receptive field covers the whole image (28×28) by the end of the network.

Figure 2. Our convolutional architecture (339 weights). LFE: local feature extraction; SGELU: Symmetrical Gaussian Error Linear Units.



In the following paragraphs, we provide a brief description of each of the components of the network as well as the rationale behind their implementation in the context of deep learning from small data sets.

Depth-wise separable convolutional layers were first introduced in a previous study by Chollet et al [21] and implemented in Google's Xception and MobileNet architectures. A depth-wise separable convolution separates the convolution process into the following 2 parts: a depth-wise convolution, and a pointwise convolution. They can allow for a reduction of parameters of up to 95% compared to classic convolutional layers [26]. While this reduction is typically desired from the perspective of lessening computational and size demands of neural networks, particularly during prediction time and for mobile hardware deployment, our rationale for using them is entirely different. In classical statistics, it is known that small samples should be fitted with models using relatively few degrees of freedom (ie, parameters) if one wants to prevent overfitting the training set. Typically, the best practice ratio is 10 to 1; ie, 10 times fewer degrees of freedom than data available. While that ideal might be too stringent when ported to modern machine learning, we still thought it was vital to keep it as a guiding principle. The fewer parameters we used, the least the network *could* overfit the data. Hence, our utilization of these layers.

SGELU activation was recently introduced in a previous study by Yu et al [22]. Yu et al took advantage of the already powerful GELU function, which represents nonlinearity by using the stochastic regularizer on an input (the cumulative distribution function derived from the Gaussian error function), which has shown several advantages over other activation functions and is currently implemented in modern natural language processing (NLP) transformer models. The new SGELU function allows activations to take on equally large negative and positive values, pushing the weights to also do so. In their investigation, they found that this new activation function performs better than all other available activation functions, but this was not the reason

that had us choose it for our task. Rather, they also reported that training becomes smoother and more stable when using SGELU and that they found preliminary evidence of better generalization of the network when trained with it. Since ours is a task that deals with a very small data set and thus probably exaggerated levels of variance, smoother more stable training can be crucial, and the capacity to generalize better could indicate greater self-regularization, which is essential when learning from a small sample.

Mean shifting [27] is a method that consists of simulating random data, similar to what an activation function might compute, and passing it through the activation function, in our case SGELU, to find the empirical mean of the activations. Once we find it, we can subtract it from 0, the desired mean for the activations, and then add (ie, shift) that difference to the activation itself. In so doing, now the empirical mean of the activation function becomes 0 (for random data). This approach has been shown to increase both convergence speed and accuracy.

Group normalization was introduced by the Facebook AI Research (FAIR) team in 2019 [23]. Its claim to fame was its capacity to produce performance results that paralleled batch normalization when using regularly sized batches, but that strongly outperformed it when using small batches. Small batches are more typical in the context of parallelization of neural networks training within computing clusters. Although we also got interested in it because of its capacity to deal with small batches, our reasoning was not computational. Instead, it has been shown that smaller batches increase regularization by, among other things, increasing stochasticity [28,29]. Importantly, we implemented group normalization *after* the SGELU activation functions for the following reason: as reported by [22], if activations are normalized *before* they hit the SGELU activation function, there is a risk that the full extent of it might not be used, particularly the nonlinear nature of both extremes of the function. We hard-coded the group norm hyperparameter, which decides the number of groups, to be always half of the number of kernels in the previous CNN layer (so 4 for all of our blocks).

The networks end with a GAP [24] layer to average the final activation map; the result of that operation is the prediction of the network. The GAP layer has come to replace fully connected layers in CNNs lately, mainly because of its capacity to reduce overfitting and drastically reduce parameters.

The full CNN model is shown in Figure 2.

After each of the 13 CNNs produce an estimation of emotion dysregulation, those estimations become the sequential data fed to the next and final architecture, to deal with the temporal aspect of our problem, which is the transformer.

Endowed with the task of decoding the sequential meaning of the participant's responses to the succession of MDP's controlled experiments, our transformer network is of course inspired by the seminal work of Vaswani and the team at Google Brain [30]. Transformers have replaced recurrent neural networks and their convolutional counterparts for an ever-increasing number of sequential learning tasks, including NLP, video classification, etc. Indeed, they can be trained faster than models based on recurrent or convolutional layers [30].

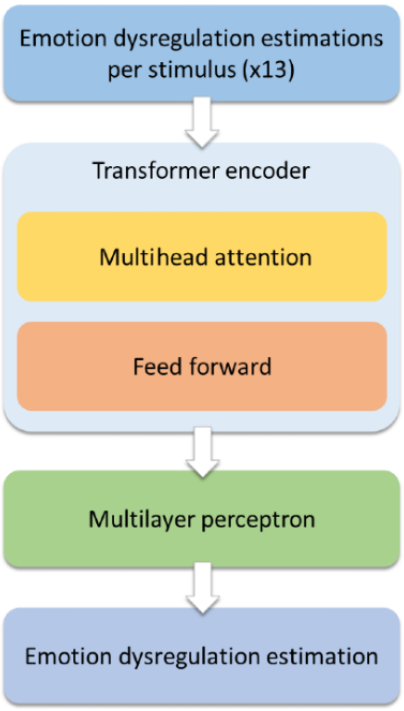
At their core is the multiheaded attention mechanism, which allows evaluating, in parallel and for each data point in a sequence, which other data points in the said sequence are relevant to the assessment. The attention heads in our encoder block are of size 13, to cover the whole MDP sequence, as opposed to the size of 64 used in the study by Vaswani et al, and we used 4 heads as opposed to 8. Our encoder block also includes residual connections, layer normalization, and dropout. The projection layers are implemented using a 1D convolution layer.

The encoder was followed by a 1D GAP layer to reduce the output tensor of the encoder to a vector of features for each data point in the current batch. Right after this is the multilayer perceptron regression head, consisting of a stack of fully connected layers with ReLU activation, followed by a final 1 neuron–sized fully connected layer with linear activation that produces the actual estimation of emotion dysregulation. We tried implementing positional encodings, as per the original paper, as well as look-ahead masking; however, both methods yielded worse results for our use case, so we discarded them.

In the original paper, Vaswani et al implemented label smoothing. Given that ours is a regression problem, we switched this for test-time augmentation (TTA), which will be described later.

The loss function for our transformer architecture was the concordance correlation coefficient (CCC) [31]. It was pioneered as a loss function by Atmaja et al, and tends to find a good balance of low error and high correlation between predictions and the ground truth [32]. Our transformer architecture can be seen in Figure 3.

Figure 3. Our transformer architecture (4223 weights).



FMAP Layer

This new kind of layer computes the average of the activations or feature maps produced by a 2D convolution layer as follows:

where *a* is a 3D "channels-last" tensor and *K* is the number of kernels of the previous convolution layer (ie, the number of channels).

It was inspired by the GAP layer, which revolutionized CNNs by drastically reducing the number of weights without sacrificing performance, while increasing regularization. However, the FMAP layer averages tensors among feature maps (ie, channels), as opposed to across the 2 dimensions of each feature map like GAP does.

If included at the end of every convolutional block, FMAP assures that the depth (ie, number of channels) of the activations flowing forward in the network remains flat (ie, 1 channel) at all depths of the network, instead of exponentially increasing, as is typically the case.

It is important to realize that a sort of weighted average *already* happens within regular convolutional layers when they calculate the dot product (ie, cross-correlations) between the kernel weights and the image pixels for each of its channels. By analogy, with FMAP, we are transforming that into a nonweighted average.

The FMAP can also be thought of as a nonlearnable version of the depth-wise convolution (ie, convolutions with kernel size 1×1 typically used to reduce the complexity of a model by merging its feature maps). By using a fixed function (average) instead of a learned one, though, we obtain a decrease in learnable weights in our model. For a depth-wise convolution, we need 1 weight and 1 bias per input feature map, whereas with FMAP, we need none. We also prevent the network from overfitting the training set during the computation.

In terms of the decrease in the number of weights for a network, in our own CNNs, the reduction is of 71% (from 1172 weights to 339). This remarkable reduction in weights has several effects, including reducing computational demands for both training and prediction, and, as we mentioned earlier, reducing the number of degrees of freedom in the model, thus reducing the potential to overfit the training set.

We believe this layer forces an ensembling effect onto the network's block in which it is inserted. It is a consensual observation that ensembles of trained neural networks generalize better than just 1 trained neural network [33]. This is because their different random initializations increase stochasticity, empowering each network in the ensemble to explore the loss landscape by taking entirely different paths toward minima, and when their predictions are averaged, they can cancel each other's overfitting tendencies out. We think that when FMAP layers are used consistently after all (or at least many) 2D convolutional layers, the same ensembling effect is introduced *within* subnetworks (ie, blocks) of the network, so that each block ending in an FMAP layer is forced to create an ensemble of subnetworks. This, we hypothesize, should introduce desirable block-wise stochasticity that increases model generalization ability without the need to train multiple entire neural networks.

Training and Test Time Data Augmentation Scheme

In our quest against overfitting, we implemented data augmentation. In its classic form, it allows for the on-the-fly creation of new training examples based on random transformations of the original ones.

With regard to our CNNs, we created a layer designed to introduce uniform random noise within the multimodal codexes.

During training, it introduces up to 10% noise for each pixel representing a feature in the multimodal codex (while it leaves all other pixels, the ones not representing any feature, alone). This meant that, for each epoch, the network saw an up to 10% different version of each image.

This procedure was especially important given that our uniformization of the ground truth variable by upsampling meant that there was a nonnegligible amount of image (multimodal codex) repetition being fed to the CNNs. So this data augmentation scheme allowed for them to be actually *somewhat* different.

Another more modern form of data augmentation is TTA [34]. This approach consists of, at prediction time, generating on the fly X-augmented data sets, predicting with each, and then averaging the results.

The way we implement TTA is innovative. We use it between our spatial (CNNs) and temporal (transformer) networks. When our 13 CNNs predict their final emotion dysregulation estimates, we do so using TTA, and moreover, we repeat the process 10 times. As a result, we provide the transformer with both better predictions and more diverse data to train on. We believe this procedure can greatly increase the generalization of the network to unseen data.

Training Procedure

We used vanilla Adam optimizer for both our CNNs and the transformer network, with default settings. We did not implement any learning rate scheduler.

We trained our CNNs for 500 epochs each. We trained our transformer network for 100 epochs. At each epoch, the models were saved. By the end of training, our code automatically selected the best model, which was the one with the highest Pearson correlation for our CNNs and that with the highest CCC for our transformer, between predictions and the ground truth on the validation set.

As we described earlier, all the aforementioned steps were implemented within each fold of a cross-validation procedure. Eight folds were utilized overall.

Analyses

Pearson correlation coefficient was calculated using SciPy, version 1.7.1 (Community Library Project). Mean absolute error and the CCC were assessed using Tensorflow, version 2.6.0 (Google Brain; code included in the associated Google Colab, see section below). Means and standard deviations were calculated using NumPy, version 1.19.5 (Community Project).

Convergent Validity Analysis and Interpretation Criteria

Convergent validity is the extent to which a measure produces results that are similar to other validated measures measuring the same construct [35]. A standard way of measuring it is by using Pearson product moment correlation [36]. We will interpret Pearson's results based on a review by Drummond et al on the best practices for interpreting validity coefficients, where a value ≥0.5 indicates very high correlation, 0.4 to 0.49 indicates high correlation, 0.21 to 0.4 indicates moderate correlation, and ≤0.2 indicates unacceptable correlation [37].

Replicability via Google Colab

We decided to port a large portion of our work from MATLAB to Tensorflow/Keras (created by François Chollet) and to prepare a Jupyter Notebook within Google Colab so that every reader can replicate our findings. The notebook can be accessed online [38]. It can be executed on Colab itself, or downloaded and run locally.

Results

The results are presented in Figure 4, Figure 5, and Table 1.

Figure 4. Scatter plot. Prediction (ie, estimation) vs Difficulties in Emotion Regulation Scale, brief version (DERS-16) for each fold. Pearson *r*, concordance correlation coefficient (CCC), and mean absolute error (MAE) are provided for each fold.

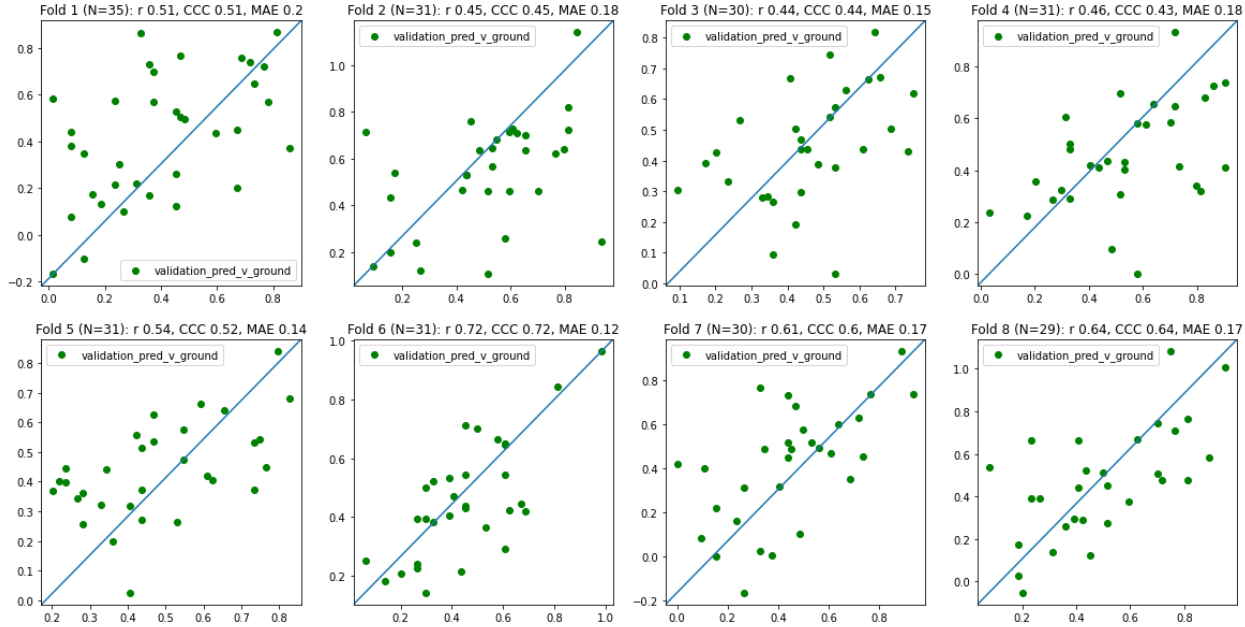


Figure 5. Eight folds' validation sets combined (N=248). Pearson *r*, concordance correlation coefficient (CCC), and mean absolute error (MAE) are provided for this combined sample.

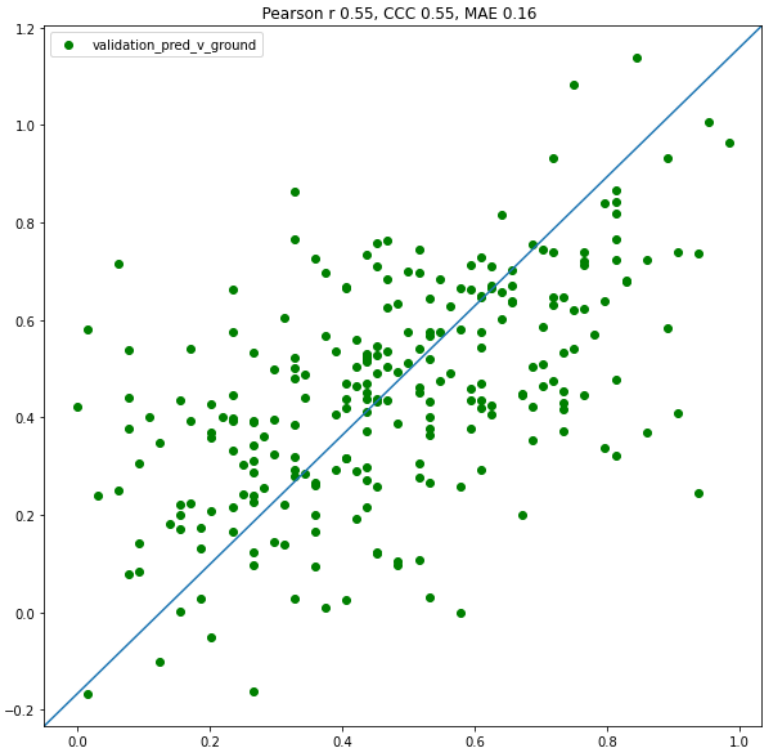


Table 1. Data per fold for our system's estimated emotion dysregulation versus the findings with the Difficulties in Emotion Regulation Scale, brief version (DERS-16; ground truth).

Variable	Number	Pearson *r*	*P* value	CCC[a]	MAE[b]
Fold					
1	35	0.51	.002	0.51	0.20
2	31	0.45	.01	0.45	0.18
3	30	0.44	.01	0.44	0.15
4	31	0.46	.01	0.43	0.18
5	31	0.54	.002	0.52	0.14
6	31	0.72	<.001	0.72	0.12
7	30	0.61	<.001	0.60	0.17
8	29	0.64	<.001	0.64	0.17
Mean value[c]	N/A[d]	0.55	<.001	0.54	0.16
SD value[e]	N/A	0.10	.01	0.10	0.02

[a]CCC: concordance correlation coefficient.

[b]MAE: mean absolute error.

[c]The mean across folds for each metric.

[d]N/A: not applicable.

[e]The mean of the standard deviations across folds for each metric.

Discussion

Principal Findings

Can computers detect emotion dysregulation in adults, by looking at their behavior and physiology during a set of controlled experiments? Can they generate "mental images" containing different sense modalities, like clinicians do? Can they do so in a sample that spans different cultures and languages? Can one train a deep multimodal fusion neural network using only a couple of thousand parameters? These are some of the questions we set out to answer in this work. This study evaluated the convergence validity of MDP's emotion dysregulation estimation with regard to DERS-16, a brief version

of the "gold standard" measure for emotion dysregulation. We interpret our results as excellent evidence for convergence validity between MDP's emotion dysregulation estimation and the DERS-16 in our sample, suggesting that scores obtained using the MDP are valid measures of emotion dysregulation in adults.

It is important to reflect on the diversity of our sample. It spanned 3 continents and 2 languages, with a broad age range, and included individuals with psychopathology to represent the higher end of the emotion dysregulation spectrum. With that in mind, we believe it is impressive that emotion dysregulation estimations were so correlated with their DERS-16 counterparts for all folds, showing similar results. We think this shows a preliminary form of cross-cultural validity for the approach, adding to the evidence we found in our prior work [13]. It also shows that the MDP is capable of assessing emotion dysregulation in adults with a psychopathology.

We think the multimodal codex approach captures quite well the mental processes that occur in the mind of a clinician while conducting an assessment interview. We attribute the success of our approach in large part to the good framing of the problem as spatiotemporal, and believe this representation of all sense modalities as a combined image is closer to the way we humans do multimodal fusion.

To our knowledge, the MDP is the first test of its kind. It is a validated exposure-based psychometric test that implements deep multimodal fusion to analyze responses within a set of controlled experiments in order to measure psychological constructs.

Its advantages over classical questionnaires and interview-based tests are manifold. They are as follows: the MDP takes less than 10 minutes to complete; it can be taken at home with a computer or tablet and is resilient to unpredictable variability in the test conditions; it is scored automatically in minutes; it is objective and replicable in its observations; it is holistic, taking into account language, voluntary and involuntary behavior, and physiology; it can be used in different cultures with only minimal translation efforts; and it can evolve over time, learning new scoring models based on different validated psychometric measures.

In terms of deep learning, we cannot stress enough how this work defies current trends and tenets within the field. In the current international race toward the trillion-parameter model, how can anyone dare to present a deep network capable of estimating very abstract psychological phenomena with only 8630 weights? In a field powered by Google, Apple, Facebook, Amazon, and other American and Asian tech giants data mining free online services for millions of data points, how can anyone dare to present a model that can be well trained with only 274 examples? We think this work should be seen as pertaining to a concurrent and perhaps literally opposite trend. Humans do not need that many examples to learn something, even something complex. Maybe machines do not need it either, provided intelligent constraints are put in place (sort of bike wheels for children) to prevent the system from falling into tendencies (memorization, ie, overfitting) that would prevent real learning. We think that at the heart of this concurrent view of machine learning, there is chaos in the form of randomness. Random noise has been added to our samples as data augmentation. There are random paths toward minima spearheaded by an increase in stochasticity due to small batches during training. There is randomness during prediction by implementing TTA. There is randomness in the random initialization of each kernel within each convolutional block, and the way the FMAP layers force them to ensemble. There is randomness in the automatic choice of the stimulus from the stimuli pool so that no single person experiences the exact same stimuli set. There is randomness in the random errors that occur in pretty much every one of the feature extraction processes implemented by the MDP software. Randomness might seem to be just noise, but what if, in reality, it is what allows us to separate signal from noise?

Limitations and Future Directions

One of the obvious limitations of our work is the size of our sample. Although we purposely set to prove that one can learn very complex and deep multimodal models that can be accurate and reliable with just a few hundred cases, this does not in any way disprove the common sense assumption that, with more data, the model would improve even more. In addition to sheer sample size, we believe it would be interesting, and quite unexplored in psychometry, to use census-based samples (data sets whose distribution in terms of sex, age, income, etc, matches the census of a given country). Online recruiting agencies are beginning to propose this as a service, and we hope we will be able to work with such a sample in the near future.

Another weak point of our study is the lack of a hold-out test set. We did not implement one primarily because of a lack of enough data. Indeed, it is known that validation sets can be overfitted, in a process some have called "model hacking" [39]. Model hacking is the extensive repetition of a cross-validation scheme for hyperparameter tuning and model development, for which we report only the best fit found. Similar to "human overfitting," our resulting model might obtain great cross-validation scores but perform more poorly in new unseen samples. This is especially true with brute-force approaches to hyperparameter tuning. Small-sized samples, such as ours, that contain high variability and an extremely diverse population are somewhat inherently protected against model hacking. Each fold's validation set will be strongly different from that of another fold, not to mention that training samples themselves will be very different from fold to fold, producing quite different models. If with such variability the model still shows stable performance across all or most folds, it might be a good indication that the methodology and the models resulting from it do generalize well. In addition, we took some empirical measures to prevent model hacking, such as having a random seed set at the beginning of our code, so that the partition of folds was always equal, and then working with the first fold for hyperparameter tuning and model tuning. Most importantly, we have not implemented any sort of automatic search algorithm for hyperparameter tuning. Instead, we chose to explore only a handful of theoretically promising options by hand.

Furthermore, we question whether a hold-out sample, proportional in size to our overall sample, would have been a

better unbiased estimator (how can a sample with a size of around 30 be taken as representative of the whole population?). In the future, we will look to the works of Martin and Corneanu [40,41] that unlock estimating generalization performance directly from the characteristics of the model itself. We are already working on a criterion inspired by their work, which we call the network engagement criterion. This criterion seems promising in estimating test error using only the training sample. Such a method would, in our opinion, close the circle, completing the set of methods and approaches we presented in this work to fully implement a cycle of unbiased learning with the sort of "small data" samples commonly found in the social sciences.

Conclusion

In this work, we successfully trained a deep neural network consisting of spatial (convolutional) and sequential (transformer) submodels, to estimate emotion dysregulation in adults. Remarkably, we were able to do so with only a small sample of 248 participants, without using transfer learning. The metrics of performance we used show not only that the network seems to generalize well, but also that its correlation with the "gold standard" DERS-16 questionnaire is such that our system is a promising alternative. Perhaps most importantly, it was confirmed that deep learning does not need to mean millions of parameters or even millions of training examples. Carefully designed experiments, diverse small data, and careful design choices that increase self-regularization might be sufficient.

Acknowledgments

We want to thank Gwenaëlle Persiaux for her recruiting efforts in Lyon, France; Nahed Boukadida for her recruiting efforts in Tunisia; Susana Tereno, Carole Marchand, Eva Hanras, and Clara Falala-Séchet for their recruiting efforts in Paris, France; and Khalid Kalalou and Dominique Januel for their recruiting efforts at Etablissement Public De Santé Ville-Evrard in Saint-Denis, France. Funding for this publication (fees) was provided by FP and the University of Bourgogne Franche-Compte.

Authors' Contributions

FP handled project funding, training scheme, network design, multimodal codex development, coding, and recruitment at Paris and the United States. YB handled remote photoplethysmography algorithm development, recruitment at Dijon, and academic review. FY handled recruitment at Dijon and academic review.

Conflicts of Interest

None declared.

References

1. Gratz K, Roemer L. Multidimensional Assessment of Emotion Regulation and Dysregulation: Development, Factor Structure, and Initial Validation of the Difficulties in Emotion Regulation Scale. Journal of Psychopathology and Behavioral Assessment 2004 Mar;26(1):41-54 [FREE Full text] [doi: 10.1023/b:joba.0000007455.08539.94]
2. Beauchaine T. Vagal tone, development, and Gray's motivational theory: toward an integrated model of autonomic nervous system functioning in psychopathology. Dev Psychopathol 2001;13(2):183-214 [FREE Full text] [doi: 10.1017/s0954579401002012] [Medline: 11393643]
3. Hayes SC, Wilson K, Gifford E, Follette V, Strosahl K. Experiential avoidance and behavioral disorders: A functional dimensional approach to diagnosis and treatment. Journal of Consulting and Clinical Psychology 1996 Dec;64(6):1152-1168 [FREE Full text] [doi: 10.1037/0022-006x.64.6.1152]
4. Mennin DS, Heimberg R, Turk C, Fresco D. Applying an emotion regulation framework to integrative approaches to generalized anxiety disorder. Clinical Psychology: Science and Practice 2002;9(1):85-90 [FREE Full text] [doi: 10.1093/clipsy.9.1.85]
5. Parra F, George C, Kalalou K, Januel D. Ideal Parent Figure method in the treatment of complex posttraumatic stress disorder related to childhood trauma: a pilot study. Eur J Psychotraumatol 2017;8(1):1400879 [FREE Full text] [doi: 10.1080/20008198.2017.1400879] [Medline: 29201286]
6. Linehan MM. Cognitive-Behavioral Treatment of Borderline Personality Disorder. New York, NY, USA: Guilford Press; 1993.
7. Bjureberg J, Ljótsson B, Tull MT, Hedman E, Sahlin H, Lundh L, et al. Development and Validation of a Brief Version of the Difficulties in Emotion Regulation Scale: The DERS-16. J Psychopathol Behav Assess 2016 Jun 14;38(2):284-296 [FREE Full text] [doi: 10.1007/s10862-015-9514-x] [Medline: 27239096]
8. Vasilev CA, Crowell S, Beauchaine T, Mead H, Gatzke-Kopp L. Correspondence between physiological and self-report measures of emotion dysregulation: a longitudinal investigation of youth with and without psychopathology. J Child Psychol Psychiatry 2009 Nov;50(11):1357-1364. [doi: 10.1111/j.1469-7610.2009.02172.x] [Medline: 19811585]
9. Hickey BA, Chalmers T, Newton P, Lin C, Sibbritt D, McLachlan CS, et al. Smart Devices and Wearable Technologies to Detect and Monitor Mental Health Conditions and Stress: A Systematic Review. Sensors (Basel) 2021 May 16;21(10):3461 [FREE Full text] [doi: 10.3390/s21103461] [Medline: 34065620]

10. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 2017 Sep;37:98-125 [FREE Full text] [doi: 10.1016/j.inffus.2017.02.003]
11. Parra F, Miljkovitch R, Persiaux G, Morales M, Scherer S. The Multimodal Assessment of Adult Attachment Security: Developing the Biometric Attachment Test. J Med Internet Res 2017 Apr 06;19(4):e100 [FREE Full text] [doi: 10.2196/jmir.6898] [Medline: 28385683]
12. Murray HA. Thematic Apperception Test Manual. Cambridge, MA, USA: Harvard University Press; 1943.
13. Parra F, Scherer S, Benezeth Y, Tsvetanova P, Tereno S. (revised May 2019) Development and cross-cultural evaluation of a scoring algorithm for the Biometric Attachment Test: Overcoming the challenges of multimodal fusion with "small data". IEEE Trans. Affective Comput 2019:1-1 [FREE Full text] [doi: 10.1109/taffc.2019.2921311]
14. Rutter M, Sroufe L. Developmental psychopathology: concepts and challenges. Dev Psychopathol 2000;12(3):265-296 [FREE Full text] [doi: 10.1017/s0954579400003023] [Medline: 11014739]
15. Bleidorn W, Hopwood C. Using Machine Learning to Advance Personality Assessment and Theory. Pers Soc Psychol Rev 2019 May;23(2):190-203 [FREE Full text] [doi: 10.1177/1088868318772990] [Medline: 29792115]
16. Alvi RH, Rahman M, Khan A, Rahman R. Deep learning approach on tabular data to predict early-onset neonatal sepsis. Journal of Information and Telecommunication 2020 Dec 25;5(2):226-246 [FREE Full text] [doi: 10.1080/24751839.2020.1843121]
17. Buturović L, Miljković D. A novel method for classification of tabular data using convolutional neural networks. bioRxiv. URL: https://www.biorxiv.org/content/10.1101/2020.05.02.074203v1 [accessed 2022-01-02]
18. Sun B, Yang L, Zhang W, Lin M, Dong P, Young C, et al. SuperTML: Two-Dimensional Word Embedding for the Precognition on Structured Tabular Data. 2019 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 16-17, 2019; Long Beach, CA, USA p. 2973-2981. [doi: 10.1109/cvprw.2019.00360]
19. Sharma A, Vans E, Shigemizu D, Boroevich K, Tsunoda T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. Sci Rep 2019 Aug 06;9(1):11399 [FREE Full text] [doi: 10.1038/s41598-019-47765-6] [Medline: 31388036]
20. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. JOSS 2018 Sep;3(29):861 [FREE Full text] [doi: 10.21105/joss.00861]
21. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI, USA p. 1800-1807. [doi: 10.1109/cvpr.2017.195]
22. Yu C, Su Z. Symmetrical Gaussian Error Linear Units (SGELUs). arXiv. 2019. URL: https://arxiv.org/abs/1911.03925 [accessed 2022-01-02]
23. Wu Y, He K. Group Normalization. Int J Comput Vis 2019 Jul 22;128(3):742-755 [FREE Full text] [doi: 10.1007/s11263-019-01198-w]
24. Lin M, Chen Q, Yan S. Network In Network. arXiv. 2014. URL: https://arxiv.org/abs/1312.4400 [accessed 2022-01-02]
25. Hamaguchi R, Fujita A, Nemoto K, Imaizumi T, Hikosaka S. Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery. 2018 Presented at: IEEE Winter Conference on Applications of Computer Vision (WACV); March 12-15, 2018; Lake Tahoe, NV, USA. [doi: 10.1109/wacv.2018.00162]
26. Wang CF. A Basic Introduction to Separable Convolutions. Towards Data Science. 2018. URL: https://towardsdatascience.com/a-basic-introduction-to-separable-convolutions-b99ec3102728 [accessed 2022-01-02]
27. Wright L. Comparison of new activation functions for deep learning. Results favor FTSwishPlus. Medium. 2019. URL: https://lessw.medium.com/comparison-of-activation-functions-for-deep-learning-initial-winner-ftswish-f13e2621847 [accessed 2022-01-02]
28. Martin CH, Mahoney MW. Traditional and Heavy-Tailed Self Regularization in Neural Network Models. arXiv. 2019. URL: https://arxiv.org/abs/1901.08276 [accessed 2022-01-02]
29. Jiang Y, Nagarajan V, Baek C, Kolter JZ. Assessing Generalization of SGD via Disagreement. arXiv. 2021. URL: https://arxiv.org/abs/2106.13799 [accessed 2022-01-02]
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA, USA p. 6000-6010. [doi: 10.5555/3295222.3295349]
31. Lin L. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 1989 Mar;45(1):255 [FREE Full text] [doi: 10.2307/2532051]
32. Atmaja BT, Akagi M. Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition. J Phys Conf Ser 2021 Apr 01;1896(1):012004 [FREE Full text] [doi: 10.1088/1742-6596/1896/1/012004]
33. Fort S, Hu H, Lakshminarayanan B. Deep Ensembles: A Loss Landscape Perspective. arXiv. 2019. URL: https://arxiv.org/abs/1912.02757 [accessed 2022-01-02]
34. Moshkov N, Mathe B, Kertesz-Farkas A, Hollandi R, Horvath P. Test-time augmentation for deep learning-based cell segmentation on microscopy images. Sci Rep 2020 Mar 19;10(1):5068 [FREE Full text] [doi: 10.1038/s41598-020-61808-3] [Medline: 32193485]

35. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. Front Public Health 2018 Jun 11;6:149 [FREE Full text] [doi: 10.3389/fpubh.2018.00149] [Medline: 29942800]
36. Swank JM, Mullen P. Evaluating Evidence for Conceptually Related Constructs Using Bivariate Correlations. Measurement and Evaluation in Counseling and Development 2017 Oct 04;50(4):270-274 [FREE Full text] [doi: 10.1080/07481756.2017.1339562]
37. Drummond RJ, Sheperis CJ, Jones KD. Assessment Procedures for Counselors and Helping Professionals. London, United Kingdom: Pearson; 2016.
38. Jupyter Notebook. Colab Research Google. URL: https://colab.research.google.com/drive/1Pz2RlzYrIjTqmz0lmyxU3C4j7CoFVCs2?usp=sharing [accessed 2022-01-02]
39. Orrù G, Monaro M, Conversano C, Gemignani A, Sartori G. Machine Learning in Psychometrics and Psychological Research. Front Psychol 2019;10:2970 [FREE Full text] [doi: 10.3389/fpsyg.2019.02970] [Medline: 31998200]
40. Martin CH, Peng T, Mahoney M. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nat Commun 2021 Jul 05;12(1):4122-4021 [FREE Full text] [doi: 10.1038/s41467-021-24025-8] [Medline: 34226555]
41. Corneanu C, Madadi M, Escalera S, Martinez A. Explainable Early Stopping for Action Unit Recognition. 2020 Presented at: 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020); November 16-20, 2020; Buenos Aires, Argentina. [doi: 10.1109/fg47880.2020.00080]

Abbreviations

BAT: Biometric Attachment Test
CCC: concordance correlation coefficient
CNN: convolutional neural network
DERS: Difficulties in Emotional Regulation Scale
FMAP: Feature Map Average Pooling
GAP: Global Average Pooling
HRV: heart rate variability
LFE: local feature extraction
MDP: Multimodal Developmental Profile
NLP: natural language processing
RPPG: remote photoplethysmography
SGELU: Symmetrical Gaussian Error Linear Units
TTA: test-time augmentation
UMAP: Uniform Manifold Approximation and Projection

Edited by G Eysenbach; submitted 19.10.21; peer-reviewed by Z Ni, V Verma; comments to author 08.11.21; revised version received 10.11.21; accepted 23.11.21; published 24.01.22.

Please cite as:
Parra F, Benezeth Y, Yang F
Automatic Assessment of Emotion Dysregulation in American, French, and Tunisian Adults and New Developments in Deep Multimodal Fusion: Cross-sectional Study
JMIR Ment Health 2022;9(1):e34333
URL: https://mental.jmir.org/2022/1/e34333
doi:10.2196/34333
PMID:35072643

Original Paper

FOCUS mHealth Intervention for Veterans With Serious Mental Illness in an Outpatient Department of Veterans Affairs Setting: Feasibility, Acceptability, and Usability Study

Benjamin Buck[1], PhD; Janelle Nguyen[2], BA; Shelan Porter[2], BA; Dror Ben-Zeev[1], PhD; Greg M Reger[1,2], PhD

[1]Behavioral Research in Technology and Engineering (BRiTE) Center, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, United States

[2]VA Puget Sound Healthcare System, Seattle, WA, United States

Corresponding Author:
Benjamin Buck, PhD
Behavioral Research in Technology and Engineering (BRiTE) Center
Department of Psychiatry and Behavioral Sciences
University of Washington
1959 NE Pacific Street
Seattle, WA, 98195
United States
Phone: 1 206 221 8518
Fax: 1 206 543 9520
Email: buckbe@uw.edu

Abstract

Background: Veterans with serious mental illnesses (SMIs) face barriers to accessing in-person evidence-based interventions that improve illness management. Mobile health (mHealth) has been demonstrated to be feasible, acceptable, effective, and engaging among individuals with SMIs in community mental health settings. mHealth for SMIs has not been tested within the Department of Veterans Affairs (VA).

Objective: This study examines the feasibility, acceptability, and preliminary effectiveness of an mHealth intervention for SMI in the context of VA outpatient care.

Methods: A total of 17 veterans with SMIs were enrolled in a 1-month pilot trial of FOCUS, a smartphone-based self-management intervention for SMI. At baseline and posttest, they completed measures examining symptoms and functional recovery. The participants provided qualitative feedback related to the usability and acceptability of the intervention.

Results: Veterans completed on an average of 85.0 (SD 96.1) interactions with FOCUS over the 1-month intervention period. They reported high satisfaction, usability, and acceptability, with nearly all participants (16/17, 94%) reporting that they would recommend the intervention to a fellow veteran. Clinicians consistently reported finding mHealth-related updates useful for informing their care. Qualitative feedback indicated that veterans thought mHealth complemented their existing VA services well and described potential opportunities to adapt FOCUS to specific subpopulations (eg, combat veterans) as well as specific delivery modalities (eg, groups). In the 1-month period, the participants experienced small improvements in self-assessed recovery, auditory hallucinations, and quality of life.

Conclusions: The FOCUS mHealth intervention is feasible, acceptable, and usable among veterans. Future work should develop and examine VA-specific implementation approaches of FOCUS for this population.

(JMIR Ment Health 2022;9(1):e26049) doi:10.2196/26049

KEYWORDS

mHealth; veterans; schizophrenia; serious mental illness; mobile phone

Introduction

Background

Serious mental illnesses (SMIs), including schizophrenia, bipolar disorder, and major depression, are associated with disruption of typical social [1] and vocational functioning [2], homelessness [3], and even premature death [4,5]. However, a significant portion of individuals with SMIs recover and enjoy long, productive, and meaningful lives [6]. A critical determinant of recovery is the capacity for symptom self-management, or coping with the illness to mitigate its negative effects. A growing body of evidence supports the effectiveness of self-management interventions for individuals with SMIs [7,8]. These interventions, which provide support and resources to facilitate coping skills and medication adherence, are associated with reductions in symptoms and risk of hospitalization, as well as increased recovery and quality of life [9]. The Department of Veterans Affairs (VA)—the nation's largest integrated health care provider—has emerged as a leader in psychosocial rehabilitation for people with SMIs [10]. VA services, which include primary care, hospital medicine, and a comprehensive collection of specialty services, reach >9 million enrolled veterans each year [11]. SMI is overrepresented in VA health care settings relative to the general population [12], and veterans with SMIs are at increased risk of negative outcomes relative to other mental disorders [13]. Specifically, veterans with SMIs are at increased risk of comorbid chronic pain conditions [14], obesity [15], and undiagnosed and untreated trauma-related symptoms [16]. The constellation of medical and psychiatric complications associated with SMIs results in these individuals losing on average >14 years of life relative to the average [17].

Several barriers limit the reach and effectiveness of self-management interventions even among veterans receiving care in integrated health care systems such as the VA. First, veterans with SMIs face many challenges with care, including transportation, recall of appointment times, and the impact of personal crises on access to services [18]. Research suggests that very few individuals with SMIs receive specialized evidence-based psychosocial care for SMIs [19], and veterans living farther from VA health care facilities have poorer use [20]. Second, even when resources are available, veterans with SMIs are susceptible to disengagement. In total, 2 studies examining veterans with SMIs being treated at the VA found that, respectively, 42% and 47% of individuals with SMIs receiving care within the VA experienced a service gap of at least a year [20,21]. Third, even when individuals have access to and motivation to engage in care, typical in-person services are provided weekly or monthly. Self-management is most effective when it is activated immediately in response to stressors.

Recent developments in web-based and mobile technology have the potential to economically expand the reach and effectiveness of self-management interventions [22]. Individuals with SMIs report similar access to technologies as the general population [23] as well as an interest in the use of these technologies for mental health support [24]. A mobile health (mHealth) intervention for individuals with SMIs—FOCUS—has demonstrated usability among individuals with SMIs [25] and feasibility in community mental health settings [26]. A recent randomized trial comparing FOCUS with an evidence-based in-clinic group intervention for symptom self-management demonstrated comparable positive clinical effects between the 2 interventions, and those randomized into FOCUS remained engaged at higher rates than those randomized into typical in-person care [27].

The VA has also demonstrated innovations in the deployment of mHealth for mental health. A recent meta-analysis [28] identified 20 mental health or addiction mobile apps developed by the VA or the Department of Defense. Although these apps cover a variety of clinical interventions (eg, cognitive behavioral therapy [CBT] for insomnia; *CBTi Coach*), self-management activities (eg, tracking; *T2 Mood Tracker*), or diagnoses (eg, posttraumatic stress disorder [PTSD]; *PTSD Coach*), few (eg, *Virtual Hope Box* and *PTSD Coach*) have been tested in randomized trials and demonstrated significant improvements relative to waitlist [29] or usual care control conditions [30]. Many veterans report openness to using digital interventions for managing mental health [31], and over half of veterans receiving care for PTSD with access to digital technologies report interest in using mHealth for a range of clinical issues [32], although knowledge of available mHealth options remains a barrier to broad uptake among veterans.

There is a lack of mHealth tools designed for SMIs available through the VA. Of the apps currently featured on the VA mobile app website, none provide content specifically designed for the management of psychosis [33]. Although early work examining mHealth for SMIs has demonstrated its feasibility and effectiveness, there may exist specific features relevant to the deployment of these tools for veterans or within VA health care settings. Veterans with schizophrenia often present with comorbid chronic pain [14], other chronic medical conditions (eg, hypertension or diabetes) [34,35], or PTSD [36], which, when co-occurring with schizophrenia, increases the risk of suicide [37]. Veterans and active duty service members with mental illnesses also appear particularly susceptible to stigma associated with mental illness [38], which could affect their willingness to engage in clinical services at brick-and-mortar facilities. Insights gleaned from the deployment of technological innovations in community settings may not generalize to VA settings given specific institutional structures and clinical workflows [39]. Taken together, these risk factors suggest a need for research that examines the feasibility and acceptability of mHealth among veterans with SMIs receiving outpatient care from a VA facility.

Objective

This study reports the results of a pilot feasibility study of FOCUS deployed in a VA outpatient clinic for individuals with SMIs (ie, a Psychosocial Rehabilitation and Recovery Center [PRRC]). This clinic provides access to ongoing group therapies, individual therapy and case management, medication management, and the option to access related VA services, including vocational support. The results aim to determine whether mHealth is (1) feasible to deploy in a VA setting and (2) acceptable to veterans with SMIs, as well as to explore the

preliminary effectiveness of this intervention among veterans with SMIs and determine whether the participants' qualitative feedback suggests changes that would make mHealth for SMIs more appropriate and effective for the VA setting or the veteran population.

Methods

Participants

The study was reviewed and approved by the VA Puget Sound Health Care System Institutional Review Board. The participants were 17 individuals receiving treatment from an outpatient psychosocial rehabilitation clinic in a VA hospital in the Pacific Northwest. Potential participants were eligible for the study if they (1) had a serious and chronic mental illness (eg, schizophrenia-spectrum or mood disorder) with (2) current or past psychotic symptoms and (3) received their services at the PRRC. They were excluded if they (1) were incapable of providing informed consent or (2) had hearing, vision, or motor impairments that made it impossible for them to use a smartphone. Clinicians first shared information about the study with prospective participants and assessed their potential interest. With veteran authorization, study clinicians provided these referrals to the research team, who then contacted participants by phone to schedule their first visit.

FOCUS mHealth Intervention

FOCUS comprises 3 components: a mobile app, a clinician dashboard, and an mHealth support specialist. The FOCUS mobile app includes brief, preprogrammed self-management interventions that can be accessed by the user on demand. Participants can do this in two ways: (1) on demand completing a brief ecological momentary assessment (EMA) item that provides them with a tailored intervention (if they indicate distress) or (2) via the *toolbox*, which provides users with access to specific skill practices without tailoring assessment. Self-management interventions are also accessed via prompts that remind participants to use FOCUS (a device notification that reads *Would you like to check in with FOCUS?*). On the basis of their responses to the EMA items, FOCUS delivers tailored in-the-moment interventions. For example, if a participant responds to a prompted assessment by selecting the option that they are bothered by the thought that their voices *know everything*, the system provides an example of a mental exercise designed to challenge the validity of that belief. These notifications are automatically deployed 3 times per day. Intervention categories include voices (cognitive and behavioral strategies to cope with auditory hallucinations), mood (behavioral activation and other cognitive exercises), sleep (sleep hygiene psychoeducation and relaxation exercises), social functioning (cognitive exercises for persecutory ideation, anger management, and social skill training), and medication use (reminders, behavioral tailoring, and psychoeducation). For the duration of the study, the FOCUS system prompted within 3 time frames daily (9 AM-1 PM, 1 PM-5 PM, and 5 PM-9 PM; exact times within those ranges were determined randomly by the system each day).

All participant use of the system was logged on the web-based clinician dashboard, which was reviewed at least weekly by the mHealth support specialist, a member of the research team tasked with tracking and supporting participant use of FOCUS and providing relevant updates to the VA mental health treatment team [40]. On weekly calls with each participant, mHealth support specialists were tasked with (1) providing technical support in case of app issues and (2) encouraging the personalized use of FOCUS skills for participants' specific concerns. These calls were designed to last between 5 and 15 minutes. In this study, the mHealth support specialist also attended weekly meetings with the psychosocial rehabilitation mental health treatment team, providing brief (ie, <1 minute) updates related to each veteran enrolled in the study including an overview of (1) their use of FOCUS, (2) their responses to FOCUS items (ie, indicating symptoms and functioning), and (3) skills and support provided during weekly mHealth calls. This ensured that the members of the clinical team were aware of progress and relevant clinical changes to inform ongoing standard treatment. The mHealth support specialist was also available as needed to the primary treatment team to answer questions about FOCUS functions and content.

Procedure

At the baseline visit, the participants were provided with a detailed overview of the study, were given the opportunity to ask questions, and provided written informed consent after completing a brief competency questionnaire. After providing consent, the participants completed baseline study assessments (described below) and then received an orientation to FOCUS. The participants were given the opportunity to use their own personal device if they had one that was compatible with FOCUS (ie, an Android device) and were lent a study device if they did not. If necessary for those using a loaned study device, the orientation also included instructions on the use of the device, for example, operations such as turning the phone on or off, how to use the touchscreen, or how to place phone calls. FOCUS notifications (ie, the daily reminders) prompted the participants to complete assessments and receive interventions tailored to the goals individually set at baseline related to areas that they identified as being relevant to their recovery. At posttest visits, the participants returned the study device (if necessary) and again completed the same battery of assessments in addition to assessments related to the usability of FOCUS and a brief semistructured interview soliciting qualitative feedback. The participants were compensated with US $40 for each of the 2 study visits.

Measures

The participants completed a modified version of the System Usability Scale (SUS) based on previous work examining the feasibility and acceptability of FOCUS [26] to assess acceptability and feasibility. In addition to the conventional 26 items, we included items that assessed whether FOCUS required adaptation for a veteran population (eg, *FOCUS is appropriate for use with veterans* or *FOCUS was well integrated into my usual care at the VA PRRC*). We administered brief questionnaires to members of the primary clinician team when a client on their caseload was involved in FOCUS to assess the feasibility and acceptability of weekly updates to the clinical team, asking (1) whether they found FOCUS updates useful

and (2) whether those updates affected their clinical care. Following the study assessment battery, the participants also responded to open-ended questions requiring them to expand upon their experience with the intervention. We reported on responses to the following items: (1) *What did you like about the app?* and (2) *What did you not like about the app?* to assess intervention acceptability and usability. For items regarding fit and adaptation to veterans, we reported on (1) *Would you recommend the app to a fellow veteran? Why or why not?* and (2) *What are ways this app could be improved for use specifically with veterans?* This interview was conducted face-to-face at the VA medical center in a private setting by a trainee clinical psychologist or a research study coordinator. Responses to each item were recorded by hand by the study coordinator.

A total of six different clinical or functional outcomes were assessed: depressive symptoms, auditory verbal hallucinations, persecutory ideation, insomnia, quality of life, and overall recovery. Depressive symptoms were assessed using the Beck Depression Inventory–Second Edition [41], a 21-item assessment of ranging symptoms of depression that is summed for an overall score. Auditory verbal hallucinations were assessed using the Hamilton Program for Schizophrenia Voices Questionnaire [42], a 13-item self-report questionnaire that assesses the frequency and severity of one's experience of auditory verbal hallucinations within the past week. The Green Paranoid Thoughts Scale [43], a 32-item questionnaire covering thoughts about intentional threats from others, provided an assessment of persecutory ideation. Sleep quality was assessed using the Insomnia Severity Index [44], a 7-item scale assessing the extent and severity of current insomnia as well as satisfaction with one's current sleep routine. Quality of life was assessed using the Quality of Life Enjoyment and Satisfaction Questionnaire [45,46], an 18-item assessment of satisfaction in various areas of one's life, including social connections, work, and leisure. Finally, recovery was assessed using the Illness Management and Recovery Scale [47], a 15-item assessment of self-management and recovery developed to be consistent with the theoretical guidelines underlying Illness Management and Recovery [48], an evidence-based treatment program focused on independent, self-directed recovery.

Data Analytic Plan

We first examined descriptive statistics among all participants on the SUS to examine the acceptability, usability, and satisfaction among veterans using the intervention. We then examined the qualitative responses to the postintervention interview prompts. In total, 2 raters (BB and JLN) reviewed all interview responses and independently created proposed response categories that unified a particular idea to analyze the participants' perspectives on the open-ended items related to the FOCUS app. Units were defined as the collection of all words in a statement that conveyed a single idea or attribute. All disagreements were reconciled through discussion between the coders.

We reported pre–post descriptive statistics and effect sizes to examine the preliminary effectiveness of FOCUS among veterans participating in psychosocial rehabilitation. Although not powered for statistical significance testing, we conducted a series of paired sample 2-tailed *t* tests to explore whether during the 1-month study period the participants experienced improvements in depressive symptoms, auditory verbal hallucinations, persecutory ideation, sleep quality, self-reported quality of life, and self-reported recovery.

Results

Demographics

Participant characteristics are reported in Table 1. Our sample was predominantly White (11/17, 65%), male (12/17, 71%), and never married (9/17, 53%); reported a high school diploma (8/17, 47%) or associate's degree (6/17, 35%) as the highest educational level; and had experienced between 1 and 5 psychiatric hospitalizations (10/17, 59%). Although the inclusion criteria encompassed a mood or schizophrenia-spectrum disorder with current or past psychotic symptoms, multiple participants reported a comorbid diagnosis of PTSD (6/17, 35%). Other frequent diagnoses were schizophrenia (4/17, 24%), schizoaffective disorder (5/17, 29%), and major depressive disorder (6/17, 35%). The participants' average age was 55.12 (SD 13.02) years.

Table 1. Demographic characteristics of the study participants (N=17).

Characteristic	Values
Age (years), mean (SD)	55.12 (13.02)
Gender, n (%)	
Female	5 (29)
Male	12 (71)
Diagnosis, n (%)	
PTSD[a]	6 (35)
Major depressive disorder	6 (35)
Schizoaffective disorder	5 (29)
Schizophrenia	4 (24)
Unspecified schizophrenia-spectrum or psychotic disorder	2 (12)
Bipolar disorder	1 (6)
Race, n (%)	
American Indian or Alaskan Native	1 (6)
Asian	2 (12)
Black or African American	3 (18)
White	11 (65)
Ethnicity, n (%)	
Hispanic	2 (12)
Non-Hispanic	15 (88)
Highest degree, n (%)	
High school diploma or GED[b]	8 (47)
Associate's degree	6 (35)
Bachelor's degree	2 (12)
Other	1 (6)
Marital status, n (%)	
Never married	9 (53)
Married	2 (12)
Divorced	6 (35)
Smartphone ownership, n (%)	
Yes	12 (71)
No	5 (29)
Lifetime hospitalizations, n (%)	
0	3 (18)
1-5	10 (59)
6-10	2 (12)
11-15	0 (0)
≥16	2 (12)

[a]PTSD: posttraumatic stress disorder.

[b]GED: General Educational Development.

Feasibility

On average, the participants completed 85.0 (SD 96.1, median 48.0) EMA interactions with FOCUS and did so on an average of 19.29 (SD 9.27) of 30 access days (mean 64.3%, SD 30.9%). These interactions directly lead to a brief intervention when users indicate distress. In addition to these interactions, the participants used the FOCUS Toolbox (ie, direct access to skills) an average of 49.0 (SD 42.5, median 33.0) times (timestamps of the FOCUS Toolbox uses were not collected, so this figure does not standardize use across participants to the first 30 days of access). All but 1 participant (16/17, 94%) completed all 4 weekly check-ins with the mHealth support specialist by phone. The participant who did not (1/17, 6%) completed 2 of the 4 possible weekly calls. With regard to weekly check-ins with the clinical team, of the 48 times a questionnaire was administered to a clinician with 1 or more clients enrolled in the program, the clinician reported that they found the FOCUS update useful all 48 (100%) times and that these updates affected their clinical care (eg, orienting toward particular clinical concerns and providing additional follow-up) 24 (50%) times.

Acceptability and Usability

The responses to all acceptability-related questions on the SUS are shown in Table 2. Overall, the participants described the intervention as highly acceptable. Nearly all participants reported that they would recommend FOCUS to a friend (16/17, 94%), and most reported that they felt satisfied with FOCUS (15/17, 88%) and would use FOCUS if they had access to it (14/17, 82%). With regard to their experience of its usability, veterans also provided overall positive feedback as nearly all veterans reported feeling comfortable (16/17, 94%) and confident (15/17, 88%) using FOCUS as well as thinking that it was easy to learn (16/17, 94%) and easy to use (16/17, 94%). Very few participants reported that they found FOCUS to be complicated (1/17, 6%) or that they needed to learn a lot (1/17, 6%) or receive technical support to use it (2/17, 12%). Most of the sample reported that they felt that FOCUS helped them manage their symptoms (12/17, 71%).

The participants provided qualitative insights in response to questions related to what they liked and did not like about FOCUS. A prominent positive theme of acceptability involved access to self-management tools. The participants reported that they liked that FOCUS was consistently available to them, that they were able to access helpful tools in the moment, and that they could provide updates about current functioning without having to wait for an upcoming appointment with an in-person provider:

> *I liked always having it on me. The only time I didn't was at church or the store. I like having it on me, documenting my symptoms. [Usually] I have to tell [my clinician] what's going on in a month. With this, it was immediate, I knew someone was reading.* [Participant 4]

> *It's like a 24/7 therapist in my pocket.* [Participant 11]

Other positive participant responses reported an increased propensity to engage in reflection and self-management when they were using FOCUS, identifying either specific skills that they found helpful or describing a general sense that they were more aware of and equipped for coping with symptoms in the moment:

> *The app helped me more quickly identify that I was hearing voices and that I needed and could do something about it.* [Participant 15]

> *I didn't feel like it was completely diffusing my symptoms, but it was like having a safety checklist that told me what I should do when I was struggling-even if I've already tried the skills.* [Participant 16]

Many participants reported that they appreciated the positive and supportive messaging provided by the intervention:

> *I like that it was supportive. It had positive messaging and positive feedback.* [Participant 10]

> *I like that it helped me get into a more positive frame of mind, even if I was reluctant about it, even if I felt reluctant to change.* [Participant 14]

When the participants reported on characteristics that they did not like about the app, fewer consistent themes emerged. The participants most commonly reported on specific design features that would have personalized FOCUS to more directly meet their needs, for example, the addition of a back button or changing particular check-in items:

> *Once you go into the main screen and select a new skill, you can't back out. Made me feel like I was reporting something that I didn't want to report. Also, make this app available for iPhone.* [Participant 12]

> *I would've changed my prompts to check in with my sleep, it would ask me "how did you sleep last night?" That's all I would change.* [Participant 11]

Some participants reported feeling bothered by prompt notifications and how responding to these notifications either felt invasive or forced them to pay closer attention to their phones:

> *Although it was useful, I sometimes wouldn't like when it would ask me to check in. Seemed like an all-day thing. Maybe should have had more information.* [Participant 1]

> *Having to reach for the phone. It got annoying to be prompted to go to the app.* [Participant 9]

Other participants reported disliking the degree of specificity of the intervention content, although some differed on whether the intervention content was too specific or too broad and general to be applied:

> *Sometimes the app felt "wishy washy" or "soft" almost too positive. I would have like to have more time with the app to play with it more.* [Participant 14]

> *Some of the wording. The way they worded sometimes not really getting to the point, but also specific, instead of being broad. That would be better [to be more broad].* [Participant 5]

Table 2. Participant usability and acceptability ratings (N=17).

Item	Disagree or strongly disagree, n (%)	Neutral, n (%)	Agree or strongly agree, n (%)
Acceptability			
I would recommend FOCUS to a friend.	1 (6)	0 (0)	16 (94)
I found the check-ins with the mHealth[a] specialist to be helpful.	0 (0)	1 (6)	16 (94)
I am satisfied with FOCUS.	1 (6)	1 (6)	15 (88)
If I have access to FOCUS, I will use it.	2 (12)	1 (6)	14 (82)
I think that I would like to use FOCUS often.	2 (12)	2 (12)	13 (77)
FOCUS is fun to use.	1 (6)	5 (29)	11 (65)
I feel I need to have FOCUS.	3 (18)	4 (24)	10 (59)
Usability			
The information provided for FOCUS was easy to understand.	0 (0)	0 (0)	17 (100)
The mHealth specialist provided useful feedback on my use of the program.	0 (0)	0 (0)	17 (100)
I felt comfortable using FOCUS.	0 (0)	1 (6)	16 (94)
It was easy to learn to use FOCUS.	0 (0)	1 (6)	16 (94)
How things appeared on the screen was clear.	0 (0)	1 (6)	16 (94)
I thought FOCUS was easy to use.	0 (0)	1 (6)	16 (94)
I felt very confident using FOCUS.	0 (0)	2 (12)	15 (88)
Overall, I am satisfied with how easy it is to use FOCUS.	0 (0)	2 (12)	15 (88)
I found that the different parts of FOCUS work well together.	1 (6)	2 (12)	14 (82)
I was able to complete the modules quickly in FOCUS.	0 (0)	3 (18)	14 (82)
It was easy to find the information I needed.	0 (0)	3 (18)	14 (82)
Whenever I made a mistake using FOCUS, I could recover easily and quickly.	4 (24)	4 (24)	9 (53)
I think that I would need the support of a technical person to be able to use FOCUS.[b]	12 (71)	3 (18)	2 (12)
I found FOCUS to be very complicated.[b]	12 (71)	4 (24)	1 (6)
I needed to learn a lot of things before I could get going with FOCUS.[b]	11 (65)	5 (29)	1 (6)
I thought that there was too much inconsistency in FOCUS.[b]	15 (88)	2 (12)	0 (0)
I found FOCUS very awkward to use.[b]	15 (88)	2 (12)	0 (0)
Veteran fit and adaptation			
FOCUS is appropriate for use with veterans.	0 (0)	1 (6)	16 (94)
I would imagine that most people would learn to use FOCUS very quickly.	1 (6)	2 (12)	14 (82)
FOCUS was interactive enough.	5 (29)	4 (24)	12 (71)
FOCUS helped me manage my symptoms.	3 (18)	2 (12)	12 (71)
FOCUS was well integrated into my usual care at the VA[c] PRRC.[d]	0 (0)	5 (29)	12 (71)
FOCUS works the way I want it to work.	2 (12)	8 (47)	7 (41)

[a]mHealth: mobile health.

[b]Reverse-coded such that disagreement denotes higher perceived usability or acceptability.

[c]VA: Department of Veterans Affairs.

[d]PRRC: Psychosocial Rehabilitation and Recovery Center.

Adaptation for Veterans

In addition to reporting that FOCUS was highly usable and acceptable, the participants provided information related to the fit of FOCUS for veterans and a VA outpatient mental health setting. Nearly all participants (16/17, 94%) reported that they felt FOCUS was appropriate for use with veterans, and most

(12/17, 71%) reported that they felt FOCUS was well-integrated into their usual care at the VA.

The participants also provided additional information about the VA-specific application of FOCUS. At the start of the qualitative questions, the participants were asked whether they would recommend FOCUS to a fellow veteran, and their responses were almost uniformly affirmative (16/17, 94%). When asked to identify how FOCUS could be adapted to improve its acceptability among veterans, most participants reported that they had no suggestions for adaptations and that FOCUS nicely paralleled their current treatment needs as the intervention provides access to similar skills to those emphasized in VA outpatient services:

> *There are vets that have seen combat, war, and this FOCUS app would be a good resource to help curb the PTSD they might develop. Helped me be more positive and helped me realize my moods, and helped remind me to take my meds. This will help open people's minds to being more open to getting help...It was a good experience and it's good for veterans and it's a positive influence tool to help the veteran in their therapy.* [Participant 12]

> *Veterans can help find a way to subside the voices, because the app will help them. They just have to listen to the app's suggestions.* [Participant 15]

The participants specifically emphasized that FOCUS was helpful in reducing negative thinking and decreasing stress and that these characteristics were particularly well-suited to a veteran population:

> *I think it would help people. If you have a lot of negative thoughts you can check in with yourself and get out of your head.* [Participant 13]

With regard to improvements and adaptations for veteran populations, the participants commonly identified adaptations that would improve FOCUS for subpopulations of veterans, for example, veterans with hearing impairments or PTSD:

> *A way for hearing and vision impaired veterans to be able to use the app. I can't think of how but a way for them to use the app too.* [Participant 14]

> *Have more solutions, more things going on. More content. Maybe for PTSD. These guys have a hard time, probably worse than I have. PTSD support.* [Participant 4]

> *Expand the voices option. I think people with PTSD hear things in their own head. That would be an improvement.* [Participant 5]

A second emergent theme involved integrating FOCUS more closely into existing VA services. Notably, on the SUS, fewer participants reported that they felt FOCUS was well-integrated into their routine services than those who reported that they enjoyed the use of the app or mHealth coaching. Some participants commented on connecting FOCUS to existing structures, including referral services or group meetings:

> *Connecting it to existing care, like having an mHealth referral service in VA. The doctor could recommend it to a veteran, and then a coordinator picks it up.* [Participant 2]

> *Hold group meetings for FOCUS, to get together with other veterans to discuss and share how everyone is managing their symptoms. We could compare notes with each other. We need more apps like this for veterans.* [Participant 15]

Clinical Outcomes

The summary statistics of the models examining clinical outcomes are reported in Table 3. Paired sample *t* tests were conducted to examine within-participant changes during the 1-month study period. Given that the primary aim of this pilot study was to establish the feasibility and acceptability of this approach in a VA setting, these analyses were underpowered to detect significant clinical effects; however, we report the effect sizes here. Small positive effects were detected for participants in self-directed recovery (Illness Management and Recovery Scale; Cohen d=0.30), quality of life (Quality of Life Enjoyment and Satisfaction Questionnaire; Cohen d=0.25), and severity of the voices (Hamilton Program for Schizophrenia Voices Questionnaire; Cohen d=0.23).

Table 3. Baseline and posttest scores of clinical outcome measures.[a]

Clinical outcome measure	Baseline score, mean (SD)	Posttest score, mean (SD)	Difference, mean (SD)	*t* test (*df*)	*P* value	Cohen *d*
Recovery (IMRS[b])	34.71 (5.65)	35.94 (6.67)	−1.24 (4.19)	1.22 (16)	.24	0.30
Quality of life (QLES-Q[c])[d]	49.44 (9.02)	51.31 (6.73)	1.88 (7.64)	0.98 (15)	.34	0.25
Voices (HPSVQ[e])	20.50 (5.68)	19.20 (6.32)	1.30 (5.35)	0.77 (9)	.46	0.24
Insomnia (ISI[f])	11.35 (6.12)	10.71 (5.75)	0.64 (5.06)	0.53 (16)	.61	0.13
Depression (BDI-II[g])[d]	25.44 (13.93)	24.50 (9.37)	−0.94 (7.69)	0.49 (15)	.63	0.12
Medication beliefs (BMQ[h])	11.00 (11.18)	11.53 (10.52)	0.53 (5.36)	0.41 (16)	.69	0.10
Paranoia (GPTS[i])[d]	67.63 (30.71)	69.94 (32.77)	2.31 (22.72)	0.41 (15)	.69	−0.10

[a]All the effects were statistically nonsignificant. Effect sizes are computed such that positive values reflect changes in the expected direction.

[b]IMRS: Illness Management and Recovery Scale.

[c]QLES-Q: Quality of Life Enjoyment and Satisfaction Questionnaire.

[d]Because of missing data from skipped items, N=16 for analyses involving the Beck Depression Inventory–Second Edition, QLES, and Green Paranoid Thoughts Scale.

[e]HPSVQ: Hamilton Program for Schizophrenia Voices Questionnaire. HPSVQ scores reported are those of participants (n=10) who reported any level of auditory verbal hallucinations at baseline and completed the study.

[f]ISI: Insomnia Severity Index.

[g]BDI-II: Beck Depression Inventory–Second Edition.

[h]BMQ: Brief Medication Questionnaire.

[i]GPTS: Green Paranoid Thoughts Scale.

Discussion

Principal Findings

This study aimed to examine the feasibility, acceptability, usability, and preliminary effectiveness of the FOCUS mHealth intervention in a VA psychosocial rehabilitation outpatient setting. The participants used FOCUS frequently during the month-long deployment period (mean 85.0, SD 96.1 assessed interactions and mean 64.3%, SD 30.9% of days enrolled in the study) and overwhelmingly reported that they found the intervention acceptable and usable. This matches previous work examining the acceptability of FOCUS in non-VA populations [49]. When asked to elaborate on adaptation for the VA setting, veterans largely found the intervention ready to deploy, but a few participants provided suggestions for improvement, including content for specific veteran subpopulations (ie, PTSD or sensory impairments) as well as integration into existing services (ie, referral services and mental health groups). The trial was underpowered to detect statistically significant changes in clinical outcomes, and the effect sizes were consistent with small improvements. Together with existing research supporting the effectiveness of a 3-month deployment of FOCUS [27], this pilot study suggests that the FOCUS mHealth intervention is appropriate for a large-scale trial in a VA setting to evaluate effectiveness.

Use statistics suggested that the participants were able to access a substantial weekly dose of the FOCUS clinical intervention during the 1-month study period. The participants also almost unanimously completed a weekly FOCUS check-in call every week that they were enrolled. This high rate of use mirrors previous studies of FOCUS, including among those with a recent psychiatric hospitalization and individuals enrolled in outpatient community mental health [50]. These use rates are particularly notable in a veteran population given the low rates of veteran use of existing VA or Department of Defense mental health apps [31]. These results suggest that a usability-tested mHealth intervention such as FOCUS, together with weekly mHealth support and coaching from a member of the study team, could sufficiently engage veterans enrolled in outpatient mental health services.

Regarding fit for veterans, many participants reported feeling that FOCUS symptom management skills closely mirrored their current mental health treatment, particularly in its impact on reducing unhelpful negative thinking. Some participants provided recommendations for VA-specific adaptations, including subpopulation-relevant content (eg, comorbid PTSD support) and creating referral pathways for mHealth provision, as well as the development of mental health groups where veterans can practice FOCUS skills in a socially supportive format. Despite a growing body of evidence, few mHealth interventions have been implemented in real-world practice. As one of the nation's largest health care providers, the VA could provide fertile ground for testing of various mHealth implementation models, including, for example, an embedded mHealth specialist in primary care or a supportive group in outpatient mental health. Future hybrid and implementation-oriented work could identify the specific organization-related variables linked with the most successful VA deployments of mHealth for SMIs.

The participants' overall ratings of the usability and acceptability of the intervention were high and closely mirrored comments regarding acceptability in non-VA community mental health settings [49]. All but 1 participant (16/17, 94%) reported that they would recommend FOCUS to a friend and that FOCUS

was appropriate for use with veterans. Qualitative responses suggested that the participants particularly appreciated the positive tone of the messages, the symptom management skills delivered, its around-the-clock availability for support, and its simplicity and straightforward design. In addition to these positive comments, the participants reported on features of the intervention that they did not enjoy, including specific design features (eg, the inability to go *back* and having limited time to respond to prompts) and being interrupted by device notifications from FOCUS, as well as suboptimal degree of specific versus broad app content (though this varied across participants as to which was preferred). On one hand, these specific points of feedback were relatively uncommon, and most participants reported high levels of satisfaction with the FOCUS app itself. In contrast, FOCUS could benefit from improved personalization and fit to the user's specific needs and preferences. Future innovations could allow for automated customization to meet this objective.

The clinical effects were smaller than those reported in other clinical trials examining FOCUS [26]. At posttest, the participants experienced small but nonsignificant improvements in recovery, quality of life, and severity of auditory hallucinations. The study sample may have affected these results. The participants enrolled in this trial were well-established in a psychosocial rehabilitation program, and FOCUS was provided as an adjunct to existing services. The participants were not required to be naive to the interventions on which FOCUS was based (eg, CBT for insomnia, CBT for psychosis, and social skill training), and many reported that the intervention content mirrored care they had already received. Furthermore, the participants received 1 month of the FOCUS intervention rather than the 3-month period that has been suggested as standard in full-scale trials [27]. It is possible that treatment effects would have been larger after a full course of the intervention.

Other study limitations warrant mention. Given the small sample size and brief study period, our findings speak primarily to the feasibility, acceptability, and usability of FOCUS in a veteran population. Conclusions related to clinical benefits cannot be drawn. Second, the clinical model for this deployment of FOCUS involved weekly calls from a member of the study team. This model may have limited generalizability to clinics where frontline clinicians may be operating in this mHealth clinical support role. Furthermore, although updates were provided to the participants' mental health clinicians, there was no specific protocol in place to make FOCUS data actionable. Given the brief length of the trial, many participants also reported that they did not meet with their primary clinician for an individual session during the intervention period; therefore, the benefits of ongoing FOCUS assessments to routine care were not explicitly examined. Finally, in general, given the multiple components of the intervention (ie, mobile app, weekly check-in calls, and communication with the primary clinical team), it will be difficult to know without more rigorous trials the extent to which any clinical gains might be attributable to particular components of the intervention. Future work should also examine whether benefits might differ in various subgroups of veterans, including those with varying degrees of digital literacy.

Conclusions

Overall, the results suggest that FOCUS is feasible, acceptable, and usable to a veteran population. Future randomized effectiveness and hybrid trials can provide insight into the specific adaptations to ensure successful implementation of mHealth for SMIs in the VA population. If effective, FOCUS could fill a critical gap in the currently available suite of VA mobile apps and has potential for significant impact on the VA. This study suggests that future work is warranted and provides initial suggestions for such efforts.

Acknowledgments

This work was supported by a Department of Veterans Affairs Puget Sound Health Care System Research and Development Seed Grant (MIRB 01624) awarded to GMR (principal investigator). BB is supported by a Mentored Patient-Oriented Research Career Development Award (K23MH122504). The contents do not represent the views of the US Department of Veterans Affairs, the National Institute of Mental Health, or the US government.

Conflicts of Interest

GMR edited *Technology and Mental Health: A Clinician's Guide to Improved Clinical Outcomes* and will receive royalties from Routledge following its publication. DBZ has an intervention content licensing agreement with Pear Therapeutics and has financial interest in FOCUS technology. He has consulted for Trusst Health Inc, eQuility, and Otsuka Pharmaceuticals Ltd. The other authors have no conflicts of interest to disclose.

References

1. Wiersma D, Wanderling J, Dragomirecka E, Ganev K, Harrison G, An Der Heiden W, et al. Social disability in schizophrenia: its development and prediction over 15 years in incidence cohorts in six European centres. Psychol Med 2000 Sep;30(5):1155-1167. [doi: 10.1017/s0033291799002627] [Medline: 12027051]
2. Tsang H, Leung A, Chung R, Bell M, Cheung W. Review on vocational predictors: a systematic review of predictors of vocational outcomes among individuals with schizophrenia: an update since 1998. Aust N Z J Psychiatry 2010 Jun;44(6):495-504. [doi: 10.3109/00048671003785716] [Medline: 20482409]

3. Folsom D, Jeste DV. Schizophrenia in homeless persons: a systematic review of the literature. Acta Psychiatr Scand 2002 Jun;105(6):404-413. [doi: 10.1034/j.1600-0447.2002.02209.x] [Medline: 12059843]
4. Hayes JF, Miles J, Walters K, King M, Osborn DP. A systematic review and meta-analysis of premature mortality in bipolar affective disorder. Acta Psychiatr Scand 2015 Jun;131(6):417-425 [FREE Full text] [doi: 10.1111/acps.12408] [Medline: 25735195]
5. Saha S, Chant D, McGrath J. A systematic review of mortality in schizophrenia. Arch Gen Psychiatry 2007 Oct 01;64(10):1123. [doi: 10.1001/archpsyc.64.10.1123]
6. Jääskeläinen E, Juola P, Hirvonen N, McGrath JJ, Saha S, Isohanni M, et al. A systematic review and meta-analysis of recovery in schizophrenia. Schizophr Bull 2013 Nov;39(6):1296-1306 [FREE Full text] [doi: 10.1093/schbul/sbs130] [Medline: 23172003]
7. Lean M, Fornells-Ambrojo M, Milton A, Lloyd-Evans B, Harrison-Stewart B, Yesufu-Udechuku A, et al. Self-management interventions for people with severe mental illness: systematic review and meta-analysis. Br J Psychiatry 2019 May;214(5):260-268 [FREE Full text] [doi: 10.1192/bjp.2019.54] [Medline: 30898177]
8. Lutgens D, Gariepy G, Malla A. Psychological and psychosocial interventions for negative symptoms in psychosis: systematic review and meta-analysis. Br J Psychiatry 2017 May;210(5):324-332. [doi: 10.1192/bjp.bp.116.197103] [Medline: 28302699]
9. Petros R, Solomon P. Reviewing illness self-management programs: a selection guide for consumers, practitioners, and administrators. Psychiatr Serv 2015 Nov;66(11):1180-1193. [doi: 10.1176/appi.ps.201400355] [Medline: 26129995]
10. Goldberg RW, Resnick SG. US Department of Veterans Affairs (VA) efforts to promote psychosocial rehabilitation and recovery. Psychiatr Rehabil J 2010;33(4):255-258. [doi: 10.2975/33.4.2010.255.258] [Medline: 20374981]
11. Veterans Health Administration. U.S. Department of Veterans Affairs. URL: https://www.va.gov/health/ [accessed 2022-01-18]
12. Wu EQ, Shi L, Birnbaum H, Hudson T, Kessler R. Annual prevalence of diagnosed schizophrenia in the USA: a claims data analysis approach. Psychol Med 2006 Nov;36(11):1535-1540. [doi: 10.1017/S0033291706008191] [Medline: 16907994]
13. Trivedi RB, Post EP, Sun H, Pomerantz A, Saxon AJ, Piette JD, et al. Prevalence, comorbidity, and prognosis of mental health among US veterans. Am J Public Health 2015 Dec;105(12):2564-2569. [doi: 10.2105/AJPH.2015.302836] [Medline: 26474009]
14. Birgenheir DG, Ilgen MA, Bohnert AS, Abraham KM, Bowersox NW, Austin K, et al. Pain conditions among veterans with schizophrenia or bipolar disorder. Gen Hosp Psychiatry 2013;35(5):480-484. [doi: 10.1016/j.genhosppsych.2013.03.019] [Medline: 23639185]
15. Breland JY, Phibbs CS, Hoggatt KJ, Washington DL, Lee J, Haskell S, et al. The obesity epidemic in the veterans health administration: prevalence among key populations of women and men veterans. J Gen Intern Med 2017 Apr;32(Suppl 1):11-17 [FREE Full text] [doi: 10.1007/s11606-016-3962-1] [Medline: 28271422]
16. Calhoun PS, Stechuchak KM, Strauss J, Bosworth HB, Marx CE, Butterfield MI. Interpersonal trauma, war zone exposure, and posttraumatic stress disorder among veterans with schizophrenia. Schizophr Res 2007 Mar;91(1-3):210-216. [doi: 10.1016/j.schres.2006.12.011] [Medline: 17276658]
17. Hjorthøj C, Stürup AE, McGrath JJ, Nordentoft M. Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. Lancet Psychiatry 2017 Apr;4(4):295-301. [doi: 10.1016/S2215-0366(17)30078-0] [Medline: 28237639]
18. Drapalski AL, Milford J, Goldberg RW, Brown CH, Dixon LB. Perceived barriers to medical care and mental health care among veterans with serious mental illness. Psychiatr Serv 2008 Aug;59(8):921-924. [doi: 10.1176/ps.2008.59.8.921] [Medline: 18678691]
19. Haddock G, Eisner E, Boone C, Davies G, Coogan C, Barrowclough C. An investigation of the implementation of NICE-recommended CBT interventions for people with schizophrenia. J Ment Health 2014 Aug;23(4):162-165. [doi: 10.3109/09638237.2013.869571] [Medline: 24433132]
20. McCarthy JF, Blow FC, Valenstein M, Fischer EP, Owen RR, Barry KL, et al. Veterans Affairs Health System and mental health treatment retention among patients with serious mental illness: evaluating accessibility and availability barriers. Health Serv Res 2007 Jun;42(3 Pt 1):1042-1060 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00642.x] [Medline: 17489903]
21. Fischer EP, McCarthy JF, Ignacio RV, Blow FC, Barry KL, Hudson TJ, et al. Longitudinal patterns of health system retention among veterans with schizophrenia or bipolar disorder. Community Ment Health J 2008 Oct;44(5):321-330. [doi: 10.1007/s10597-008-9133-z] [Medline: 18401711]
22. Ben-Zeev D, Razzano LA, Pashka NJ, Levin CE. Cost of mHealth versus clinic-based care for serious mental illness: same effects, half the price tag. Psychiatr Serv 2021 Apr 01;72(4):448-451. [doi: 10.1176/appi.ps.202000349] [Medline: 33557599]
23. Torous J. Technology and smartphone ownership, interest, and engagement among those with schizophrenia. Biological Psychiatry 2018 May;83(9):S62. [doi: 10.1016/j.biopsych.2018.02.170]
24. Gay K, Torous J, Joseph A, Pandya A, Duckworth K. Digital technology use among individuals with schizophrenia: results of an online survey. JMIR Ment Health 2016 May 04;3(2):e15 [FREE Full text] [doi: 10.2196/mental.5379] [Medline: 27146094]

25. Ben-Zeev D, Kaiser SM, Brenner CJ, Begale M, Duffecy J, Mohr DC. Development and usability testing of FOCUS: a smartphone system for self-management of schizophrenia. Psychiatr Rehabil J 2013 Dec;36(4):289-296 [FREE Full text] [doi: 10.1037/prj0000019] [Medline: 24015913]
26. Ben-Zeev D, Brenner CJ, Begale M, Duffecy J, Mohr DC, Mueser KT. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. Schizophr Bull 2014 Nov;40(6):1244-1253 [FREE Full text] [doi: 10.1093/schbul/sbu033] [Medline: 24609454]
27. Ben-Zeev D, Brian RM, Jonathan G, Razzano L, Pashka N, Carpenter-Song E, et al. Mobile health (mHealth) versus clinic-based group intervention for people with serious mental illness: a randomized controlled trial. Psychiatr Serv 2018 Sep 01;69(9):978-985. [doi: 10.1176/appi.ps.201800063] [Medline: 29793397]
28. Gould CE, Kok BC, Ma VK, Zapata AM, Owen JE, Kuhn E. Veterans Affairs and the Department of Defense mental health apps: a systematic literature review. Psychol Serv 2019 May;16(2):196-207. [doi: 10.1037/ser0000289] [Medline: 30431306]
29. Kuhn E, Kanuri N, Hoffman JE, Garvert DW, Ruzek JI, Taylor CB. A randomized controlled trial of a smartphone app for posttraumatic stress disorder symptoms. J Consult Clin Psychol 2017 Mar;85(3):267-273. [doi: 10.1037/ccp0000163] [Medline: 28221061]
30. Bush NE, Smolenski DJ, Denneson LM, Williams HB, Thomas EK, Dobscha SK. A virtual hope box: randomized controlled trial of a smartphone app for emotional regulation and coping with distress. Psychiatr Serv 2017 Apr 01;68(4):330-336. [doi: 10.1176/appi.ps.201600283] [Medline: 27842473]
31. Reger GM, Harned M, Stevens ES, Porter S, Nguyen J, Norr AM. Mobile applications may be the future of veteran mental health support but do veterans know yet? A survey of app knowledge and use. Psychol Serv. Preprint posted online on June 3, 2021. [doi: 10.1037/ser0000562] [Medline: 34081527]
32. Erbes CR, Stinson R, Kuhn E, Polusny M, Urban J, Hoffman J, et al. Access, utilization, and interest in mHealth applications among veterans receiving outpatient care for PTSD. Mil Med 2014 Nov;179(11):1218-1222. [doi: 10.7205/MILMED-D-14-00014] [Medline: 25373044]
33. US Department of Veterans Affairs (VA). App Store Preview. URL: https://apps.apple.com/ai/developer/us-department-of-veterans-affairs-va/id430646305 [accessed 2022-01-18]
34. Gierz M, Jeste DV. Physical comorbidity in elderly veterans affairs patients with schizophrenia and depression. Am J Geriatr Psychiatry 1993;1(2):165-170. [doi: 10.1097/00019442-199300120-00010] [Medline: 28531032]
35. Lambert B, Cunningham F, Miller D, Dalack G, Hur K. Diabetes risk associated with use of olanzapine, quetiapine, and risperidone in veterans health administration patients with schizophrenia. Am J Epidemiol 2006 Oct 01;164(7):672-681. [doi: 10.1093/aje/kwj289] [Medline: 16943266]
36. Magruder KM, Yeager DE. The prevalence of PTSD across war eras and the effect of deployment on PTSD: a systematic review and meta-analysis. Psychiatr Ann 2009 Aug 01;39(8):778-788. [doi: 10.3928/00485713-20090728-04]
37. Strauss J, Calhoun P, Marx C, Stechuchak K, Oddone E, Swartz M, et al. Comorbid posttraumatic stress disorder is associated with suicidality in male veterans with schizophrenia or schizoaffective disorder. Schizophr Res 2006 May;84(1):165-169. [doi: 10.1016/j.schres.2006.02.010] [Medline: 16567080]
38. Hoge CW, Grossman SH, Auchterlonie JL, Riviere LA, Milliken CS, Wilk JE. PTSD treatment for soldiers after combat deployment: low utilization of mental health care and reasons for dropout. Psychiatr Serv 2014 Aug 01;65(8):997-1004. [doi: 10.1176/appi.ps.201300307] [Medline: 24788253]
39. Jameson J, Farmer M, Head K, Fortney J, Teal C. VA community mental health service providers' utilization of and attitudes toward telemental health care: the gatekeeper's perspective. J Rural Health 2011;27(4):425-432. [doi: 10.1111/j.1748-0361.2011.00364.x] [Medline: 21967387]
40. Jonathan GK, Pivaral L, Ben-Zeev D. Augmenting mHealth with human support: notes from community care of people with serious mental illnesses. Psychiatr Rehabil J 2017 Sep;40(3):336-338 [FREE Full text] [doi: 10.1037/prj0000275] [Medline: 28891660]
41. BDI-II, Beck Depression Inventory: Manual. San Antonio, TX: Psychological Corporation; 1996.
42. Van Lieshout RJ, Goldberg JO. Quantifying self-reports of auditory verbal hallucinations in persons with psychosis. Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement 2007;39(1):73-77. [doi: 10.1037/cjbs2007006]
43. Green CE, Freeman D, Kuipers E, Bebbington P, Fowler D, Dunn G, et al. Measuring ideas of persecution and social reference: the Green et al. Paranoid Thought Scales (GPTS). Psychol Med 2008 Jan;38(1):101-111. [doi: 10.1017/S0033291707001638] [Medline: 17903336]
44. Bastien C, Vallières A, Morin C. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med 2001 Jul;2(4):297-307. [doi: 10.1016/s1389-9457(00)00065-4] [Medline: 11438246]
45. Endicott J, Nee J, Harrison W, Blumenthal R. Quality of life enjoyment and satisfaction questionnaire. APA PsycTests 1993 [FREE Full text] [doi: 10.1037/t49981-000]
46. Ritsner M, Kurs R, Gibel A, Ratner Y, Endicott J. Validity of an abbreviated quality of life enjoyment and satisfaction questionnaire (Q-LES-Q-18) for schizophrenia, schizoaffective, and mood disorder patients. Qual Life Res 2005 Sep;14(7):1693-1703. [doi: 10.1007/s11136-005-2816-9] [Medline: 16119181]

47. Färdig R, Lewander T, Fredriksson A, Melin L. Evaluation of the illness management and recovery scale in schizophrenia and schizoaffective disorder. Schizophr Res 2011 Nov;132(2-3):157-164. [doi: 10.1016/j.schres.2011.07.001] [Medline: 21798718]
48. Mueser KT, Corrigan PW, Hilton DW, Tanzman B, Schaub A, Gingerich S, et al. Illness management and recovery: a review of the research. Psychiatr Serv 2002 Oct;53(10):1272-1284. [doi: 10.1176/appi.ps.53.10.1272] [Medline: 12364675]
49. Jonathan G, Carpenter-Song EA, Brian RM, Ben-Zeev D. Life with FOCUS: a qualitative evaluation of the impact of a smartphone intervention on people with serious mental illness. Psychiatr Rehabil J 2019 Jun;42(2):182-189. [doi: 10.1037/prj0000337] [Medline: 30589278]
50. Achtyes ED, Ben-Zeev D, Luo Z, Mayle H, Burke B, Rotondi AJ, et al. Off-hours use of a smartphone intervention to extend support for individuals with schizophrenia spectrum disorders recently discharged from a psychiatric hospital. Schizophr Res 2019 Apr;206:200-208. [doi: 10.1016/j.schres.2018.11.026] [Medline: 30551981]

Abbreviations

CBT: cognitive behavioral therapy
EMA: ecological momentary assessment
mHealth: mobile health
PRRC: Psychosocial Rehabilitation and Recovery Center
PTSD: posttraumatic stress disorder
SMI: serious mental illness
SUS: System Usability Scale
VA: Department of Veterans Affairs

Edited by J Torous; submitted 25.11.20; peer-reviewed by C Arnold, T Campellone, S Byrne, G Jonathan; comments to author 22.01.21; revised version received 15.02.21; accepted 04.10.21; published 28.01.22.

Please cite as:
Buck B, Nguyen J, Porter S, Ben-Zeev D, Reger GM
FOCUS mHealth Intervention for Veterans With Serious Mental Illness in an Outpatient Department of Veterans Affairs Setting: Feasibility, Acceptability, and Usability Study
JMIR Ment Health 2022;9(1):e26049
URL: https://mental.jmir.org/2022/1/e26049
doi:10.2196/26049
PMID:35089151

Original Paper

Social Equity in the Efficacy of Computer-Based and In-Person Brief Alcohol Interventions Among General Hospital Patients With At-Risk Alcohol Use: A Randomized Controlled Trial

Jennis Freyer-Adam[1,2], PhD; Sophie Baumann[3], PhD; Gallus Bischof[4], PhD; Andreas Staudt[3,5], PhD; Christian Goeze[1], Dipl -Ing; Beate Gaertner[6], PhD; Ulrich John[2,7], PhD

[1]Institute for Medical Psychology, University Medicine Greifswald, Greifswald, Germany

[2]German Centre for Cardiovascular Research (DZHK), Greifswald, Germany

[3]Department of Methods in Community Medicine, Institute of Community Medicine, University Medicine Greifswald, Greifswald, Germany

[4]Department of Psychiatry and Psychotherapy, Medical University of Lübeck, Luebeck, Germany

[5]Institute and Policlinic of Occupational and Social Medicine, Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

[6]Department of Epidemiology and Health Monitoring, Robert Koch Institute Berlin, Berlin, Germany

[7]Department of Prevention Research and Social Medicine, Institute of Community Medicine, University Medicine Greifswald, Greifswald, Germany

Corresponding Author:
Jennis Freyer-Adam, PhD
Institute for Medical Psychology
University Medicine Greifswald
Walther-Rathenau-Str. 48
Greifswald, 17475
Germany
Phone: 49 3834865606
Fax: 49 3834865605
Email: Jennis.Freyer-Adam@med.uni-greifswald.de

Abstract

Background: Social equity in the efficacy of behavior change intervention is much needed. While the efficacy of brief alcohol interventions (BAIs), including digital interventions, is well established, particularly in health care, the social equity of interventions has been sparsely investigated.

Objective: We aim to investigate whether the efficacy of computer-based versus in-person delivered BAIs is moderated by the participants' socioeconomic status (ie, to identify whether general hospital patients with low-level education and unemployed patients may benefit more or less from one or the other way of delivery compared to patients with higher levels of education and those that are employed).

Methods: Patients with nondependent at-risk alcohol use were identified through systematic offline screening conducted on 13 general hospital wards. Patients were approached face-to-face and asked to respond to an app for self-assessment provided by a mobile device. In total, 961 (81% of eligible participants) were randomized and received their allocated intervention: computer-generated and individually tailored feedback letters (CO), in-person counseling by research staff trained in motivational interviewing (PE), or assessment only (AO). CO and PE were delivered on the ward and 1 and 3 months later, were based on the transtheoretical model of intentional behavior change and required the assessment of intervention data prior to each intervention. In CO, the generation of computer-based feedback was created automatically. The assessment of data and sending out feedback letters were assisted by the research staff. Of the CO and PE participants, 89% (345/387) and 83% (292/354) received at least two doses of intervention, and 72% (280/387) and 54% (191/354) received all three doses of intervention, respectively. The outcome was change in grams of pure alcohol per day after 6, 12, 18, and 24 months, with the latter being the primary time-point of interest. Follow-up interviewers were blinded. Study group interactions with education and employment status were tested as predictors of change in alcohol use using latent growth modeling.

Results: The efficacy of CO and PE did not differ by level of education (P=.98). Employment status did not moderate CO efficacy (Ps≥.66). Up to month 12 and compared to employed participants, unemployed participants reported significantly greater drinking reductions following PE versus AO (incidence rate ratio 0.44, 95% CI 0.21-0.94; P=.03) and following PE versus CO (incidence rate ratio 0.48, 95% CI 0.24–0.96; P=.04). After 24 months, these differences were statistically nonsignificant (Ps≥.31).

Conclusions: Computer-based and in-person BAI worked equally well independent of the patient's level of education. Although findings indicate that in the short-term, unemployed persons may benefit more from BAI when delivered in-person rather than computer-based, the findings suggest that both BAIs have the potential to work well among participants with low socioeconomic status.

Trial Registration: ClinicalTrials.gov NCT01291693; https://clinicaltrials.gov/ct2/show/NCT01291693

(JMIR Ment Health 2022;9(1):e31712) doi:10.2196/31712

KEYWORDS

brief alcohol intervention; electronic; eHealth; digital; motivational interviewing; socioeconomic status; equity; social inequality; transtheoretical model; moderator; mental health; public health; alcohol interventions; digital intervention; digital health intervention; alcohol use

Introduction

People with low socioeconomic status (SES) have a greater risk of cancer, cardiovascular, and all-cause mortality [1]. Social inequality in health and mortality is increasing [2-4], and alcohol-related mortality plays a crucial role [5]. People with low SES have a 1.7-fold increased risk of dying from alcohol-attributable causes [6]. Alcohol-related causes are responsible for 5% of social inequality in total mortality in European men aged 35 to 79 years, and in some Eastern and Northern European countries, they account for 10% or more [7]. In addition, SES moderates the effect of alcohol use on harm (ie, even when alcohol use is uniform, alcohol-attributable harm is greater in people with low SES [8]).

To close the social inequity gap, behavior change interventions need positive social equity impact (ie, greater reach and greater efficacy in low vs high SES people [5]). To prevent the further widening of the social inequality gap, interventions need neutral impact (ie, equal reach and equal efficacy in low and high SES people). Interventions with greater reach and greater efficacy in high than in low SES people have a negative social equity impact. As reach and efficacy constitute two dimensions of the public health impact of interventions [9], achieving positive or neutral social equity impact at least is a crucial challenge for preventive efforts directly targeting behavior change on the population level.

However, while effective brief alcohol interventions (BAI) have been developed as supported by numerous systematic reviews and meta-analyses [10-15], research findings on the social equity impact of BAI are less encouraging. Firstly, intervention trials, including our own, often report a lower reach of people with low SES or low education, an SES indicator [16,17]. Secondly, little research has been done on the moderating effects of SES indicators, such as level of education and employment status, on intervention efficacy in general. Particularly, little is known about the effect of unemployment status [18]. Thirdly, in some studies, efficacy was found to be reduced in people with lower levels of education than in people with higher levels of education [19,20], indicating that behavior change interventions may have a negative impact on social equity. Reviews revealed a neutral impact once the participants had been recruited [17,21].

Moreover, the development of digital behavior change interventions is advancing. Computer-based interventions have been found to reduce alcohol use in health care [22-24] and beyond [21,25-28]. As they require fewer resources than in-person delivered interventions, their potential impact on public health and social equity may be considered high. Among general hospital patients, our research group showed that computer-based BAI was no less effective than in-person BAI in reducing alcohol use and improving measures of health over two years [29-31]. Thus, computer-based BAI appears to be incorporable into a broader health care program. However, little is known about whether computer-based and in-person delivered interventions work differently for people with low versus high SES.

The aim of this study was to investigate two indicators of SES as moderators of BAI efficacy, namely level of education and employment status. Specifically, we aimed to investigate 3 questions: (1) Does the efficacy of computer-based BAI differ between persons with low versus high levels of education and between unemployed versus employed persons? (2) Does the efficacy of in-person BAI differ between persons with low versus high levels of education and between unemployed versus employed persons? (3) Does the comparative efficacy of computer-based versus in-person BAI differ between persons with low versus high levels of education and between unemployed versus employed persons?

Methods

Overview

The data used for these analyses are from the three-arm randomized controlled trial (RCT) entitled "Testing delivery channels of individualized motivationally tailored alcohol interventions among general hospital patients: in-person versus computer-based, PECO" (ClinicalTrials.gov: NCT01291693). The local ethics committee approved the study (BB 07/10, BB 05/13), and the study was conducted as planned.

Sample recruitment took place from February 2011 to July 2012 on four medical departments (internal medicine, surgical medicine, trauma surgery, and ear-nose-throat wards) of the University Medicine Hospital Greifswald [16,31]. All consecutively admitted patients aged 18 to 64 years were first approached face-to-face and asked to respond to an app for self-assessment of health behaviors provided by a mobile device. Patients were excluded from screening if they were cognitively or physically incapable or terminally ill, discharged or transferred within the first 24 hours, already recruited, employed at the conducting research institute, or if they had highly

infectious diseases or insufficient language skills. Computer literacy was not required. If needed, participants received a quick introduction about handling the mobile device and assessment app. Patients screening positive for at-risk alcohol use (ie, women or men with ≥4 or ≥5 points on the Alcohol Use Disorders Identification Test [AUDIT]-Consumption) [32,33] and negative for more severe alcohol problems (ie, persons with <20 on the AUDIT) [34,35] were eligible for the PECO trial.

As described in more detail elsewhere [31], enrolment was done by research assistants. Patients who provided informed written consent to participate in the trial were asked to respond to more questions on alcohol use and motivation using the app for self-assessment and were allocated to computer-based BAI (CO), in-person BAI (PE), or assessment only (AO). A sample size of 975 participants with an allocation ratio of 2:2:1 was calculated to be sufficient to detect small intervention effects concerning reduced gram of pure alcohol use, the primary outcome of the RCT [31]. Allocation was computerized and depended on the week and ward to avoid the exchange of information between study groups. Recruitment was stopped after the intended sample size was reached within the planned recruitment time of 18 months.

Interventions

As described in more detail elsewhere, CO and PE were designed to be comparable in terms of intervention dose and content and primarily differed in method of delivery [16,31,36].

The CO group received individually tailored feedback letters at baseline, 1, and 3 months. Based on electronic and standardized data assessment, 3 to 4-page letters were created automatically by an expert system software. The software was programmed in MS Access and handled by the research staff. For the 1-month and 3-month interventions, participants were first phoned by research assistants and asked to respond to computer-assisted telephone interviews. Afterward, the software selected text modules and graphical visualizations based on the participant's assessment data and predefined selection rules [37]. In accordance with the transtheoretical model of intentional behavior change, feedback depended on each participant's current motivational stage of change [38]. Participants also received normative feedback, specifically feedback on (1) their current alcohol use in comparison to others of the same gender and (2) according to theoretical constructs such as processes of change, decisional balance, and self-efficacy [39] in comparison to others in the same motivational stage. At baseline, individually tailored text modules were selected from a pool of 120 text modules. At months 1 and 3, the pool was comprised of about 270 text modules as the participants also received ipsative feedback (ie, feedback on how the participant's current data on drinking and motivation compared to the participant's previous data). Information on the limits of low-risk drinking was provided at all time points [40]. The letters were then handed or sent out by research assistants along with a stage-matched self-help manual. Of the CO participants, 89% (345/387) received at least two feedback letters, and 72% (280/387) received all three feedback letters [16].

The PE group received in-person counseling at baseline (face-to-face on the ward) and 1 and 3 months later (via telephone). Counseling was delivered by research staff trained in motivational interviewing [41] techniques and supervised on a regular basis. Like CO, PE was stage-matched and included normative and ipsative feedback on alcohol use and theoretical constructs and information on the limits of low-risk drinking. Counselors received a one-page manual, including the same computer-generated feedback information as the letters used in CO, to ensure comparability. Over 3 months, PE participants received a total of 35 minutes (median) of counseling, with 83% (292/354) of them being counseled over at least two consultations and 54% (191/354) over three consultations. PE was delivered with acceptable adherence to motivational interviewing [16,31].

Participants in the AO group received minimal assessment at baseline (including sociodemographics, alcohol use, and motivational stage) and were not contacted at months 1 and 3.

Measures

The outcome in this study was grams of pure alcohol consumed per day. At baseline and at all follow-ups, grams per day were assessed by 2 questions concerning the previous month. The frequency question ("In [month], how often did you have an alcoholic drink?") included 5 response categories: never (0), once (1), 2 to 4 times (3), 2 to 3 times per week (10), and 4 times or more per week (22). The quantity question ("In [month], how many drinks did you typically have on a drinking day?") separately asked for the numbers of drinks containing beer (0.25 L), wine or sparkling wine (0.125 L), and spirits (0.04 L). The numbers of drinks were multiplied with their associated amount of pure alcohol (9.5 g/10.9 g/10.5 g) and summed up. A quantity-frequency product was determined, divided by 30.5, and rounded.

Moderators were assessed at baseline. Education was categorized as low, middle, and high levels. Categorization was derived from the assessment of different types of school education in Germany. Participants with 9 or fewer years of schooling were allocated to low education, participants with 10 to 11 years to middle education, and those with 12 or more years to high education. Six participants, reporting to be still in school, were allocated to high education. Employment status was differentiated between employed and unemployed participants. Categorization was derived from the assessment of 2 questions: (1)"Are you currently employed?" with two response options (yes/ no) and (2) among participants who responded "no" were asked which of 6 response options applied (unemployed, pupil, college student, retired, housewife or house-husband, or other). The category "employed" included participants responding "yes" in the first question and participants providing any response other than "unemployed" in the second question to investigate the effect of actual unemployment.

Covariates included gender, age, medical department, self-rated health assessed by the single-item (ie, "Would you say your health in general is: excellent, very good, good, fair, or poor?" [42]). Mental health was assessed by the 5-item Mental Health Inventory [43,44], specifically having a partner (yes, including being married, or no), the number of cigarettes per day, alcohol problem severity assessed by the AUDIT [35], and motivational stage of change measured by a 4-item staging algorithm [16].

Follow-Ups

Follow-ups were conducted between August 2011 and November 2014. All trial participants were followed-up 6, 12, 18, and 24 months after baseline, primarily via computer-assisted telephone interviews. Interviewers were blinded to group allocations; some of them were involved in sample recruitment 12 to 24 months earlier. Incentives were paid before (month 12: self-selected 5€ voucher) or after participation (months 6, 18, and 24: 10, 15, and 20€ voucher, respectively). An average currency exchange rate of €1 = US $1.34 was applicable during this time.

Statistical Analysis

Data were analyzed using Mplus version 7.31(Muthén and Muthén) [45]. Two latent growth models were used to test differential BAI effects on alcohol use per day. Latent growth models afford to reflect nonlinearity and heterogeneity in the outcome growth trajectory and to handle incomplete data properly [46]. In this study, a maximum likelihood estimator with robust standard errors using numerical integration was chosen. Thus, both models were estimated under a missing at random [47] assumption using all available data and including all participants regardless of attrition. Repeated measures of alcohol per day were treated as indicators of latent growth factors that represented the alcohol growth trajectory over 24 months. As data were characterized by a large proportion of zeros with the remaining values being highly positively skewed, alcohol use per day was regressed on the growth factors using a negative binomial model. To handle nonlinearity, the model included 3 growth factors (intercept, linear, and quadratic growth factor). The variance of the quadratic growth factor was fixed to zero.

Interaction terms between the study groups and the two moderator variables (school education and employment status) were included as predictors of the growth factors to test differences in the efficacy of CO and PE. If rescaled likelihood ratio tests indicated significantly improved model fit due to the inclusion of the interaction terms, moderator level-specific net changes in alcohol use were calculated. Net changes were given in incidence rate ratios (IRRs), indicating study group differences in the percentage change in alcohol use per day between baseline and follow-up at 6, 12, 18, and 24 months, respectively. The 24-month follow-up was considered the primary time-point of interest. *P* values below .05 were considered statistically significant. Both analyses were adjusted for all baseline covariates reported above and for the remaining moderator variable.

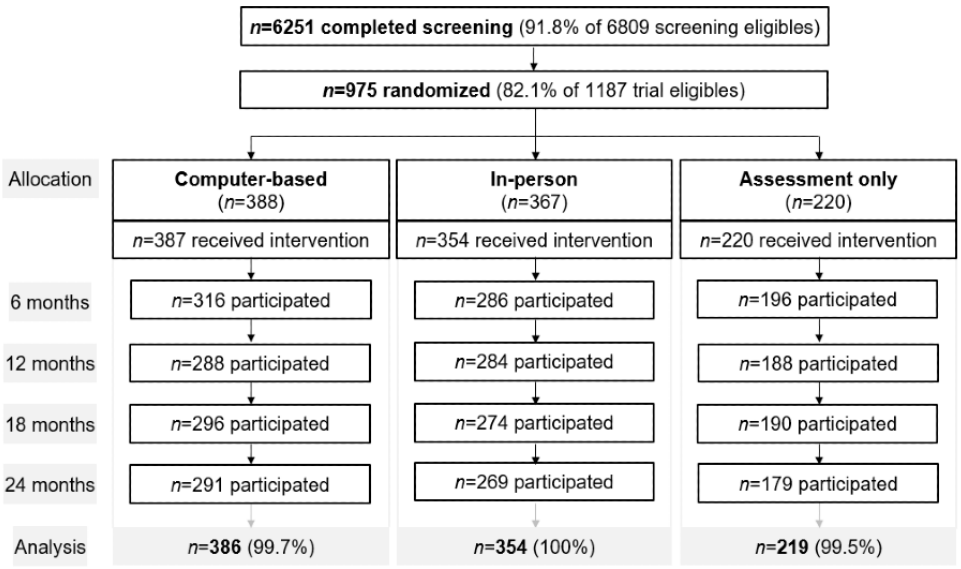
The adjustment for the medical department also took into account potential clustering effects. Different from common cluster-randomized trials, no severe loss of power was expected: (1) all wards provided participants for each study group and (2) with the large number of 140 clusters and the small average number of 7 participants per cluster, only a small design effect (if at all) was expected [48].

Results

Study Sample at Baseline

Of the 6809 patients eligible for screening, 6251 (92%) completed screening (Figure 1). Of the 1187 patients who screened positive for at-risk alcohol use but negative for more severe alcohol problems, 975 (82%) participated in the trial, and 961 (81%) received their allocated intervention. Follow-up participation rates were 83% (798/961) at month 6, 79% (760/961) at month 12, 79% (760/961) at month 18, and 77% (739/961) at month 24. For a detailed CONSORT flow chart, please see elsewhere [16,31]. Two participants (0.2%), 1 with missing baseline covariate data and 1 with unreasonably high alcohol data, were excluded from the analysis.

Figure 1. Participant flow by study group.



As described in more detail elsewhere [16,31], the final sample (N=959) comprised of 719 (75%) men and 240 (25%) women, with a mean age of 40.9 years (SD 14.1). Among the participants, 190 (20%), 532 (55%), and 237 (25%) had low, middle, and high levels of education, respectively. Participants consumed on average 15.2 g of pure alcohol per day (SD 19.8)

at baseline. As depicted in Table 1, a total of 136 (14%) participants were unemployed, and 823 (86%) were employed, also including 96 (12%) retired persons, 61 (7%) college students or pupils, and 41 (41%) others (eg, housewives or house-husbands). Nonparticipants were older and had lower levels of education but did not differ significantly concerning any of the other characteristics [16].

Table 1. Moderator characteristics at baseline stratified by study group (N=959).

Moderators	Computer-based intervention (n=386)	In-person intervention (n=354)	Assessment only (n=219)
Level of education, n (%)			
Low	84 (21.7)	60 (16.9)	46 (21.0)
Middle	211 (54.7)	207 (58.5)	114 (52.1)
High	91 (23.6)	87 (24.6)	59 (26.9)
Employment status, n (%)			
Unemployed	65 (16.8)	37 (10.5)	34 (15.5)
Employed	321 (83.2)	317 (89.5)	185 (84.5)

Moderation Analyses

Rescaled likelihood ratio tests indicated that model fit was not significantly improved by the inclusion of interaction terms between the study group and level of education (*P*=.98). Model fit was significantly improved by including study group x employment status interactions (*P*=.04). These findings are described in more detail.

The effect of CO versus AO by employment status is depicted in Figure 2. Among employed participants, those who received CO reported significantly greater drinking reductions up to month 18 than those who received AO (IRR 0.76, 95% CI 0.58-0.99; *P*=.04). Among unemployed participants, IRRs were comparable but not statistically significant (*P*s≥.27). The efficacy of CO did not differ significantly between employed and unemployed participants (*P*s≥.66; Table 2).

Figure 2. Effects of the computer-based intervention versus assessment only by employment status.

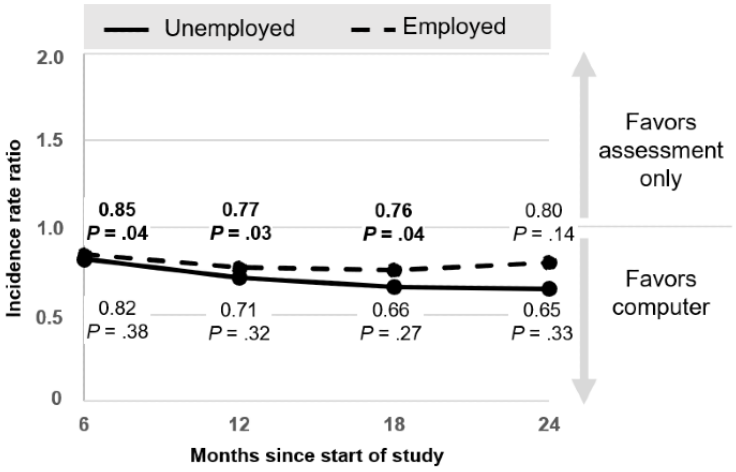


Table 2. Net changes in alcohol use in employed versus unemployed patients (n=959).[a]

	CO[b] versus AO[c]			PE[d] versus AO			PE versus CO		
	IRR[e]	95% CI	*P*	IRR	95% CI	*P*	IRR	95% CI	*P*
Month 0 to 6	0.97	0.60-1.56	.90	0.58	0.35-0.95	.03	0.60	0.38-0.96	.03
Month 0 to 12	0.92	0.45-1.89	.83	0.44	0.21-0.94	.03	0.48	0.24-0.96	.04
Month 0 to 18	0.87	0.39-1.93	.73	0.44	0.18-1.06	.07	0.50	0.23-1.10	.09
Month 0 to 24	0.81	0.32-2.04	.66	0.57	0.19-1.69	.31	0.70	0.27-1.81	.46

[a]Adjusted for gender, age, having a partner, school education, medical department, self-rated health, smoking, alcohol use problem severity, and motivational stage of change.

[b]CO: computer-based intervention.

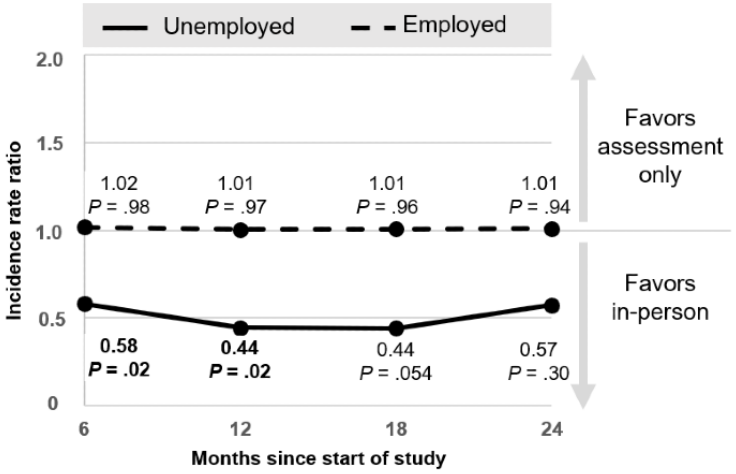
[c]AO: assessment only.

[d]PE: in-person intervention.

[e]IRR: incidence rate ratio.

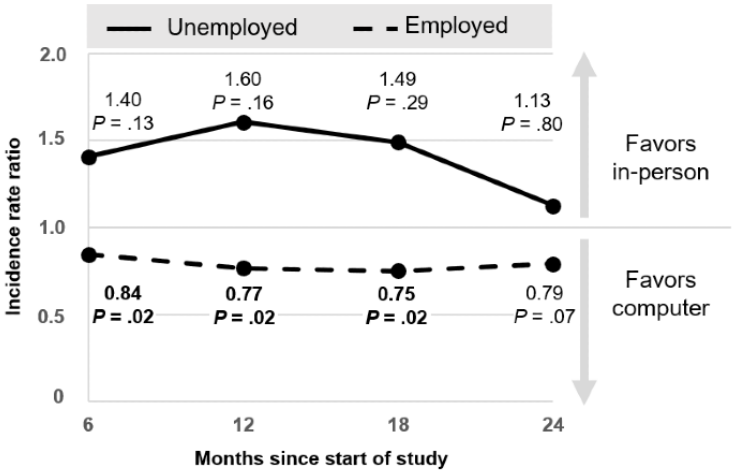
The effect of PE versus AO by employment status is depicted in Figure 3. Among unemployed participants, those who received PE reported significantly greater drinking reductions up to month 12 than those who received AO (IRR=0.44, 95% CI 0.22-0.90; *P*=.02). The difference was marginally significant at month 18 (IRR=0.44, 95% CI 0.19-1.02; *P*=.054) and nonsignificant at month 24 (*P*=.30). Among employed participants, no statistically significant differences were found (*P*s≥.94). As depicted in Table 2, unemployed participants reported significantly greater drinking reductions following PE versus AO than employed participants up to month 12 (IRR 0.44, 95% CI 0.21-0.94; *P*=.03). This difference was marginally significant after 18 months (IRR 0.44, 95% CI 0.18-1.06; *P*=.07) and nonsignificant after 24 months (IRR 0.57, 95% CI 0.19-1.69; *P*=.31).

Figure 3. Effects of the in-person intervention versus assessment only by employment status.



The effect of PE versus CO by employment status is depicted in Figure 4. Among employed participants, those who received CO reported significantly greater drinking reductions up to month 18 than those who received PE (IRR 0.75, 95% CI 0.59-0.95; *P*=.02). The difference was marginally significant at month 24 (IRR 0.79, 95% CI 0.61-1.02; *P*=.07). Among unemployed participants, differences between PE and CO were not statistically significant (*P*s≥.13). As depicted in Table 2, up to month 12, unemployed participants reported significantly greater drinking reductions following PE versus CO than employed participants, while the latter rather benefitted from CO than from PE (IRR 0.48, 95% CI 0.24-0.96; *P*=.04). This difference was marginally significant after 18 months (IRR 0.50, 95% CI 0.23-1.10; *P*=.09) and not significant after 24 months (IRR 0.70, 95% CI 0.27-1.81; *P*=.46).

Figure 4. Comparative effect of computer-based versus in-person intervention by employment status.



Discussion

Overview

This was the first study on the moderating effects of education and employment status on the efficacy and on the comparative efficacy of in-person versus computer-based delivered BAI. It revealed three encouraging findings. Firstly, the efficacy of computer-based BAI was neither moderated by the patients' level of education nor by their employment status. Secondly, in-person BAI had a greater impact on reduced drinking up to month 12 in unemployed versus employed patients. Thirdly, the short-term superiority of in-person BAI over computer-based BAI in unemployed patients and of computer-based BAI over in-person BAI in employed patients was no longer significant after 2 years.

Principal Results and Comparison With Prior Work

The finding that BAI efficacy was not moderated by the level of education is in line with previous reviews showing that once the participants had been recruited, there is no difference in effect [17,21]. While previous studies have often been limited to follow-ups of 12 months or less, our findings demonstrate that comparable efficacy was also observed in the long term. Our findings also add that, although the level of education may not make a difference, other indicators of SES may.

Although after 2 years, we found no differences in efficacy for unemployed versus employed patients, in the first year, the benefits from CO and PE were significantly reversed, indicating that unemployed patients may benefit sooner (ie, within the first year) from in-person delivered BAI, while employed patients may benefit sooner from computer-based feedback. Although these differences attenuate over time, an earlier onset of behavior change may also have other positive consequences for patients, such as earlier reduction of adverse consequences from drinking and earlier improvement of quality of life. Until now, employment status has only rarely been investigated as a moderator of behavior change interventions in general [18].

We may only speculate on why a moderation effect was found for employment status but not for school education. It is possible people with current or particularly heavy strain (as unemployment is likely to be) especially appreciate in-person conversations characterized by compassion, acceptance, partnership, and evocation as transported by motivational interviewing [49], or unemployed people especially appreciate in-time conversations also when they are more time-consuming as they provide the opportunity for answering questions. In contrast, employed people may especially appreciate the independence from the time that may be involved with computer-based feedback. However, in line with other findings on moderating effects as found in this same RCT [50,51], these findings suggest that in-person interventions may not be completely replaceable, particularly for persons with a greater strain who may require in-person rather than computer-based BAI to achieve BAI benefit as soon as possible.

Concerning the question of whether alcohol screening and BAI has at least a neutral social equity impact, the reach of the intervention investigated must also be considered. Although our approach resulted in a significantly lower reach of patients with low levels of education [16], overall reach was satisfying: 81% (961/1187) of the total target population and 79% (723/907) of those with low levels of education had been reached with our recruitment strategy. Lower-effort recruitment results in much larger selection and discrepancies. For example, a large-scale population-based intervention study in Denmark reached 53% of the total target population and 43% of those with low education [52]. With proactive recruitment, as used in our study, the extent of selectivity and discrepancy can be diminished to a great extent but may not be excluded completely. Any self-selection may result in the participation of the "(rather) healthy well-educated," and nonsystematic selection may be driven by socially-unfavorable selection mechanisms, such as stigma. For example, although a population survey in England revealed that general practitioners approached low SES patients twice as likely as high SES patients for BAI, the selection mechanism was highly selective as less than 1 in 10 participants who would have met the eligibility criteria were approached to begin with [53].

In light of all findings on reach and on the moderators of efficacy from this RCT, we may conclude that proactive selection (ie, systematic alcohol screening) and BAI has the potential to have at least a neutral social equity impact. Equity impact may be optimized by providing computer-based BAI to

the vast majority of patients with lower strain (eg, to employed patients) and by providing in-person BAI to the minority of patients with heavier strain (eg, to unemployed patients). To improve the reach of low SES people and to improve the cost-efficiency of BAI, the implementation of screening and BAI in social settings such as job agencies has been found to be promising [54].

Strengths and Limitations

The study provides several strengths. First, the findings are based on a sample of general hospital patients representing 81% (959/1187) of the eligible patients with at-risk alcohol use. Second, the investigation of 2 indicators of SES, including employment status, which has rarely been investigated as a moderator of intervention efficacy [18], provided the opportunity to obtain a more detailed picture of the role different indicators may play in BAI efficacy. Third, the BAIs tested were theory-based, adequately delivered, and intervention retention was high [16]. The finding that intervention retention was particularly high in those receiving computer-based feedback is encouraging and is discussed in more detail elsewhere [16]. Fourth, the 4 follow-ups from 6 to 24 months provided the opportunity to investigate not only short-term changes as usual but also long-term changes by SES groups. Monetary incentives were used to reduce selection bias at follow-ups, resulting in satisfactory follow-up participation of 77%-83%. It appears unlikely that incentives have distorted study results as they were provided to participants at follow-ups only, independent of the study group, individual intervention retention, and behavior change. And fifth, latent growth modeling allowed the capture of individual differences in change over 5 time points to depict nonlinear trajectories of change and include all baseline participants in the analysis, regardless of their adherence to intervention or follow-up.

Several limitations are to be noted. First, it must be acknowledged that the RCT was powered to detect treatment effects in the total sample rather than differential treatment effects between subgroups. Therefore, potential effects did not reach statistical significance. This was particularly obvious concerning the small group of 136 unemployed participants. Second, as applies to most eHealth and BAI trials, findings are based on self-report and may be biased in terms of recall and social desirability. We cannot rule out that, as a result of receiving more attention, intervention participants responded in a more socially desirable way than assessment-only participants [55]. However, alcohol self-reports offer a minorly invasive and low-cost way of obtaining alcohol use data with acceptable validity [56], particularly among persons without severe alcohol problems, as targeted in our study [57]. Third, as also applies to most eHealth trials, participants were not blinded. Fourth, findings may be limited to those patients who agree to participate in an intervention study. Although overall reach was high, including among patients with low levels of education, nonparticipants had lower education levels and were older compared to participants [16]. The analyses were adjusted for education levels and age to account for the potential effects of these characteristics. Fifth, the generalizability of our findings may be limited to proactively recruited populations and may not apply to convenience samples given different initial characteristics in terms of problem severity and motivation to change [58].

Conclusions

To advance the development of behavior change interventions with public health and equity impact, we, as intervention researchers, are asked to put social equity impact [5] into focus in addition to the impact of interventions on the behavioral level. To identify whether certain vulnerable members of the population benefit more or less from one or the other way of delivery, we critically investigated computer-based and in-person delivered BAIs that showed not only positive effects on reduced alcohol use but also long-term effects on health in the total sample over 2 years. The findings are encouraging with respect to reach and efficacy independent of education levels. But the study also identified that the small subgroup of unemployed patients might benefit sooner from BAI when delivered in person. These findings also highlight that, in the future, differences in intervention reach (and retention, if applicable) and efficacy or effectiveness by indicators of SES should not only be reported as descriptive measures (although it would be a good starting point) but should rather be treated as core outcome measures of behavior change interventions.

Acknowledgments

JFA, BG, and UJ received funding from the German Cancer Aid to conduct the randomized controlled trial and to prepare the paper (grant numbers 108376, 109737, 110676, 110543, 111346, and 70110543). Statistical analysis was supported by funding from the German Research Foundation provided to SB (grant numbers BA 5858/2-1 and BA 5858/2-3). The funders had no role in study design; collection, analysis and interpretation of data; writing the report; and the decision to submit the report for publication.

We acknowledge support for the Article Processing Charge from the DFG (German Research Foundation, grant number 393148499) and the Open Access Publication Fund of the University of Greifswald.

Authors' Contributions

All authors made substantial contributions to the conception and design of the study (JFA, SB, BG, and UJ), or acquisition of data (SB and CG), or analysis and interpretation of data (JFA, SB, GB, AS, BG, and UJ), including drafting the article (JFA and SB) or critically revising it for important intellectual content (GB, CG, AS, BG, and UJ). All authors granted final approval of the version to be published.

Conflicts of Interest

JFA and GB are members of the Motivational Interviewing Network of Trainers. The authors have no financial conflicts of interest to disclose.

Multimedia Appendix 1
CONSORT eHEALTH Checklist (V 1.6.1).
[PDF File (Adobe PDF File), 464 KB - mental_v9i1e31712_app1.pdf]

References

1. Stringhini S, Carmeli C, Jokela M, Avendaño M, Muennig P, Guida F, et al. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. The Lancet 2017 Mar;389(10075):1229-1237. [doi: 10.1016/s0140-6736(16)32380-7]
2. Chetty R, Stepner M, Abraham S, Lin S, Scuderi B, Turner N, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. JAMA 2016 Apr 26;315(16):1750-1766 [FREE Full text] [doi: 10.1001/jama.2016.4226] [Medline: 27063997]
3. Mackenbach J, Kulhánová I, Menvielle G, Bopp M, Borrell C, Costa G, Eurothine and EURO-GBD-SE consortiums. Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 European countries. J Epidemiol Community Health 2015 Mar;69(3):207-17; discussion 205. [doi: 10.1136/jech-2014-204319] [Medline: 24964740]
4. Olshansky SJ, Antonucci T, Berkman L, Binstock RH, Boersch-Supan A, Cacioppo JT, et al. Differences in life expectancy due to race and educational differences are widening, and many may not catch up. Health Aff (Millwood) 2012 Aug;31(8):1803-1813. [doi: 10.1377/hlthaff.2011.0746] [Medline: 22869659]
5. Mackenbach J, Kulhánová I, Artnik B, Bopp M, Borrell C, Clemens T, et al. Changes in mortality inequalities over two decades: register based study of European countries. BMJ 2016 Apr 11;353:i1732 [FREE Full text] [doi: 10.1136/bmj.i1732] [Medline: 27067249]
6. Probst C, Roerecke M, Behrendt S, Rehm J. Socioeconomic differences in alcohol-attributable mortality compared with all-cause mortality: a systematic review and meta-analysis. Int. J. Epidemiol 2014 Mar 11;43(4):1314-1327 [FREE Full text] [doi: 10.1093/ije/dyu043] [Medline: 24618188]
7. Mackenbach J, Kulhánová I, Bopp M, Borrell C, Deboosere P, Kovács K, et al. Inequalities in Alcohol-Related Mortality in 17 European Countries: A Retrospective Analysis of Mortality Registers. PLoS Med 2015 Dec;12(12):e1001909 [FREE Full text] [doi: 10.1371/journal.pmed.1001909] [Medline: 26625134]
8. Katikireddi SV, Whitley E, Lewsey J, Gray L, Leyland AH. Socioeconomic status as an effect modifier of alcohol consumption and harm: analysis of linked cohort data. The Lancet Public Health 2017 Jun;2(6):e267-e276 [FREE Full text] [doi: 10.1016/S2468-2667(17)30078-6] [Medline: 28626829]
9. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
10. Kaner EF, Beyer F, Muirhead C, Campbell F, Pienaar E, Bertholet N, et al. Effectiveness of brief alcohol interventions in primary care populations. Cochrane Database Syst Rev 2018 Feb 24;2:CD004148 [FREE Full text] [doi: 10.1002/14651858.CD004148.pub4] [Medline: 29476653]
11. Frost H, Campbell P, Maxwell M, O'Carroll RE, Dombrowski SU, Williams B, et al. Effectiveness of Motivational Interviewing on adult behaviour change in health and social care settings: A systematic review of reviews. PLoS One 2018 Oct 18;13(10):e0204890 [FREE Full text] [doi: 10.1371/journal.pone.0204890] [Medline: 30335780]
12. O'Connor E, Perdue L, Senger C, Rushkin M, Patnode C, Bean S, et al. Screening and Behavioral Counseling Interventions to Reduce Unhealthy Alcohol Use in Adolescents and Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA 2018 Nov 13;320(18):1910-1928. [doi: 10.1001/jama.2018.12086] [Medline: 30422198]
13. Donoghue K, Patton R, Phillips T, Deluca P, Drummond C. The effectiveness of electronic screening and brief intervention for reducing levels of alcohol consumption: a systematic review and meta-analysis. J Med Internet Res 2014 Jun 02;16(6):e142 [FREE Full text] [doi: 10.2196/jmir.3193] [Medline: 24892426]
14. Álvarez-Bueno C, Rodríguez-Martín B, García-Ortiz L, Gómez-Marcos MA, Martínez-Vizcaíno V. Effectiveness of brief interventions in primary health care settings to decrease alcohol consumption by adult non-dependent drinkers: a systematic review of systematic reviews. Prev Med 2015 Jul;76 Suppl:S33-S38. [doi: 10.1016/j.ypmed.2014.12.010] [Medline: 25514547]
15. Mdege ND, Fayter D, Watson J, Stirk L, Sowden A, Godfrey C. Interventions for reducing alcohol consumption among general hospital inpatient heavy alcohol users: a systematic review. Drug Alcohol Depend 2013 Jul 01;131(1-2):1-22. [doi: 10.1016/j.drugalcdep.2013.01.023] [Medline: 23474201]
16. Freyer-Adam J, Baumann S, Haberecht K, Tobschall S, Schnuerer I, Bruss K, et al. In-person and computer-based alcohol interventions at general hospitals: reach and retention. Eur J Public Health 2016 Oct;26(5):844-849. [doi: 10.1093/eurpub/ckv238] [Medline: 26748101]

17. Littlejohn C. Does socio-economic status influence the acceptability of, attendance for, and outcome of, screening and brief interventions for alcohol misuse: a review. Alcohol Alcohol 2006;41(5):540-545. [doi: 10.1093/alcalc/agl053] [Medline: 16855002]
18. Alcántara C, Diaz SV, Cosenzo LG, Loucks EB, Penedo FJ, Williams NJ. Social determinants as moderators of the effectiveness of health behavior change interventions: scientific gaps and opportunities. Health Psychology Review 2020 Feb 12;14(1):132-144. [doi: 10.1080/17437199.2020.1718527]
19. Paz Castro R, Haug S, Kowatsch T, Filler A, Schaub MP. Moderators of outcome in a technology-based intervention to prevent and reduce problem drinking among adolescents. Addictive Behaviors 2017 Sep;72:64-71. [doi: 10.1016/j.addbeh.2017.03.013]
20. Riper H, Kramer J, Keuken M, Smit F, Schippers G, Cuijpers P. Predicting Successful Treatment Outcome of Web-Based Self-help for Problem Drinkers: Secondary Analysis From a Randomized Controlled Trial. J Med Internet Res 2008 Nov 22;10(4):e46. [doi: 10.2196/jmir.1102]
21. Riper H, Hoogendoorn A, Cuijpers P, Karyotaki E, Boumparis N, Mira A, et al. Effectiveness and treatment moderators of internet interventions for adult problem drinking: An individual patient data meta-analysis of 19 randomised controlled trials. PLoS Med 2018 Dec 18;15(12):e1002714. [doi: 10.1371/journal.pmed.1002714]
22. Beyer F, Lynch E, Kaner E. Brief Interventions in Primary Care: an Evidence Overview of Practitioner and Digital Intervention Programmes. Curr Addict Rep 2018;5(2):265-273 [FREE Full text] [doi: 10.1007/s40429-018-0198-7] [Medline: 29963364]
23. Nair NK, Newton NC, Shakeshaft A, Wallace P, Teesson M. A Systematic Review of Digital and Computer-Based Alcohol Intervention Programs in Primary Care. Curr Drug Abuse Rev 2015 Sep 28;8(2):111-118. [doi: 10.2174/1874473708666150916113538] [Medline: 26373848]
24. Ramsey AT, Satterfield JM, Gerke DR, Proctor EK. Technology-Based Alcohol Interventions in Primary Care: Systematic Review. J Med Internet Res 2019 Apr 08;21(4):e10859 [FREE Full text] [doi: 10.2196/10859] [Medline: 30958270]
25. Dedert EA, McDuffie JR, Stein R, McNiel JM, Kosinski AS, Freiermuth CE, et al. Electronic Interventions for Alcohol Misuse and Alcohol Use Disorders: A Systematic Review. Ann Intern Med 2015 Aug 04;163(3):205-214 [FREE Full text] [doi: 10.7326/M15-0285] [Medline: 26237752]
26. Kaner EF, Beyer FR, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev 2017 Sep 25;9:CD011479 [FREE Full text] [doi: 10.1002/14651858.CD011479.pub2] [Medline: 28944453]
27. Sundström C, Blankers M, Khadjesari Z. Computer-Based Interventions for Problematic Alcohol Use: a Review of Systematic Reviews. Int J Behav Med 2017 Oct;24(5):646-658 [FREE Full text] [doi: 10.1007/s12529-016-9601-8] [Medline: 27757844]
28. Tansil K, Esser M, Sandhu P, Reynolds J, Elder R, Williamson R, Community Preventive Services Task Force. Alcohol Electronic Screening and Brief Intervention: A Community Guide Systematic Review. Am J Prev Med 2016 Nov;51(5):801-811 [FREE Full text] [doi: 10.1016/j.amepre.2016.04.013] [Medline: 27745678]
29. Freyer-Adam J, Baumann S, Bischof G, John U, Gaertner B. Sick days in general hospital patients two years after brief alcohol intervention: Secondary outcomes from a randomized controlled trial. Preventive Medicine 2020 Oct;139:106106. [doi: 10.1016/j.ypmed.2020.106106]
30. Freyer-Adam J, Baumann S, Haberecht K, Bischof G, Meyer C, Rumpf H, et al. Can brief alcohol interventions in general hospital inpatients improve mental and general health over 2 years? Results from a randomized controlled trial. Psychol. Med 2018 Sep 04;49(10):1722-1730. [doi: 10.1017/s0033291718002453]
31. Freyer-Adam J, Baumann S, Haberecht K, Tobschall S, Bischof G, John U, et al. In-person alcohol counseling versus computer-generated feedback: Results from a randomized controlled trial. Health Psychol 2018 Jan;37(1):70-80. [doi: 10.1037/hea0000556] [Medline: 28967769]
32. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Alcohol Use Disorders Identification Test. Arch Intern Med 1998 Sep 14;158(16):1789-1795. [doi: 10.1001/archinte.158.16.1789] [Medline: 9738608]
33. Reinert DF, Allen JP. The alcohol use disorders identification test: an update of research findings. Alcohol Clin Exp Res 2007 Feb;31(2):185-199. [doi: 10.1111/j.1530-0277.2006.00295.x] [Medline: 17250609]
34. Donovan D, Kivlahan D, Doyle S, Longabaugh R, Greenfield S. Concurrent validity of the Alcohol Use Disorders Identification Test (AUDIT) and AUDIT zones in defining levels of severity among out-patients with alcohol dependence in the COMBINE study. Addiction 2006 Dec;101(12):1696-1704. [doi: 10.1111/j.1360-0443.2006.01606.x] [Medline: 17156168]
35. Saunders J, Aasland O, Babor T, de la Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption--II. Addiction 1993 Jun;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
36. Freyer-Adam J, Baumann S, Schnuerer I, Haberecht K, John U, Gaertner B. Persönliche vs. computerbasierte Alkoholintervention für Krankenhauspatienten: Studiendesign. SUCHT 2015 Dec;61(6):347-355. [doi: 10.1024/0939-5911.a000394]

37. Bischof G, Reinhardt S, Grothues J, John U, Rumpf HJ. The Expert Test and Report on Alcohol (EXTRA): Development and evaluation of a computerized software program for problem drinkers. In: D. R. B, editor. New Research on Alcoholism. New York: Nova Science Publishers, Incorporated; Nov 2007:155-177.
38. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997 Sep;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
39. Baumann S, Gaertner B, Schnuerer I, Bischof G, John U, Freyer-Adam J. How well do TTM measures work among a sample of individuals with unhealthy alcohol use that is characterized by low readiness to change? Psychol Addict Behav 2013 Sep;27(3):573-583. [doi: 10.1037/a0029368] [Medline: 22867296]
40. Seitz HK, Bühringer G, Mann K. Empfehlungen des wissenschaftlichen Kuratoriums der DHS. In: Deutsche Hauptstelle für Suchtfragen , editor. Jahrbuch Sucht 2008. Geesthacht: Neuland; 2008:205-209.
41. Miller WR, Rollnick S. Motivational Interviewing. Preparing people for change. New York, NY: The Guilford Press; 2002.
42. Idler EL, Benyamini Y. Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. Journal of Health and Social Behavior 1997 Mar;38(1):21. [doi: 10.2307/2955359]
43. Berwick DM, Murphy JM, Goldman PA, Ware JE, Barsky AJ, Weinstein MC. Performance of a five-item mental health screening test. Med Care 1991 Feb;29(2):169-176. [doi: 10.1097/00005650-199102000-00008] [Medline: 1994148]
44. Rumpf HJ, Meyer C, Hapke U, John U. Screening for mental health: validity of the MHI-5 using DSM-IV Axis I psychiatric disorders as gold standard. Psychiatry Research 2001 Dec;105(3):243-253. [doi: 10.1016/s0165-1781(01)00329-8]
45. Muthén LK, Muthén BO. Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén; 2012.
46. Preacher KJ. Latent growth curve models. In: Hancok GR, Mueller RO. eds. The reviewer's guide to quantitative methods in the social sciences. New York, NY: Taylor & Francis; 2010.
47. Little RJ, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Jon Wiley & Sons; 2002.
48. Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Ann Fam Med 2004 May 01;2(3):204-208 [FREE Full text] [doi: 10.1370/afm.141] [Medline: 15209195]
49. Miller WR, Rollnick S. Motivational Interviewing. Helping people change. New York, NY: The Guilford Press; 2013.
50. Baumann S, Gaertner B, Haberecht K, Bischof G, John U, Freyer-Adam J. How alcohol use problem severity affects the outcome of brief intervention delivered in-person versus through computer-generated feedback letters. Drug Alcohol Depend 2018 Feb 01;183:82-88. [doi: 10.1016/j.drugalcdep.2017.10.032] [Medline: 29241105]
51. Baumann S, Gaertner B, Haberecht K, Meyer C, Rumpf HJ, John U, et al. Does impaired mental health interfere with the outcome of brief alcohol intervention at general hospitals? J Consult Clin Psychol 2017 Jun;85(6):562-573. [doi: 10.1037/ccp0000201] [Medline: 28333511]
52. Bender AM, Jørgensen T, Helbech B, Linneberg A, Pisinger C. Socioeconomic position and participation in baseline and follow-up visits: the Inter99 study. Eur J Prev Cardiol 2014 Jul 11;21(7):899-905. [doi: 10.1177/2047487312472076] [Medline: 23233551]
53. Angus C, Brown J, Beard E, Gillespie D, Buykx P, Kaner E, et al. Socioeconomic inequalities in the delivery of brief interventions for smoking and excessive drinking: findings from a cross-sectional household survey in England. BMJ Open 2019 May 01;9(4):e023448 [FREE Full text] [doi: 10.1136/bmjopen-2018-023448] [Medline: 31048422]
54. Freyer-Adam J, Baumann S, Schnuerer I, Haberecht K, Bischof G, John U, et al. Does stage tailoring matter in brief alcohol interventions for job-seekers? A randomized controlled trial. Addiction 2014 Nov 01;109(11):1845-1856. [doi: 10.1111/add.12677] [Medline: 24981701]
55. Saitz R. The best evidence for alcohol screening and brief intervention in primary care supports efficacy, at best, not effectiveness: you say tomāto, I say tomăto? That's not all it's about. Addict Sci Clin Pract 2014 Aug 27;9:14 [FREE Full text] [doi: 10.1186/1940-0640-9-14] [Medline: 25168288]
56. Del Boca FK, Darkes J. The validity of self-reports of alcohol consumption: state of the science and challenges for research. Addiction 2003 Dec;98 Suppl 2:1-12. [doi: 10.1046/j.1359-6357.2003.00586.x] [Medline: 14984237]
57. Babor TF, Steinberg K, Anton R, Del Boca F. Talk is cheap: measuring drinking outcomes in clinical trials. J Stud Alcohol 2000 Jan;61(1):55-63. [doi: 10.15288/jsa.2000.61.55] [Medline: 10627097]
58. Prochaska JO. Multiple Health Behavior Research represents the future of preventive medicine. Prev Med 2008 Mar;46(3):281-285. [doi: 10.1016/j.ypmed.2008.01.015] [Medline: 18319100]

Abbreviations

AO: assessment only
AUDIT: Alcohol Use Disorder Identification Test
BAI: brief alcohol intervention
CO: computer-based intervention
IRR: incidence rate ratio
PE: in-person intervention
RCT: randomized controlled trial
SES: socioeconomic status

Edited by J Torous; submitted 02.07.21; peer-reviewed by J Bruthans, Y Yu; comments to author 08.11.21; revised version received 12.11.21; accepted 12.11.21; published 28.01.22.

Please cite as:
Freyer-Adam J, Baumann S, Bischof G, Staudt A, Goeze C, Gaertner B, John U
Social Equity in the Efficacy of Computer-Based and In-Person Brief Alcohol Interventions Among General Hospital Patients With At-Risk Alcohol Use: A Randomized Controlled Trial
JMIR Ment Health 2022;9(1):e31712
URL: https://mental.jmir.org/2022/1/e31712
doi:10.2196/31712
PMID:35089156

Original Paper

Problematic Internet Use Before and During the COVID-19 Pandemic in Youth in Outpatient Mental Health Treatment: App-Based Ecological Momentary Assessment Study

Meredith Gansner[1], BA, MD; Melanie Nisenson[1], BA, MSc; Vanessa Lin[1], BA, MSc; Sovannarath Pong[1], BA, MSc; John Torous[2], BSc, MD, MBI; Nicholas Carson[1], BA, MD

[1]Department of Psychiatry, Cambridge Health Alliance, Cambridge, MA, United States

[2]Department of Digital Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, United States

Corresponding Author:
Meredith Gansner, BA, MD
Department of Psychiatry
Cambridge Health Alliance
1493 Cambridge Street
Cambridge, MA, 02139
United States
Phone: 1 617 575 5498
Email: meredith.gansner@gmail.com

Abstract

Background: Youth with existing psychiatric illness are more apt to use the internet as a coping skill. Because many "in-person" coping skills were not easily accessible during the COVID-19 pandemic, youth in outpatient mental health treatment may have been particularly vulnerable to the development of problematic internet use (PIU). The identification of a pandemic-associated worsening of PIU in this population is critical in order to guide clinical care; if these youth have become dependent upon the internet to regulate their negative emotions, PIU must be addressed as part of mental health treatment. However, many existing studies of youth digital media use in the pandemic do not include youth in psychiatric treatment or are reliant upon cross-sectional methodology and self-report measures of digital media use.

Objective: This is a retrospective cohort study that used data collected from an app-based ecological momentary assessment protocol to examine potential pandemic-associated changes in digital media youth in outpatient mental health treatment. Secondary analyses assessed for differences in digital media use dependent upon personal and familial COVID-19 exposure and familial hospitalization, as well as factors associated with PIU in this population.

Methods: The participants were aged 12-23 years and were receiving mental health treatment in an outpatient community hospital setting. All participants completed a 6-week daily ecological momentary assessment protocol on their personal smartphones. Questions were asked about depression (PHQ-8 [8-item Patient Health Questionnaire]), anxiety (GAD-7 [7-item General Anxiety Disorder]), PIU (PIU-SF-6 [Problematic Internet Use Short Form 6]), digital media use based on Apple's daily screen time reports, and personal and familial COVID-19 exposure. The analyses compared screen time, psychiatric symptoms, and PIU between cohorts, as well as between youth with personal or familial COVID-19 exposures and those without. The analyses also assessed for demographic and psychiatric factors associated with clinically significant PIU-SF-6 scores.

Results: A total of 69 participants completed the study. The participants recruited during the pandemic were significantly more likely to meet the criteria for PIU based on their average PIU-SF-6 score (P=.02) and to spend more time using social media each day (P=.049). The overall amount of daily screen time did not differ between cohorts. Secondary analyses revealed a significant increase in average daily screen time among subjects who were exposed to COVID-19 (P=.01). Youth with clinically significant PIU-SF-6 scores were younger and more likely to have higher PHQ-8 (P=.003) and GAD-7 (P=.003) scores. No differences in scale scores or media use were found between subjects based on familial COVID-19 exposure or hospitalization.

Conclusions: Our findings support our hypothesis that PIU may have worsened for youth in mental health treatment during the pandemic, particularly the problematic use of social media. Mental health clinicians should incorporate screening for PIU into routine clinical care in order to prevent potential familial conflict and subsequent psychiatric crises that might stem from unrecognized PIU.

(JMIR Ment Health 2022;9(1):e33114) doi:10.2196/33114

KEYWORDS

COVID-19; problematic internet use; ecological momentary assessment; internet; app; youth; young adult; teenager; outpatient; mental health; treatment; pilot; cohort; change

Introduction

Significant concerns exist that youth mental health worsened during the COVID-19 pandemic. While the US emergency department visits for pediatric ailments such as asthma or otitis media decreased during the pandemic, the proportion of youth presentations related to mental health crises increased in 2020 [1]. Specifically, more recent national data show that emergency department visits for suicidal ideation increased for youth aged 12-17 years, especially for adolescent girls [2]. International data also support these concerns, including a meta-analysis of 29 studies that demonstrated increased levels of youth anxiety and depression during the pandemic [3].

Researchers leading these studies have been careful to note that their study designs do not allow for the determination of causality. The pandemic has not been proven to be the direct cause of worsening psychiatric illness, despite growing evidence that pandemic restrictions likely had a significant impact on youth mental health. News outlets have featured stories from youth explicitly stating that pandemic-associated stressors such as online schooling and cancelled extracurricular activities led to a worsening of anxiety or depression [4]. Moreover, access to some mental health services, such as those available through schools or other community-based supports, also became limited during the early months of the pandemic [5]. The compounding of these 2 factors may have created an opportune environment for an additional influencer of youth mental health, that of problematic screen time.

Elevated daily screen time and problematic internet use (PIU), an excessive, uncontrollable drive to continue use of the internet despite negative consequences, are both well-associated with numerous psychiatric comorbidities, including depression, anxiety, substance use, self-injurious behavior, and suicidality [6-9]. While increased levels of screen time were a recognized consequence of the pandemic for individuals across the developmental lifespan, youth are the largest consumers of digital media and the most likely population to develop PIU. Thus, it is potentially unsurprising that emerging studies have identified a comorbid increase in youth screen time and severity of psychiatric symptoms during the pandemic [10,11].

However, not all youth appear equally susceptible to PIU and the negative effects of screen time. Youth with existing psychiatric illness, for example, may be especially vulnerable to PIU; our prior longitudinal studies in this specific population have highlighted momentary negative correlations between cell phone engagement, PIU, and mood symptom severity, suggesting that these youth use digital media to relieve psychiatric symptoms, subsequently risking PIU development [12,13]. Therefore, youth in mental health treatment may have developed a particularly complicated relationship with digital media as a result of the COVID-19 pandemic.

This study assessed the digital media habits of 2 separate cohorts of youth (1 before and 1 during the COVID-19 pandemic) who were receiving mental health treatment in a single community health setting. Data were obtained from an existing ecological momentary assessment (EMA) smartphone protocol that collected daily information about a participant's qualitative digital media use, PIU, and symptoms of anxiety and depression over a 6-week period. Through the examination of these collected data, our study (1) assessed how digital media use and mental health may have changed for youth in mental health treatment during the pandemic and (2) explored how personal or familial exposure to the novel coronavirus might have impacted digital media habits and mental health. Due to more limited access to nondigital coping skills during the pandemic, we hypothesized that youth in the pandemic cohort would have higher rates of PIU and spend more time on screens and social media. We also hypothesized that within the COVID-19 cohort, youth personally exposed to COVID-19 might have significantly higher amounts of daily screen and social media time due to increased awareness of the disease and subsequent avoidance of in-person pastimes.

The clinical implications of this study are significant. If these high-risk youth developed a more pathological relationship with digital media during the pandemic, they may have a particularly difficult time separating from digital devices when COVID-19 restrictions are eventually rolled back in favor of in-person activities and services. Because forced separation from devices is often a trigger for parent-child conflict and can precipitate a psychiatric crisis [14], mental health professionals need to be aware of this increased risk to their patients and be prepared to help parents and guardians safely facilitate device separation.

Methods

Participants

The study participants were initially recruited as part of a separate app-based EMA pilot study investigating PIU in this population [12] and were all patients of outpatient mental health clinics within the network of a large community-based hospital in the greater Boston area. The participants were eligible for this separate EMA study if they were between 12 and 23 years old at the beginning of the study and owned a personal smartphone. If a potential participant was under 18 years of age, an informed consent was obtained from the parent or guardian. The participants were excluded if parental or guardian consent was not obtained (if <18 years old) or if they were unable to read English at a 6th grade level (due to lack of app availability in languages other than English). The pre–COVID-19 cohort was passively recruited at the clinics through posted fliers and actively recruited via referral to the study team from the participant's mental health care provider. All participants referred from providers assented to the referral. For the COVID-19 sample, the participants were actively recruited by

the study team through the hospital's electronic health record (EHR) system. Notably, study recruitment was paused temporarily the day after the state's declaration of emergency due to COVID-19 on March 10, 2020, due to the requisite need to switch to remote recruitment methods only. Recruitment began again in September 2020 once the Institutional Review Board approval was granted for remote study recruitment and changes were made in the protocol to include questions about COVID-19 exposure. For these analyses, the participants were retroactively categorized into pre–COVID-19 and COVID-19 cohorts based on whether they were recruited before or after the halt in recruitment on March 11, 2020. The participants were compensated with a $25 Amazon gift card at the beginning and end of the study period. Compensation was not dependent on the level of engagement with the app. All parts of this study were reviewed and approved by the hospital's Institutional Review Board and conformed to the latest version of the Declaration of Helsinki.

Procedure

Data for this study had previously been collected by these authors as part of a separate app-based EMA pilot protocol that used mindLAMP for daily assessment and data collection over a period of 6 weeks. MindLAMP is a free-rein research platform that includes both an online portal system and a smartphone app [15]. All study participants downloaded the mindLAMP app onto their smartphones prior to the start of their study period. The participants were reminded to complete daily surveys via a push notification sent via the app. A time of day was selected so that the participants would likely be at home, allowing for more privacy to complete the surveys.

The surveys included 3 clinical scales to measure PIU (PIU-SF-6 [Problematic Internet Use Short Form 6]), depression (PHQ-8 [8-item Patient Health Questionnaire]), and anxiety (GAD-7 [7-item General Anxiety Disorder]). The PHQ-8 is a modified version of the PHQ-9, which omits the final question assessing suicidality due to the fact that positive responses could not be actively monitored remotely. The PIU-SF-6 is a scale validated for the measurement of PIU in youth (=.77) [16]. Wording of the 3 scales was also adjusted to account for their being administered on a daily basis. It does not appear that daily administration impacts scale validity [17]. Each participant who owned an iPhone was asked to input the following information provided daily by the Apple screen time report feature of iOS: total screen time, total time on social media, and top 3 apps used that day. As part of the screen time feature, daily time spent on social media is automatically identified, categorized, and calculated. Media that are considered social media include both website browser and app visits to sites such as Facebook, WhatsApp, Instagram, or Apple Messenger. Screen time reporting for Android was not available at the time the initial study protocol was approved; however, the majority of the study participants had iPhones. Each time the participant connected to Wi-Fi, mindLAMP uploaded de-identified survey data to a secure server compliant with the Health Insurance Portability and Accountability Act of 1996.

Participant psychiatric diagnoses were obtained from the most recent mental health visit notes in the participant's EHR. Youth completing the study during the COVID-19 pandemic were asked to respond to a brief survey regarding their personal and familial exposure to the novel coronavirus at the beginning and end of the study period. For the purpose of this study, family was defined as whomever the youth considered to be family, not just those individuals living with the participant.

Processing and Analysis

Data were downloaded from mindLAMP in the form of daily scale scores, screen and social media time, and the 3 most commonly used apps on a daily basis. For each participant, daily PIU-SF-6, PHQ-8, and GAD-7 scale scores were transformed into average scale scores. This average scale score was then used to create a secondary binary variable, which described if a participant's average score met or exceeded standardized cut-off values when screening for clinical illness. For the PIU-SF-6, this threshold was a score of ≥15 [16], and ≥10 for the PHQ-8 and GAD-7 scales. Gender was defined as the current gender identity at study enrollment. Participant diagnoses obtained from the EHR were transformed into 2 binary variables: the presence of an anxiety disorder ("yes" or "no") and the presence of a depressive disorder ("yes" or "no") to compare preexisting diagnoses across samples. Average daily times (in minutes) spent using a smartphone or social media were calculated for each participant.

Addressing the study's first goal, logistic regressions compared the number of participants whose average scale scores met clinical cut-off values for the PHQ-8, GAD-7, and PIU-SF-6 across pre–COVID-19 and COVID-19 cohorts. While the sample size of the study is a notable limitation for this model, logistic regression models were chosen for this analysis in order to adjust for confounders of age and gender. Age is a known confounder positively associated with both youth screen time and social media use; therefore, the significant difference in age between pre–COVID-19 and COVID-19 cohorts needed to be considered in this first analysis. Due to the nonnormality of the data, Mann-Whitney tests were performed to assess for differences in mean daily screen and social media times across these groups. To correct for age differences when comparing screen and social media times across cohorts, these specific data underwent intercept adjustment for age and gender prior to conducting Mann-Whitney tests.

A second set of analyses assessed for differences in screen and social media times as well as average scale scores based on COVID-19 exposure within the COVID-19 cohort alone. These analyses used the Fisher exact and Mann-Whitney tests due to the small sample size. Given the unique stress of the pandemic upon minority populations, we also assessed whether youth of color in the COVID-19 cohort had significantly different digital media use compared with nonminority youth.

Finally, we sought to characterize our sample with PIU comparing participants with average PIU-SF-6 scores meeting the clinical cut-off score of ≥15 to those with scores of <15. We assessed associations between age, gender, minority status, and likelihood of having clinically significant GAD-7 and PHQ-8 scores and preexisting anxiety or depressive disorder diagnoses using the Fisher exact tests. As mentioned, Mann-Whitney tests assessed for differences in age and average daily screen or social

media time. All analyses were performed using Stata (version 1.2.5033, Stata Corp) [18] and Rstudio (version 1.2.5033, RStudio Inc) [19].

Results

A total of 69 participants completed the 6-week study. Symptom scale data were obtained from all 69 participants, and 77% (n=53) of the participants had iPhones and provided information about their smartphone use from daily screen time reports. Demographic information is summarized in Table 1. While the participants in the COVID-19 sample were significantly older than youth in the pre–COVID-19 sample, there were no differences in gender or prevalence of preexisting anxiety or depressive disorders between the groups.

Table 1. Participant demographic information.

Characteristics	Values		
	Pre–COVID-19	COVID-19	*P* value
Total	27	42	N/A[a]
Age (years), mean (SD)	15.30 (2.74)	16.95 (1.94)	.003
Gender, n (%)			.07
Male	13 (48.1)	10 (23.8)	
Female	14 (51.9)	32 (76.2)	
Race, n (%)			.24
White	14 (51.9)	21 (50.0)	
Hispanic or Latinx	7 (25.9)	10 (23.8)	
Black	5 (18.5)	10 (23.8)	
Asian	1 (3.7)	1 (2.4)	

[a]N/A: not applicable.

Controlling for age and gender, youth in our COVID-19 cohort were more likely to meet criteria for PIU based on their PIU-SF-6 scores averaged over the 6-week study (*P*=.02) (Table 2). The averaged PHQ-8 and GAD-7 scale scores were also higher in the COVID-19 cohort, but these increases did not reach statistical significance. Social media apps were the most popular type of app used both prior to and during the pandemic (Figure 1); however, again controlling for age and gender, the amount of time spent daily on social media was significantly higher in those youth who completed the study during the pandemic (*P*=.049).

Figure 1. Most frequently used types of apps based on Apple screen time reports.

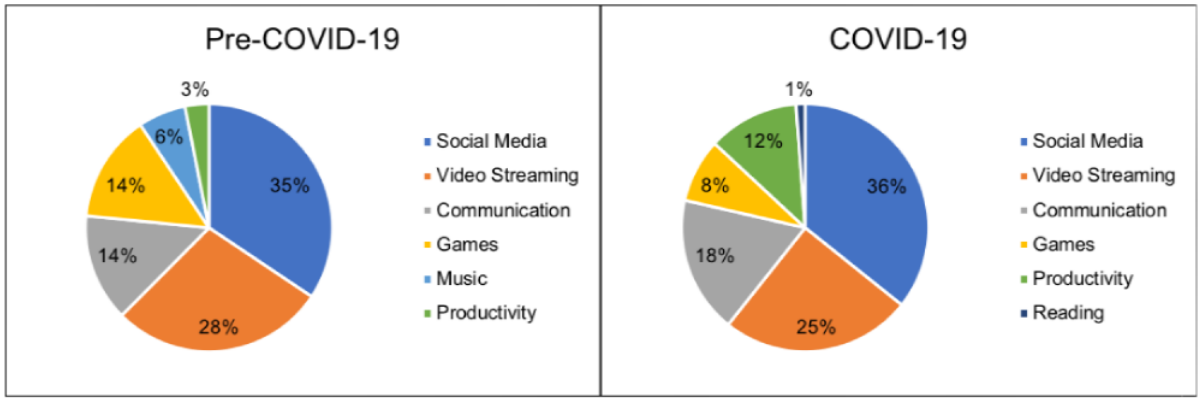


Table 2. Comparison of daily survey scores between pre–COVID-19 and COVID-19 cohorts.

Surveys	Values			
	Pre–COVID-19	COVID-19	β	*P* value
Average PIU-SF-6[a] ≥15, n (%)	1 (3.7)	10 (23.8)	2.99	.02
Average PHQ-8[b] ≥10, n (%)	8 (29.6)	16 (41.0)	.22	.72
Average GAD-7[c] ≥10, n (%)	2 (7.4)	11 (28.2)	1.44	.11
Average daily screen time, mean minutes (SD)	351.46 (204.32)	380.47 (135.23)	—[d]	.10
Average daily social media time, mean minutes (SD)	123.14 (77.58)	173.23 (84.80)	—	.049

[a]PIU-SF-6: Problematic Internet Use Short Form 6.

[b]PHQ-8: 8-item Patient Health Questionnaire.

[c]GAD-7: 7-item General Anxiety Disorder.

[d]Not available.

Of those participants in the COVID-19 cohort, youth with a personal history of COVID-19 exposure reported a significantly higher average daily screen time (*P*=.01) (Table 3). However, familial COVID-19 diagnoses and hospitalizations did not appear to be related to changes in digital media use or higher daily PIU-SF-6, PHQ-8, or GAD-7 scores. There were no significant differences between age and gender between youth with personal or family exposure to COVID-19 and those without. Of those participants who reported COVID-19 exposure, the majority (67% [n=4]) were youth of color. By contrast, among participants without a history of COVID-19 exposure, only 47% (n=17) identified as youth of color. Youth of color in the COVID-19 pandemic did not have significantly higher rates of PIU, screen time, or social media time (*P*=.72, *P*=.12, and *P*=.45, respectively). Overall, youth with PIU-SF-6 scores of ≥15 in our study population were significantly younger and more likely to have comorbid clinically elevated PHQ-8 and GAD-7 scores (*P*=.003) (Table 4).

Table 3. Associations between personal and familial COVID-19 exposure, psychiatric symptoms, and daily smartphone use.

Averaged surveys for infections and hospitalization	Values		
	No	Yes	*P* value
Personal COVID-19 infection			
Total	36	6	N/A[a]
Average PIU-SF-6[b] ≥15, n (%)	9 (25.0)	1 (16.7)	.99
Average PHQ-8[c] ≥10, n (%)	13 (38.2)	3 (60.0)	.63
Average GAD-7[d] ≥10, n (%)	9 (26.5)	2 (40.0)	.61
Average daily screen time, mean minutes (SD)	356 (126)	548 (54)	.01
Average daily social media time, mean minutes (SD)	166 (85)	221 (76)	.26
Familial COVID-19 infection			
Total	14	23	N/A
Average PIU-SF-6 ≥15, n (%)	3 (15.8)	7 (30.4)	.31
Average PHQ-8 ≥10, n (%)	6 (35.3)	10 (45.5)	.74
Average GAD-7 ≥10, n (%)	5 (29.4)	6 (27.3)	.99
Average daily screen time, mean minutes (SD)	365 (102)	391 (156)	.55
Average daily social media time, mean minutes (SD)	143 (72)	192 (89)	.16
Familial COVID-19 hospitalization			
Total	27	10	N/A
Average PIU-SF-6 ≥15, n (%)	7 (21.9)	3 (30)	.68
Average PHQ-8 ≥10, n (%)	11 (37.9)	5 (50)	.71
Average GAD-7 ≥10, n (%)	9 (31)	2 (20)	.69
Average daily screen time, mean minutes (SD)	356 (136)	443 (119)	.11
Average daily social media time, mean minutes (SD)	161 (89)	204 (68)	.07

[a]N/A: not applicable.

[b]PIU-SF-6: Problematic Internet Use Short Form 6.

[c]PHQ-8: 8-item Patient Health Questionnaire.

[d]GAD-7: 7-item General Anxiety Disorder.

Table 4. Participant comparisons based on average Problematic Internet Use Short Form 6 scores.

Characteristics	Average PIU-SF-6[a] <15	Average PIU-SF-6 ≥15	*P* value
Age (years), mean (SD)	16.4 (2.5)	15.7 (1.8)	.27
Female/male, n (%)	37/21 (63.8/36.2)	9/2 (81.8/18.2)	.31
Minority youth, n (%)	28 (48.3)	6 (54.5)	.75
Preexisting anxiety disorder diagnosis, n (%)	36 (62.1)	6 (54.5)	.74
Preexisting depressive disorder diagnosis, n (%)	38 (65.5)	7 (63.6)	.99
Average PHQ-8[b] ≥10, n (%)	16 (28.6)	8 (80.0)	.003
Average GAD-7[c] ≥10, n (%)	7 (12.5)	6 (60.0)	.003
Average daily screen time, mean (SD)	359.2 (166.5)	424.1 (152.8)	.12
Average daily social media time, mean (SD)	146.5 (83.1)	188.8 (91.3)	.19

[a]PIU-SF-6: Problematic Internet Use Short Form 6.

[b]PHQ-8: 8-item Patient Health Questionnaire.

[c]GAD-7: 7-item General Anxiety Disorder.

Discussion

Our results demonstrate that the COVID-19 pandemic may have altered digital media habits in youth with psychiatric illness. Study participants assessed over 6 weeks during the pandemic were significantly more likely to endorse consistent feelings of problematic dependency on the internet. Because rates of preexisting anxiety and depressive disorders did not differ significantly between cohorts, the higher rates of PIU seen in the COVID-19 cohort were likely not attributable to the participants' preexisting psychiatric disorders. However, prior studies have consistently emphasized positive correlations between PIU and active psychiatric symptoms, including in youth in mental health treatment [7-9,12]. This connection between active psychiatric distress and PIU is also supported by our study's finding that participants with PIU were more likely to report experiencing clinically significant symptoms of anxiety and depression during the study period. Thus, in the presence of active psychiatric symptoms, youth in our study population may be predisposed to develop PIU in environments of increased stress, such as a pandemic.

Some studies have suggested that excessive internet use and PIU directly cause adverse mental health outcomes [20]. However, our previous pilot studies using ecological momentary assessment and digital phenotyping have shown that for youth in mental health treatment, screen time and PIU are linked to temporary improvements in anxiety and depressive symptoms [12,13]. Existing research indicates that youth with mental health difficulties are more inclined to turn to online peer support to manage health issues; for example, youth with moderate-to-severe depressive symptoms are more likely to seek out peers' health-related stories posted online [21]. Because in-person supports were more challenging to access during the pandemic, especially mental health services offered through school or in-home visits, youth in our study population may have gone online to help regulate their negative emotions. This hypothesis is further reinforced by our finding that youth with psychiatric diagnoses in the COVID-19 cohort spent a significantly larger percentage of their daily screen time on social media. Social media platforms specifically can offer interpersonal connection and external validation, opportunities for which were more limited during the pandemic. Thus, without their usual mental health treatments or coping skills consistently available, these youth may have been at particularly high risk of developing or reinforcing habitual reliance upon social media as a primary coping skill.

The fact that average screen time did not also increase during the pandemic in our study population may reflect our participants' high baseline rates of digital media use compared with youth without active psychiatric symptoms [21,22]. However, in the COVID-19 cohort, youth who reported a history of COVID-19 exposure used their smartphones significantly more on a daily basis; these youths' iPhone screen time summaries indicated 54% more minutes of daily screen time than participants without such history. Notably, all participants with a known history of COVID-19 exposure were exposed before their 6-week study period, and only 50% of exposed participants subsequently contracted the virus, suggesting that illness and requisite quarantine were unlikely to be the sole contributors to this increase in phone use. We hypothesize that youth exposed to COVID-19 may have been more likely to appreciate the risks associated with the virus and therefore relied on virtual rather than in-person pastimes out of fear of contracting the virus. Additionally, the majority of our participants exposed to COVID-19 were youth of color, compared to youth who were not exposed, where the majority were White and non-Hispanic/Latinx. It has been well established that ethnic and racial minority communities are at greater risk of COVID-19 due to systemic racism impacting health care access, housing, and occupation [23], and communities with higher rates of COVID-19 transmission may have been particularly limited in the ability to provide in-person mental health services or safe spaces for in-person interactions. Moreover, many adults in these communities were essential workers and unable to stay home with children to monitor and provide guidance surrounding the amount of daily screen time. As youth of color did not have higher rates of screen time or PIU during the pandemic, the combination of multiple psychosocial stressors and COVID-19 exposure may have been necessary for triggering increased smartphone engagement.

These findings have significant implications for the treatment of youth with psychiatric diagnoses. While it is always important for clinicians to revisit a patient's digital media habits periodically throughout the course of treatment, the pandemic may necessitate additional screening for changes in media use. Youth struggling with their psychiatric symptoms or with a history of COVID-19 exposure may also benefit from PIU screening specifically, and their parents or guardians should be asked about conflicts arising surrounding separation from devices, particularly smartphones. A positive screen will allow for the careful development of a thoughtful, gradated media plan to help youth move back into healthier patterns of digital media use and begin intentional practice of coping skills that are independent of screens. Ideally, these youth will be more successful re-adjusting to aspects of screen-free daily living if the transition is predictable and gradual and involves youth input.

Finally, social media research in this population is challenging; even in adults, the recall accuracy of daily screen time is limited [24], and the finding that many younger populations use digital media continuously [25] likely further impacts recall accuracy. This study's use of EMA data afforded us a better opportunity to appreciate ecologically valid and objective changes in youth digital media use through longitudinal sampling and procurement of Apple screen time summaries. By asking our participants to provide us with their daily screen time reports, we were able to gather both qualitative and quantitative data regarding smartphone use in a population subset where protocol adherence can be challenging. Assessing the feasibility of app-based EMA as a clinical intervention was not the primary goal of this study. However, monthly visits are standard of care in pediatric psychiatry, and the majority of our participants provided psychiatric symptom updates on at least a weekly basis; therefore, there may be a clinical role for app-based EMA in this population, particularly to track changes in digital media use and associated mood symptoms.

Our research findings cannot conclude that the pandemic was the root cause of worsening youth mental health or PIU, and the study's small sample size is a notable limitation. Our data suggest that youth in mental health treatment were at increased risk of PIU development during the pandemic, specifically those with more severe symptoms of anxiety and depression. Moreover, those youth in mental health treatment exposed to COVID-19 endorsed greater amounts of daily smartphone use than those without a history of exposure. Based on our results, we recommend that clinicians screen high-risk pediatric patients for potential pandemic-associated changes in digital media habits as this may prevent psychiatric crises secondary to digital media-related conflict in the home or at school. From a systems standpoint, such crisis prevention measures may ease the burdens placed on our already overwhelmed psychiatric crisis teams and emergency rooms as we continue to navigate the COVID-19 pandemic.

Acknowledgments

The authors would like to thank Benjamin Cook for his assistance in project conceptualization and analyses. This study was supported by the DuPont Warren Fellowship and Livingston Award, awarded to MG by the Department of Psychiatry, Harvard Medical School.

Authors' Contributions

MG was involved in conceptualization, methodological design, data collection and curation, formal analysis, funding acquisition, writing the original manuscript draft, and subsequent revisions. MN, VL, and SP were involved in data collection and curation, formal analysis, writing the original manuscript draft, and subsequent revisions. NC and JT were involved in conceptualization, methodological design, and revising the manuscript.

Conflicts of Interest

JT receives research support from Otsuka Pharmaceuticals for unrelated work.

References

1. Leeb RT, Bitsko RH, Radhakrishnan L, Martinez P, Njai R, Holland KM. MMWR Morb Mortal Wkly Rep 2020 Nov 13;69(45):1675-1680 [FREE Full text] [doi: 10.15585/mmwr.mm6945a3] [Medline: 33180751]
2. Yard E, Radhakrishnan L, Ballesteros MF, Sheppard M, Gates A, Stein Z, et al. Emergency Department Visits for Suspected Suicide Attempts Among Persons Aged 12-25 Years Before and During the COVID-19 Pandemic - United States, January 2019-May 2021. MMWR Morb Mortal Wkly Rep 2021 Jun 18;70(24):888-894 [FREE Full text] [doi: 10.15585/mmwr.mm7024e1] [Medline: 34138833]
3. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global Prevalence of Depressive and Anxiety Symptoms in Children and Adolescents During COVID-19: A Meta-analysis. JAMA Pediatr 2021 Nov 01;175(11):1142-1150. [doi: 10.1001/jamapediatrics.2021.2482] [Medline: 34369987]
4. Furfaro H. Scientists are racing to unravel the pandemic's toll on kids' brains. The Seattle Times Internet. 2021 Sep 01. URL: https://www.seattletimes.com/education-lab/scientists-are-racing-to-unravel-the-pandemics-toll-on-kids-brains/ [accessed 2021-08-25]
5. Ellison K. Children's mental health badly harmed by the pandemic. Therapy is hard to find. The Washington Post. 2021 Aug 25. URL: https://www.washingtonpost.com/health/child-psychiatrist-counselor-shortage-mental-health-crisis/2021/08/13/844a036a-f950-11eb-9c0e-97e29906a970_story.html [accessed 2022-01-26]
6. Twenge JM, Campbell WK. Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. Prev Med Rep 2018 Dec;12:271-283 [FREE Full text] [doi: 10.1016/j.pmedr.2018.10.003] [Medline: 30406005]
7. Kaess M, Durkee T, Brunner R, Carli V, Parzer P, Wasserman C, et al. Pathological Internet use among European adolescents: psychopathology and self-destructive behaviours. Eur Child Adolesc Psychiatry 2014 Nov;23(11):1093-1102 [FREE Full text] [doi: 10.1007/s00787-014-0562-7] [Medline: 24888750]
8. Fuchs M, Riedl D, Bock A, Rumpold G, Sevecke K. Pathological Internet Use—An Important Comorbidity in Child and Adolescent Psychiatry: Prevalence and Correlation Patterns in a Naturalistic Sample of Adolescent Inpatients. BioMed Research International 2018;2018:1-10. [doi: 10.1155/2018/1629147]
9. Gansner M, Belfort E, Cook B, Leahy C, Colon-Perez A, Mirda D, et al. Problematic Internet Use and Associated High-Risk Behavior in an Adolescent Clinical Sample: Results from a Survey of Psychiatrically Hospitalized Youth. Cyberpsychology, Behavior, and Social Networking 2019 May;22(5):349-354. [doi: 10.1089/cyber.2018.0329]
10. Chen I, Chen C, Pakpour AH, Griffiths MD, Lin C. Internet-Related Behaviors and Psychological Distress Among Schoolchildren During COVID-19 School Suspension. J Am Acad Child Adolesc Psychiatry 2020 Oct;59(10):1099-1102.e1 [FREE Full text] [doi: 10.1016/j.jaac.2020.06.007] [Medline: 32615153]
11. Alheneidi H, AlSumait L, AlSumait D, Smith AP. Loneliness and Problematic Internet Use during COVID-19 Lock-Down. Behav Sci (Basel) 2021 Jan 06;11(1):5 [FREE Full text] [doi: 10.3390/bs11010005] [Medline: 33418914]

12. Gansner M, Nisenson M, Carson N, Torous J. A pilot study using ecological momentary assessment via smartphone application to identify adolescent problematic internet use. Psychiatry Res 2020 Nov;293:113428. [doi: 10.1016/j.psychres.2020.113428] [Medline: 32889344]
13. Gansner M, Nisenson M, Lin V, Carson N, Torous J. Piloting Smartphone Digital Phenotyping to Understand Problematic Internet Use in an Adolescent and Young Adult Sample. Child Psychiatry Hum Dev 2022 Jan 19:e. [doi: 10.1007/s10578-022-01313-y] [Medline: 35044580]
14. Gansner M, Belfort E, Leahy C, Mirda D, Carson N. An Assessment of Digital Media-related Admissions in Psychiatrically Hospitalized Adolescents. APS 2020 Jan 10;9(3):220-231. [doi: 10.2174/2210676609666190221152018]
15. LAMP: Learn, Assess, Manage, Prevent. The Division of Digital Psychiatry at BIDMC. URL: https://www.digitalpsych.org/lamp.html [accessed 2022-01-26]
16. Demetrovics Z, Király O, Koronczai B, Griffiths MD, Nagygyörgy K, Elekes Z, et al. Psychometric Properties of the Problematic Internet Use Questionnaire Short-Form (PIUQ-SF-6) in a Nationally Representative Sample of Adolescents. PLoS One 2016 Aug 9;11(8):e0159409 [FREE Full text] [doi: 10.1371/journal.pone.0159409] [Medline: 27504915]
17. Bauer AM, Baldwin SA, Anguera JA, Areán PA, Atkins DC. Comparing Approaches to Mobile Depression Assessment for Measurement-Based Care: Prospective Study. J Med Internet Res 2018 Jun 19;20(6):e10001. [doi: 10.2196/10001]
18. Stata: Software for Statistics and Data Science. Stata. URL: https://www.stata.com/ [accessed 2022-01-26]
19. RStudio. URL: http://www.rstudio.com/ [accessed 2022-01-26]
20. Twenge JM, Joiner TE, Rogers ML, Martin GN. Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time. Clinical Psychological Science 2017 Nov 14;6(1):3-17. [doi: 10.1177/2167702617723376]
21. Rideout V, Fox S, Well Being Trust. Digital Health Practices, Social Media Use, and Mental Well-Being Among Teens and Young Adults in the U.S. Providence St. Joseph Health Digital Commons 2018;1:1-95.
22. Riehm KE, Feder KA, Tormohlen KN, Crum RM, Young AS, Green KM, et al. Associations Between Time Spent Using Social Media and Internalizing and Externalizing Problems Among US Youth. JAMA Psychiatry 2019 Dec 01;76(12):1266-1273 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.2325] [Medline: 31509167]
23. Health Equity Considerations and Racial and Ethnic Minority Groups. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html [accessed 2022-01-26]
24. Araujo T, Wonneberger A, Neijens P, de Vreese C. How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use. Communication Methods and Measures 2017 Apr 27;11(3):173-190. [doi: 10.1080/19312458.2017.1317337]
25. Jiang J. How Teens and Parents Navigate Screen Time and Device Distractions. Pew Research Center. 2018 Aug 22. URL: https://www.pewresearch.org/internet/2018/08/22/how-teens-and-parents-navigate-screen-time-and-device-distractions/ [accessed 2022-01-26]

Abbreviations

EHR: electronic health record
EMA: ecological momentary assessment
GAD-7: 7-item General Anxiety Disorder
PHQ-8: 8-item Patient Health Questionnaire
PIU: problematic internet use
PIU-SF-6: Problematic Internet Use Short Form 6

Edited by G Eysenbach; submitted 25.08.21; peer-reviewed by A Teles, YH Yaw; comments to author 10.11.21; revised version received 06.12.21; accepted 19.12.21; published 28.01.22.

Please cite as:
Gansner M, Nisenson M, Lin V, Pong S, Torous J, Carson N
Problematic Internet Use Before and During the COVID-19 Pandemic in Youth in Outpatient Mental Health Treatment: App-Based Ecological Momentary Assessment Study
JMIR Ment Health 2022;9(1):e33114
URL: https://mental.jmir.org/2022/1/e33114
doi:10.2196/33114
PMID:35089157

Original Paper

Acoustic and Facial Features From Clinical Interviews for Machine Learning–Based Psychiatric Diagnosis: Algorithm Development

Michael L Birnbaum[1,2,3*], MD; Avner Abrami[4*], MSc; Stephen Heisig[5], BSc; Asra Ali[1,2], MA; Elizabeth Arenare[1,2], BA; Carla Agurto[4], PhD; Nathaniel Lu[1,2], MA; John M Kane[1,2,3*], MD; Guillermo Cecchi[4*], PhD

[1]Department of Psychiatry, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY, United States

[2]The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY, United States

[3]The Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, United States

[4]Computational Biology Center, IBM Research, Yorktown Heights, NY, United States

[5]Icahn School of Medicine at Mount Sinai, New York City, NY, United States

[*]these authors contributed equally

Corresponding Author:
Michael L Birnbaum, MD
Department of Psychiatry
The Zucker Hillside Hospital
Northwell Health
75-59 263rd St
Glen Oaks, NY, 11004
United States
Phone: 1 7184708305
Email: mbirnbaum@northwell.edu

Abstract

Background: In contrast to all other areas of medicine, psychiatry is still nearly entirely reliant on subjective assessments such as patient self-report and clinical observation. The lack of objective information on which to base clinical decisions can contribute to reduced quality of care. Behavioral health clinicians need objective and reliable patient data to support effective targeted interventions.

Objective: We aimed to investigate whether reliable inferences—psychiatric signs, symptoms, and diagnoses—can be extracted from audiovisual patterns in recorded evaluation interviews of participants with schizophrenia spectrum disorders and bipolar disorder.

Methods: We obtained audiovisual data from 89 participants (mean age 25.3 years; male: 48/89, 53.9%; female: 41/89, 46.1%): individuals with schizophrenia spectrum disorders (n=41), individuals with bipolar disorder (n=21), and healthy volunteers (n=27). We developed machine learning models based on acoustic and facial movement features extracted from participant interviews to predict diagnoses and detect clinician-coded neuropsychiatric symptoms, and we assessed model performance using area under the receiver operating characteristic curve (AUROC) in 5-fold cross-validation.

Results: The model successfully differentiated between schizophrenia spectrum disorders and bipolar disorder (AUROC 0.73) when aggregating face and voice features. Facial action units including cheek-raising muscle (AUROC 0.64) and chin-raising muscle (AUROC 0.74) provided the strongest signal for men. Vocal features, such as energy in the frequency band 1 to 4 kHz (AUROC 0.80) and spectral harmonicity (AUROC 0.78), provided the strongest signal for women. Lip corner–pulling muscle signal discriminated between diagnoses for both men (AUROC 0.61) and women (AUROC 0.62). Several psychiatric signs and symptoms were successfully inferred: blunted affect (AUROC 0.81), avolition (AUROC 0.72), lack of vocal inflection (AUROC 0.71), asociality (AUROC 0.63), and worthlessness (AUROC 0.61).

Conclusions: This study represents advancement in efforts to capitalize on digital data to improve diagnostic assessment and supports the development of a new generation of innovative clinical tools by employing acoustic and facial data analysis.

(JMIR Ment Health 2022;9(1):e24699) doi:10.2196/24699

KEYWORDS

audiovisual patterns; speech analysis; facial analysis; psychiatry; schizophrenia spectrum disorders; bipolar disorder; symptom prediction; diagnostic prediction; machine learning; audiovisual; speech; schizophrenia; spectrum disorders

Introduction

Approximately 20% of individuals aged 15 years and older experience psychiatric illness annually [1-3]. Psychiatrists may see as many as 8 patients hourly and are often unable to obtain the detailed information necessary to make effective, evidence-based, and personalized clinical decisions [4-6]. In contrast to all other areas of medicine, psychiatry is still nearly entirely reliant on subjective assessments such as patient self-report and clinical observation [7,8]. There are few valid and reliable tests, biomarkers, and objective sources of collateral information available to support diagnostic procedures and assess health status. The lack of objective information on which to base clinical decisions can contribute to reduced quality of care, underrecognized signs and symptoms, and poorer treatment outcomes, including higher dropout rates, reduced medication adherence, and persistent substance abuse [9,10]. Behavioral health clinicians need access to objective and reliable, easily collected, and interpretable patient data to enable quick, effective, and targeted interventions [11,12].

In recent years, progress has been made in audiovisual data processing [13-21]. Advances in this technology could play a pivotal role in supporting automated methods of collecting objective adjunctive patient data to inform diagnostic procedures, psychiatric symptom identification, and psychiatric symptom monitoring. Speech analysis, in particular, has been studied [22-36] because changes in both the content and acoustic properties of speech are known to be associated with several psychiatric conditions: disorganized speech in schizophrenia, pressured speech in mania, and slowed speech in depression [7]. Moreover, speech represents a universal, easily extracted, and clinically meaningful biological process and is therefore well positioned to serve as an objective marker of psychiatric illness [27]. Prior research has demonstrated the potential for the use of speech properties to distinguish between individuals with and without a variety of psychiatric disorders with high degrees of accuracy [22-36]. Acoustic analysis, for instance, has demonstrated that participants with schizophrenia tend to exhibit less total time talking, reduced speech rate, and higher pause duration [23,27,33-40] than healthy participants and that participants with bipolar disorder demonstrate increases in tonality [41-43].

Concurrently, alterations in facial expressivity accompany several psychiatric illnesses: flat or inappropriate affect in individuals with schizophrenia, euphoric or labile affect in mania, and slowed or diminished facial movements in depression [7]. Video analysis has accordingly emerged as a potentially objective and reliable method for capturing subtle head, face, and eye movements with greater precision than by clinical observation alone [16,44-46]. Alterations in facial expressivity have demonstrated success in predicting the presence of various psychiatric illnesses including schizophrenia spectrum disorders [47-49], mood disorders [49-51], and autism spectrum disorders [48].

Audiovisual patterns represent an easily extractable, naturalistic, universal, and objective data that could serve as viable digital biomarkers in psychiatry, contributing adjunctive information about a patient, beyond what can be assessed solely through traditional means. No study, to the best of our knowledge, has explored the potential for using audiovisual data to discriminate between a diagnosis of schizophrenia or bipolar disorder, a task which can be challenging for behavioral health clinicians given significant symptom overlap [52,53], especially during the early course of illness development. Additionally, few studies [19,54] have explored the relationship between audiovisual data and psychiatric symptoms, commonly used as primary outcome measures, to more efficiently and more effectively identify the presence of a specific psychiatric sign or symptom. Furthermore, research thus far has largely explored individual data sources in isolation [19,20], however, advancing this critical work will now require integrating multiple streams of digital data.

We aimed to differentiate between schizophrenia spectrum disorders and bipolar disorder using audiovisual data alone. We hypothesized that physiological data from voice acoustics and facial action units could be used to distinguish between individuals with schizophrenia spectrum disorders and individuals with bipolar disorder and that these signals would be associated with specific psychiatric signs and symptoms.

Methods

Recruitment

Participants between the ages of 15 and 35 years old diagnosed with schizophrenia spectrum disorders or bipolar disorder were recruited from Northwell Health Zucker Hillside Hospital's inpatient and outpatient psychiatric departments. Diagnoses were based on clinical assessment of the most recent episode and were extracted from participant's medical record at the time of consent. Most participants with schizophrenia spectrum disorders were recruited from the Early Treatment Program, which is a specialized outpatient early psychosis intervention clinic. Individuals with psychiatric comorbidities (such as substance use disorders) were included. Participants with known physical impairments (such as paralysis or severe laryngitis) capable of impacting facial movements or acoustic capabilities were excluded. Eligible participants were recruited by a research staff member. Healthy volunteers who had already been screened for prior studies were also recruited. Recruitment occurred between September 2018 and July 2019. The study was approved by the institutional review board (18-0137) of Northwell Health. Written informed consent was obtained from adult participants and legal guardians of participants under 18 years. Assent was obtained from minors. All participants received treatment as usual.

Interviews

Participants were assessed at baseline and invited to return for optional quarterly assessments thereafter for a maximum of 12

months. Healthy volunteers were assessed at baseline and invited to return for optional assessments at month 6 and month 12. At each visit, all participants, including healthy volunteers, were interviewed by a trained and reliable research rater utilizing the Brief Psychiatric Rating Scale (BPRS) [55], Scale for the Assessment of Negative Symptoms (SANS) [56], Hamilton Depression Rating Scale (HAMD) [57], and Young Mania Rating Scale (YMRS) [58]. In addition, at each visit, participants were asked a series of 5 emotionally neutral, open-ended questions designed to encourage speech production. For example, participants were asked to describe a typical dinner, discuss a television show or movie that they had watched, or talk about a current or prior pet. Participants were instructed to talk freely and prompted to continue to talk as much as they liked for each response. Similar methods for speech extraction have been successfully implemented in prior research [34]. Both participant and the interviewer wore headsets with microphones connected to a 2 by 2 amplifier (TASCAM) to record audio. Video was recorded with an iPad Pro (Apple Inc) focused on participants' facial expressions.

Raw data were stored in a firewalled server and were never shared outside of Northwell Health. The processing of high-level features was implemented locally, and only those features were used for further analysis outside the raw data server. High-level feature data remained within Health Insurance Portability and Accountability Act–compliant servers.

Data Preprocessing

Before extracting acoustic features, saturation, if present, was removed by identifying time points with amplitudes higher than 99.99% of the maximum value, and given that recordings involved the use of two audio channels (one each, for participant and interviewer), we extracted only the participant's voice.

Acoustic features were extracted using the OpenSMILE open-source toolbox [59]. We used a predefined feature set [60] for low-level descriptors. This configuration encompasses 150 features, which were computed with a fixed window size (ie, mel-frequency cepstral coefficients -25 ms) but with a sampling rate of 10 ms (Multimedia Appendix 1).

For facial features, we used openFace software [61]. This tool detects the presence and intensity of 18 facial expressions called action units (Multimedia Appendix 2). The video sampling rate was 30 Hz.

Both facial action units and acoustic time series were downsampled to 10 Hz (by taking the average value in each consecutive 0.1-second window) and aligned. We then fragmented each interview into consecutive 1.5-minute blocks. In each block, we derived 2 sets of aggregate features (one that was computed when the participant was listening, the other while speaking) to help ensure that the silence between answers did not have an effect on acoustic feature values and that the dynamics of facial action units in both conditions were captured by the models. Mean value and standard deviation were computed for each feature and for each 1.5-minute block. For better classification generalization and to reduce overfitting, we augmented each interview 25 times by selecting only 1 out of 2 consecutive blocks randomly for each block in the sequence.

Classification Tasks

We explored 2 main classification tasks: differential diagnosis, assigning an interview as belonging to a specific group (either schizophrenia spectrum disorders or bipolar disorder) based purely on physiological patterns, and symptom detection, predicting the presence of a psychiatric sign or symptom. In total, 75 classification tasks were run, each corresponding to the 75 unique psychiatric signs and symptoms assessed with the BPRS (18 items), SANS (22 items), YMRS (11 items), and HAMD (24 items). For each classification task, participants were assigned to the positive class if their symptom score exceeded the clinical threshold of at least mild severity: score ≥3 on BPRS items (range 1-7), score ≥2 on SANS items (range 0-5), score ≥2 or ≥4 on YMRS items (with ranges 0-4 and 0-8, respectively), and score ≥2 or ≥1 on HAMD items (with ranges 0-4 and 0-2, respectively). Total scores could range from 18 to 126 for the BPRS, 0 to 110 for the SANS, 0 to 60 for the YMRS, and 0 to 76 for the HAMD.

For each classification task, we computed 2 independent models for both men and women. This was done to prevent possible sex-specific physiological confounds in voice and face to impact the results, as the bipolar disorder group was composed of a majority of women. Additionally, we aimed to build models that were not individual-dependent.

All inferences were undertaken using a gradient boosting classifier [62] (Python; Scikit-learn library [63]) (fixed seed 0, deviance loss, 0.1 learning rate, 100 weak learners, with 10% of all samples selected randomly used for fitting the individual base learners). All inferences were run in stratified 5-fold cross-validation (participants were divided in 5 nonoverlapping groups and each group was used once as a validation, while the 4 remaining groups formed the training set). Only the most predictive features—those achieving a leave-one-out area under the receiver operating characteristic curve [AUROC] greater than 0.6 on the training set of each fold—were used by the gradient boosting classifier.

Finally, we ensured that each group (both in the positive and negative class) had similar average interview durations, We removed the final few minutes from the end of the lengthier interviews (corresponding to the difference between the average length in each class) to ensure that interview duration was not a confounding factor in classification performance, because longer interviews would provide greater statistical sampling of the features.

Aggregating Different Modalities

We investigated 3 different models including a Face model (all relevant facial action units features), a Voice model (all relevant acoustic features), and a Face–Voice model, which was constructed by averaging the probability outputs of the Face model and the Voice model. For each inference, 5-fold AUROC, accuracy, accuracy chance (the accuracy one would get by randomly attributing the classes), and F scores (for both classes of the classification) were calculated. A threshold of 0.5 was used to compute accuracy and F scores. To rank features (to assess which ones were most predictive), we used a 5-fold AUROC for each feature sequence alone. We report the most

successful models per modality (voice alone, face alone, or combined voice and face).

Results

General

In total, 89 participants (mean age 25.3 years; male: 48/89, 53.9%; female: 41/89, 46.1%) with schizophrenia spectrum disorders (n=41), bipolar disorder (n=21), and healthy volunteers (n=27) were included (Table 1), resulting in 146 interviews (mean 1.64, SD 0.84 interviews per participant). Total scores (representing aggregate scores from individual items) indicated that participants were predominantly in remission at the time of the assessments (Table 2); however, several participants scored moderate or severe on 1 or more items in the BPRS (schizophrenia spectrum disorders: 22/41, 54%; bipolar disorder: 8/21, 38%), SANS (schizophrenia spectrum disorders: 33/41, 80%; bipolar disorder: 14/21, 67%), YMRS (schizophrenia spectrum disorders: 18/41, 44%; bipolar disorder: 8/21, 38%), and HAMD (schizophrenia spectrum disorders: 32/41, 78%; bipolar disorder: 10/21, 48%). Participant assessments, including speech extraction and symptom rating scales, lasted a mean duration of 27 minutes (SD 11).

Table 1. Demographic and clinical characteristics.

Characteristic	Schizophrenia spectrum disorders (n=41)	Bipolar disorder (n=21)	Healthy volunteers (n=27)	Full sample (n=89)
Age (in years), mean (SD)	23.7 (3.97)	25.3 (4.24)	28.5 (5.15)	25.5 (4.83)
Sex, n (%)				
Male	29 (71)	7 (33)	12 (44)	48 (54)
Female	12 (29)	14 (67)	15 (56)	41 (46)
Race/ethnicity, n (%)				
African American/Black	24 (58)	3 (14)	8 (30)	35 (39)
Asian	6 (15)	4 (19)	6 (22)	16 (18)
Caucasian	10 (24.)	9 (43)	10 (37)	29 (33)
Mixed race/other	1 (2)	5 (24)	2 (7)	8 (9)
Pacific Islander	0 (0)	0 (0)	1 (4)	1 (1)
Hispanic	5 (12)	3 (14)	1 (4)	9 (10)
Diagnosis (most recent episode), n (%)				
Schizophrenia	19 (46)	N/A[a]	N/A	19 (21.)
Schizophreniform	10 (24)	N/A	N/A	10 (11)
Schizoaffective	7 (17)	N/A	N/A	7 (8)
Unspecified schizophrenia spectrum disorders	5 (12)	N/A	N/A	5 (6)
Bipolar disorder (manic)	N/A	16 (76)	N/A	16 (18)
Bipolar disorder (depressed)	N/A	3 (14)	N/A	3 (3)
Bipolar disorder (mixed)	N/A	2 (10)	N/A	2 (2)
Interviews, n				
Baseline	41	21	27	89
Follow up	33	17	7	57
Interview length, mean (SD)	29.5 (13.1)	29.5 (9.3)	20.7 (6.1)	27 (11)

[a]N/A: not applicable.

Table 2. Symptom rating scale scores for diagnostic and sex groups.

Group	Brief Psychiatric Rating Scale score[a], mean (SD)	Scale for the Assessment of Negative Symptoms score[b], mean (SD)	Young Mania Rating Scale score[c], mean (SD)	Hamilton Depression Rating Scale score[d], mean (SD)
Schizophrenia spectrum disorders				
All	26.5 (6.8)	22.6 (12.3)	3.9 (3.6)	8.7 (6.3)
Men	28.1 (7.0)	25.5 (11.2)	4.6 (3.8)	9.8 (6.7)
Women	22.8 (4.4)	15.8 (12.1)	2.3 (2.1)	6.0 (4.1)
Bipolar disorder				
All	26.8 (8.3)	14.0 (9.2)	7.5 (8.5)	9.4 (7.9)
Men	25.9 (5.7)	10.5 (8.8)	8.9 (9.1)	9.8 (10.3)
Women	27.3 (9.5)	16.2 (8.7)	6.7 (8.1)	9.2 (5.9)

[a]The total score can range from 18-126.

[b]The total score can range from 0-110.

[c]The total score can range from 0-60.

[d]The total score can range from 0-76.

Differential Diagnosis

Differential diagnosis classification performed well (5-fold AUROC 0.73) when aggregating features from both face and voice (Table 3). Facial action units, such as AU17 (Figure 1A), provided the strongest signal in discrimination between men with schizophrenia spectrum disorders and men with bipolar disorder. Men with schizophrenia spectrum disorders activated their chin-raising muscle (AU17: 5-fold AUROC 0.74) and lip corner–pulling muscle (AU12: 5-fold AUROC 0.61) more frequently than men with bipolar disorder, while demonstrating reduced activation of their cheek-raising muscle (AU6: 5-fold AUROC 0.64). In contrast, voice features, such as mean energy in the in the frequency band 1-4 kHz (Figure 1B), performed best for women. Women with schizophrenia spectrum disorders demonstrated reduced energy in the frequency band 1-4 kHz (5-fold AUROC 0.80), reduced spectral harmonicity (5-fold AUROC 0.78), and increased spectral slope (5-fold AUROC 0.77) compared with women with bipolar disorder. When comparing participants with schizophrenia spectrum disorders to healthy volunteers and bipolar disorder to healthy volunteers, we achieved a 5-fold AUROC of 0.78 for both classification tasks.

We identified some features that discriminated well between schizophrenia spectrum disorders and bipolar disorder across both sexes: lip-corner pulling (AU12), which represented the movement of lip corners pulled diagonally by the zygomaticus major muscle (5-fold AUROC men: 0.61; women: 0.62) for which the mean value was higher on average for participants with schizophrenia spectrum disorders than for participants with bipolar disorder (Figure 2). The timing of this feature was observed to be important to classification performance—AU12 values were higher on average at the beginning of the interview and decreased over time.

Table 3. Diagnostic classification.

Features	AUROC[a]	Accuracy	Accuracy chance	F score	
				Schizophrenia spectrum disorders	Bipolar disorder
Voice	0.65	0.71	0.55	0.80	0.46
Face	0.68	0.72	N/A[b]	0.80	0.56
Face and voice	0.73	0.72	N/A	0.80	0.56

[a]AUROC: area under the receiver operating characteristic curve.

[b]N/A: not applicable.

Figure 1. Sex-specific features that discriminate between schizophrenia spectrum disorders and bipolar disorder: (A) mean activation of AU17 (chin raising while speaking), and (B) mean value of the energy in the frequency band 1-4 kHz. BD: bipolar disorder; SSD: schizophrenia spectrum disorders.

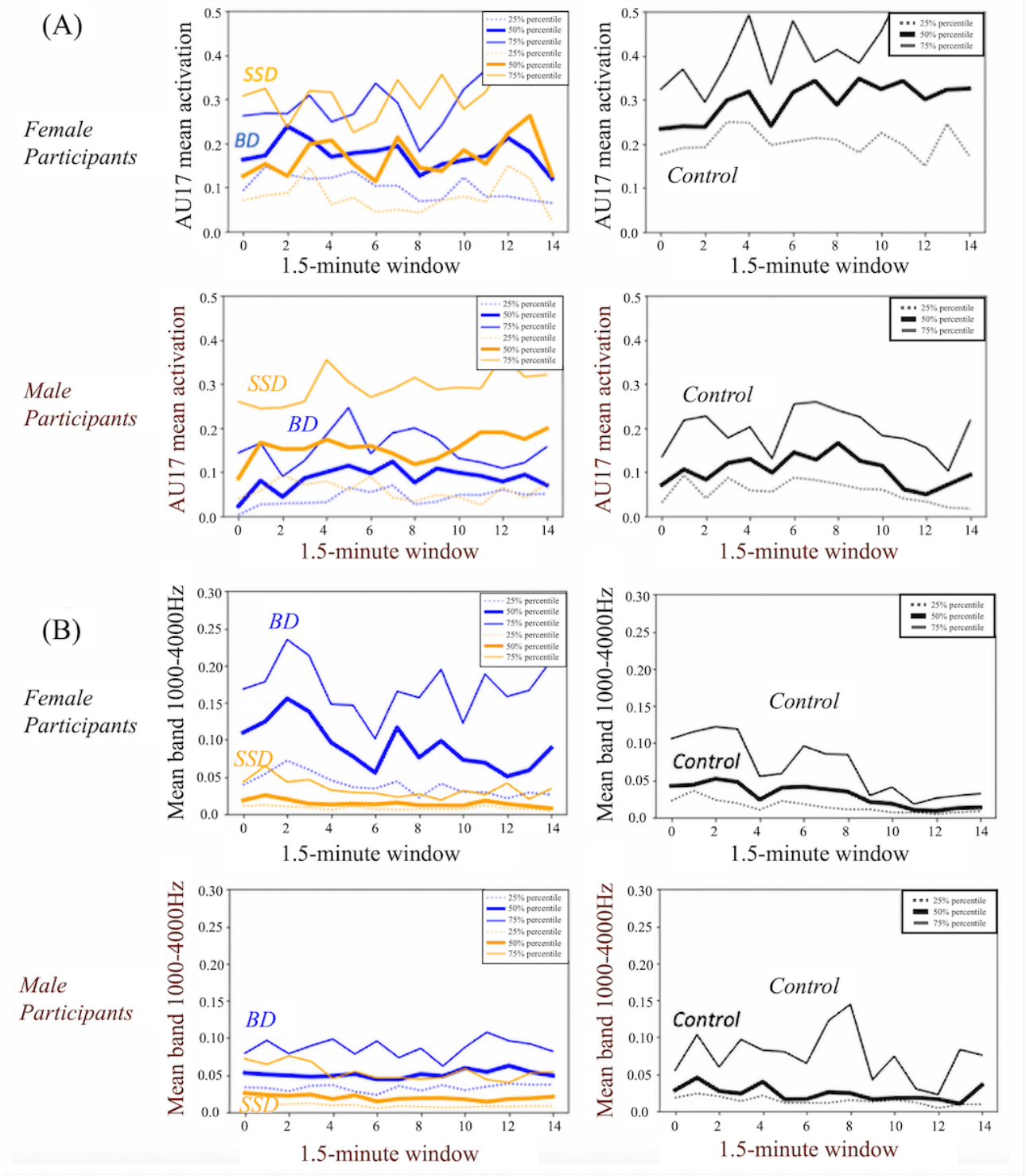
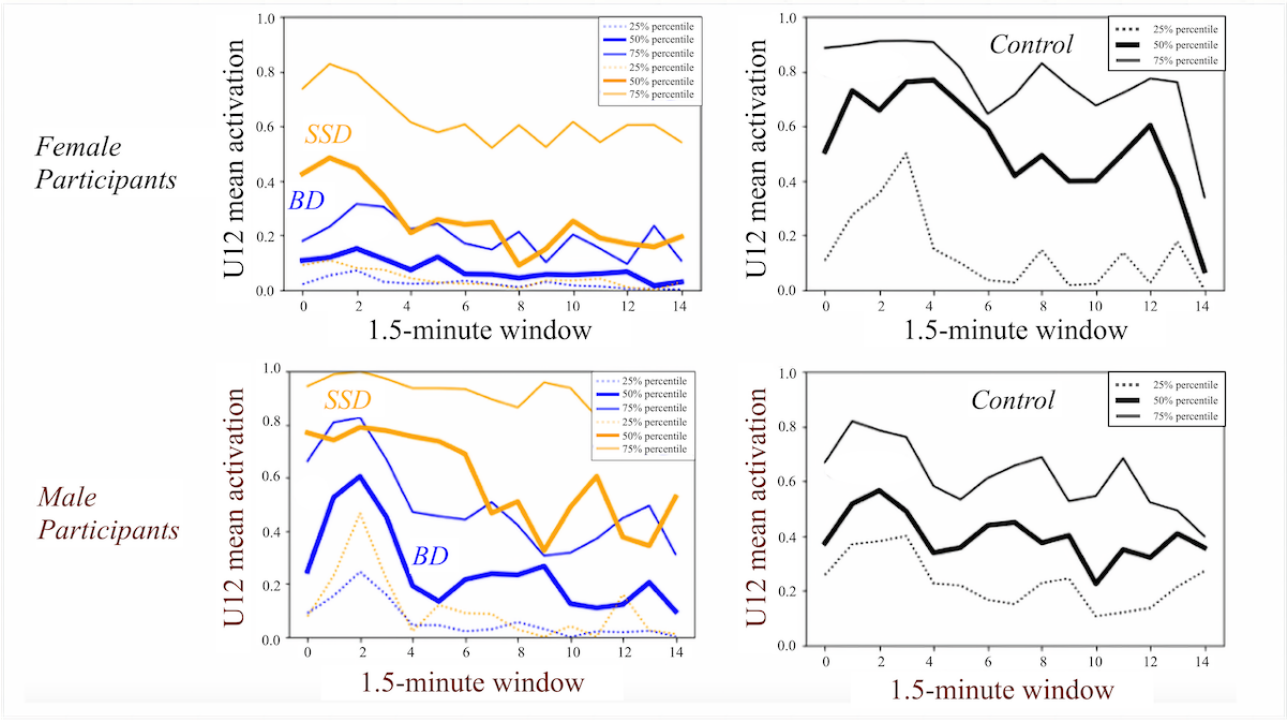


Figure 2. AU12 (lip-corner pulling while speaking) feature. For each signal, the 25th percentile, median, and 75th percentile values are shown for each 1.5-minute window. Bipolar disorder is represented in blue, schizophrenia spectrum disorders is represented in yellow, and on the adjacent plot, healthy volunteers is represented in black. BD: bipolar disorder; SSD: schizophrenia spectrum disorders.



Symptom Classification

Best performing models were derived from the SANS scale, predominantly from the affective flattening and blunting subgroup (global affective flattening, vocal inflection, paucity of expression, unchanging facial), avolition/apathy subgroup (physical anergia, role function level, global avolition), and asociality/anhedonia subgroup (sexual interest, asociality, intimacy). Two items passed the performance threshold from the BPRS (blunted affect and motor retardation), and 2 others were derived from the HAMD scale (work interest and worthlessness). No signs or symptoms from the YMRS passed the performance threshold criteria.

Voice outperformed facial action units for blunted affect (5-fold AUROC 0.81), whereas facial action units outperformed voice for unchanging facial expression (5-fold AUROC 0.64) (Table 4). Synergy between both modalities was observed for paucity of expression (5-fold AUROC 0.81).

Voice alone outperformed facial action units for several items including asociality (5-fold AUROC 0.63) and work and interests (5-fold AUROC 0.64) (Table 5). Facial action units alone outperformed voice for worthlessness (5-fold AUROC 0.61). Synergy between both modalities was observed for several other symptoms including avolition (5-fold AUROC 0.72) and anergia (5-fold AUROC 0.68). Importantly, given that these symptoms represent self-reported experiences, their relationship with measured physiological signals is likely indirect and one hypothesis is that they are linked to observable symptoms. For example, we found a correlation (r=0.35; P<.001) between work and interests and blunted affect, and a correlation (r=0.31; P<.001) between avolition and affective flattening.

Among the top acoustic features (Figure 3) for objectively observed symptoms (Table 4), the mean value of the energy in the frequency band 1-4 kHz was most indicative of paucity of expression (r= –0.27, P=.004). Specifically, a reduction in the average amount of energy in high frequencies was associated with the presence of this symptom. In addition to affecting voice quality or timber (in the form vocal overtones), high frequencies (1-4 kHz) are typical in shaping consonants through rapid air motion from the mouth and through the teeth. In contrast, vowels are generally in the lower frequencies (500 Hz) and contain the majority of the voice energy. Clinically, mismatch between the acoustic frequencies of vowels and consonants jeopardizes the natural sound of the voice and leads to a reduction in speech intelligibility. This observation is stable across sex.

Among the top facial action unit features (Figure 4) for the objectively observed symptoms, the standard deviation of cheek raising muscle activation, often activated to form a smile, was most indicative of blunted affect for both men and women (r= –0.26, P=.002 during speaking). When the symptom is present, the standard deviation of this feature is decreased.

Among the top features for self-reported symptoms (Table 5), the mean value of AU45 (blinking) during speaking is higher when the symptom feature worthlessness is present (r=0.30, P=.001, calculated over all participants) (Figure 5).

Table 4. Objectively observed item classification.

Symptom	Modality	AUROC[a]	Accuracy (random)	F score	
				Above clinical threshold	Below clinical threshold
Brief Psychiatric Rating Scale					
Blunted affect	Voice	0.81	0.95 (0.87)	0.40 |	0.97
Motor retardation	Face	0.68	0.94 (0.88)	0.36	0.97
Scale for the Assessment of Negative Symptoms					
Paucity of expression	Voice, face	0.81	0.80 (0.66)	0.42	0.88
Global affective flattening	Voice, face	0.79	0.82 (0.71)	0.44	0.89
Lack of vocal inflection	Voice, face	0.71	0.88 (0.78)	0.43	0.94
Unchanging facial	Face	0.64	0.83 (0.70)	0.39	0.90

[a]AUROC: area under the receiver operating characteristic curve.

Table 5. Self-reported items classification.

Symptom	Modality	AUROC[a]	Accuracy (random)	F score	
				Above clinical threshold	Below clinical threshold
Scale for the Assessment of Negative Symptoms					
Global avolition	Voice, face	0.72	0.66 (0.53)	0.75	0.49
Physical anergia	Voice, face	0.68	0.63 (0.51)	0.70	0.53
Role function level	Voice, face	0.65	0.63 (0.58)	0.75	0.31
Sexual interest	Voice, face	0.64	0.62 (0.52)	0.46	0.70
Intimacy	Voice	0.64	0.63 (0.51)	0.56	0.67
Asociality	Voice	0.63	0.60 (0.51)	0.54	0.65
Hamilton Depression Rating Scale					
Work and interests	Voice	0.62	0.65 (0.52)	0.73	0.52
Worthlessness	Face	0.61	0.88 (0.82)	0.32	0.94

[a]AUROC: area under the receiver operating characteristic curve.

Figure 3. Paucity of expression score as a function of the mean value of the energy in the high frequency band 1-4 KHz (log-scale) for healthy volunteers (blue), patient participants with symptom rating scale scores below symptom threshold (orange), and patient participants with symptom rating scale scores above symptom threshold (green). A lower value of this feature is indicative of a more severe symptom across sex. The black line indicates the median value of the feature.

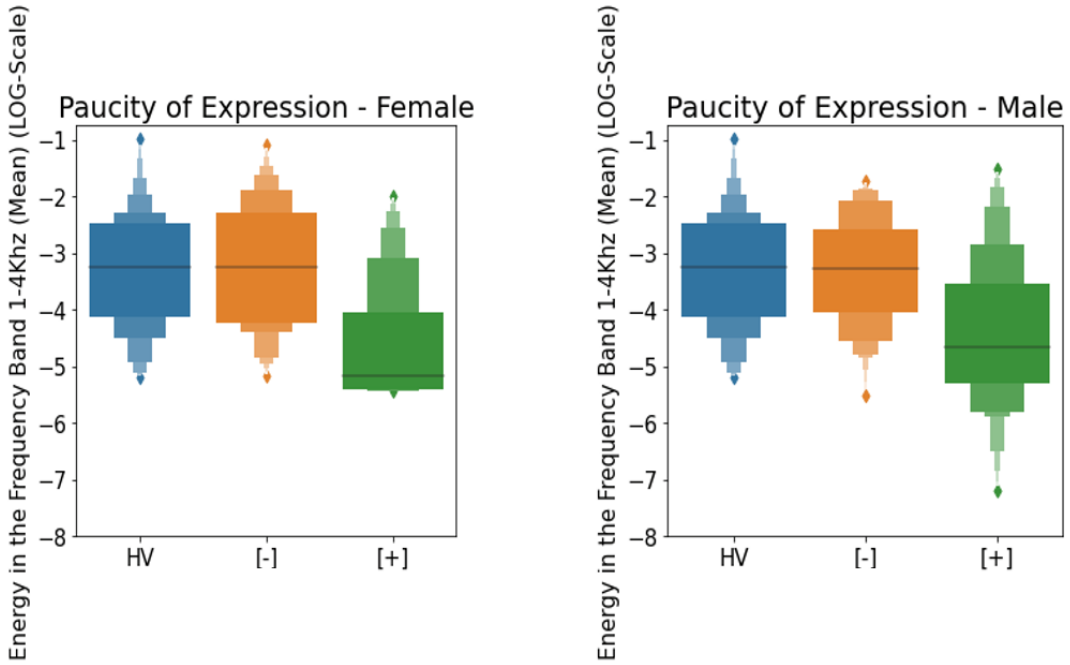


Figure 4. Blunted affect score as a function of the standard deviation of cheek raising (AU06) for healthy volunteers (blue), patient participants with symptom rating scale scores below symptom threshold (orange), and patient participants with symptom rating scale scores above symptom threshold (green). A lower value of this feature is indicative of a more severe symptom across sex. The black line indicates the median value of the feature.

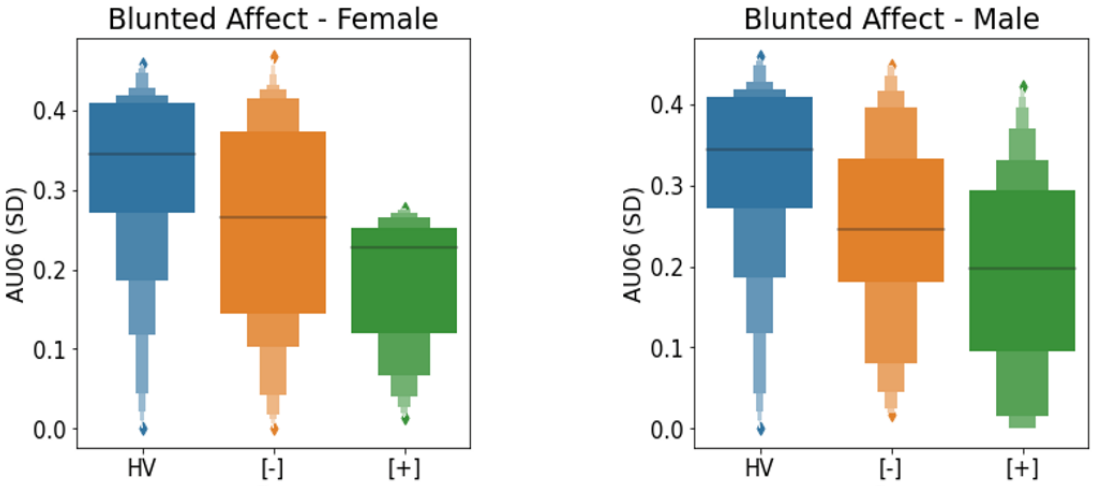
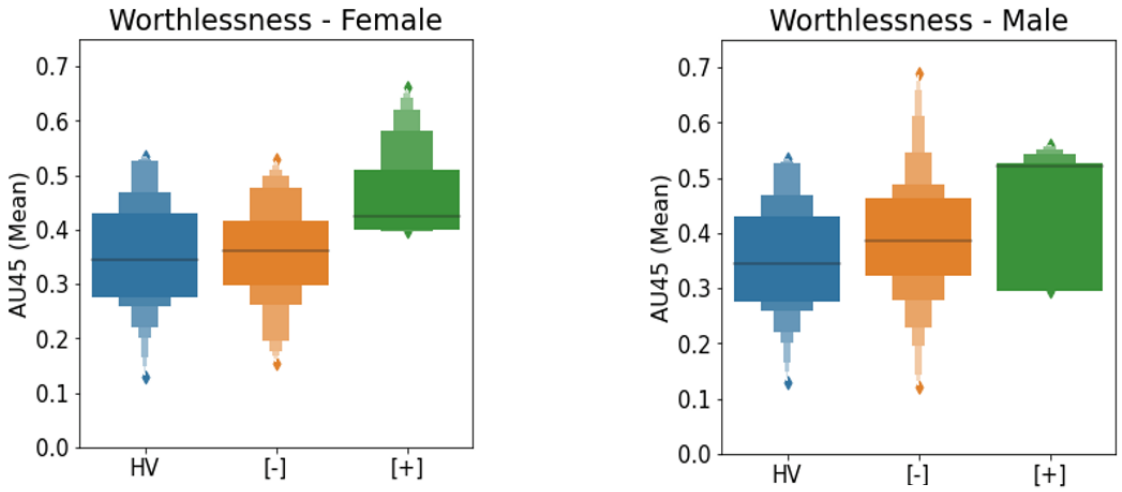


Figure 5. Worthlessness score as a function of the mean value of blinking (AU45) for healthy volunteers (blue), patient participants with symptom rating scale scores below symptom threshold (orange), and patient participants with symptom rating scale scores above symptom threshold (green). A higher value of this feature is indicative of a more severe symptom across sex. The black line indicates the median value of the feature.



Discussion

We aimed to explore the feasibility of utilizing audiovisual data extracted from participant interviews for psychiatric diagnoses and to predict the presence of psychiatric signs and symptoms. Our results indicate that computational algorithms developed from vocal acoustics and facial action units can successfully differentiate between participants with either schizophrenia spectrum disorders or bipolar disorder, as well as identify the presence of several psychiatric signs and symptoms with high degrees of accuracy. Both acoustic and facial action unit features could be independently used to differentiate between participants with schizophrenia spectrum disorders and bipolar disorder in our data set, and integrating the two modalities produced the strongest signal, as previously seen in studies of depression [64-66], suggesting a synergistic interaction. Importantly, different top features were identified for men and women. Specifically, the strongest signals separating men with schizophrenia spectrum disorders from men with bipolar disorder were derived from facial features, while the strongest signals for women were derived from acoustic features. These physiological differences may be partially explained by different distributions of psychiatric signs and symptoms among the diagnostic categories. For example, men with schizophrenia spectrum disorders rated higher on average on the BPRS and SANS than men with bipolar disorder, while women with schizophrenia spectrum disorders on average scored lower than women with bipolar disorder on all rating scales. Alternatively, notable sex-specific variations in the prevalence, onset, symptom profiles, and outcome have been identified in the literature and have been attributed to differences in premorbid functioning, psychosocial response to symptoms, and differing levels of circulating hormones and receptors [67-70]. Audiovisual data may therefore detect subtle physiological differences unique to each sex and present in the expression of psychiatric disorders. In either scenario, sex differences are clearly of utmost importance when performing voice and facial analyses and must be taken into consideration when conducting future research.

We also identified audiovisual features common to both sexes that successfully differentiated between diagnostic categories. In line with prior work demonstrating altered facial expressivity in individuals with psychiatric disorders [47-51,54,71,72], we found that participants with schizophrenia spectrum disorders were much more likely to activate the facial muscle responsible for pulling the corners of their lips than participants with bipolar disorder. While this muscle is activated for several reasons, including the formation of certain words while speaking, it is also commonly used to form a smile. Interestingly, many patients with schizophrenia spectrum disorders, including the participants in our sample, experienced facial blunting and diminished facial expressivity, and one would, therefore, expect reduced facial activity compared to that of participants with bipolar disorder. While this finding may initially appear counterintuitive, it is important to note that the presence of blunted affect was associated with reduced variation in the cheek-raising muscle, which is also activated during the formation of a smile. Participants with schizophrenia spectrum disorders, therefore, activate lip corner–pulling muscles more than participants with bipolar disorder (perhaps to form a smile), though the range of activation of cheek movement was reduced if blunting was present. These findings warrant additional research particularly to understand the clinical significance of increased activation of certain facial muscles alongside decreased variability throughout the interview and its relationship to a diagnosis of schizophrenia spectrum disorders.

Some top features contributing to the diagnostic classification remained stable throughout the course of the interview, while others changed depending on the temporal pattern. For example, AU12 (lip-corner pulling), demonstrated a consistent downward trend for all participants, whereas the energy of the voice signal in the frequency band 1-4 kHz remained mostly flat. These same trends were noted in healthy volunteers as well, suggesting that the identified differences in facial activity and voice represent subtle pathological variations in the frequency or intensity of otherwise healthy activity. The amount of high frequency energy in the voice, for example, may represent a subtle state marker of psychiatric illness or perhaps a physiological response to

certain medications, impacting speech intelligibility. Additionally, activating lip corner–pulling muscles more at the start of an assessment (perhaps to produce a smile) may represent a healthy behavior (as it was seen in the healthy volunteers population as well), though the frequency and degree of activation is what separates those with schizophrenia spectrum disorders from those with bipolar disorder.

Our findings suggest that a tool capable of extracting and analyzing audiovisual data from newly identified psychiatric patients might offer valuable collateral clinical information, supporting a more reliable approach to differential diagnoses. Accurately diagnosing someone as having either schizophrenia spectrum disorders or bipolar disorder is a critical first step in selecting appropriate medications and therapeutic interventions, and a task that is often challenging to behavioral health clinicians given significant symptom overlap [52,53], especially during the early course of illness development. Leveraging audiovisual signals holds promise to overcome many of the challenges associated with current assessment methods [73-76], including inaccuracies and biases in self-report and recall, as well as substantial time constraints that limit the ability to effectively obtain necessary clinical information. Diagnoses, however, are complex entities, based on multiple psychiatric symptoms, each likely corresponding to several unique audiovisual features that will need to be integrated to achieve an accurate and reliable measure. Furthermore, each symptom may correspond to various alterations in audiovisual characteristics depending on multiple factors including the frequency and intensity of the experience, as well as the individual experiencing them. Future research will therefore require large clinical and computerized collaborative efforts to characterize psychiatric symptoms and diagnoses in an accurate and objective manner.

Several psychiatric signs and symptom inferences were accurately made using features extracted from voice and face either individually or combined. Similar to the findings of prior studies [36,45,71], the most successful models were derived from the SANS, and greater accuracy was achieved with externally observable psychiatric signs and symptoms such as blunted affect and lack of vocal inflection. Integrating audiovisual data into symptom assessment might, therefore, offer more efficient and objective methods to identify and track changes in negative symptoms, beyond what can be achieved through traditional clinical observation alone. A more challenging task will be to provide greater objectivity to the assessment of symptoms such as hallucinations, delusions, and suicidal thoughts. In contrast to the findings of prior research, we did not find association between brow movements and delusions or depression [54,72]. One possibility is that the prevalence of negative symptoms (such as blunted affect and affective flattening) in our sample masked the expression (and, therefore, detection) of subtle physiological signals associated with these symptoms. Our findings do, however, suggest that audiovisual data can be representative of subjectively experienced symptoms, including worthlessness and avolition, though further research is required to uncover their complex correlational structure. For instance, the observed associations between audiovisual features and psychiatric symptoms may be justly considered as purely epiphenomenal, yet a mechanistic understanding of how the symptom is expressed in the feature is not obvious and may provide insights into the diagnostic conditions. When the severity of one symptom changes, it may affect the distribution of the other symptoms in a deterministic way. Consequently, it is possible to find correlations between symptoms and physiological data even if they are not causally linked. Those correlations, if confirmed in larger studies, would be very valuable as they offer indirect proxies to more subjective experiences that are not directly quantifiable. Further research is required to determine the clinical significance of physiological changes in voice and face, as well as how they might correspond to a particular psychiatric symptom to effectively incorporate audiovisual data into clinical care. A critical, though challenging, task for future research would be maximize the level of isolated psychiatric symptoms while containing other symptoms to avoid confounding the signals that we aim to capture. Accordingly, comparing participants to themselves longitudinally as symptoms fluctuate over the course of various pathological states would also help reduce potential confounds in the signals. Future research should consider how physiological differences in facial expression and voice may manifest in other clinical settings and structured tasks as well, such as emotion elicitation [77]. Lastly, follow-up studies should consider exploring participant response times, and other measures of interviewer–interviewee interaction by recording and analyzing the voice and facial expressions of the interviewer as well.

There are several noteworthy limitations to our study. First, while prior analyses using machine learning on audio and visual features have enrolled comparable sample sizes [19,25,48], a power analysis was not conducted given the exploratory nature of this project, and additional research with more participants is necessary to support generalizability. Second, many patients included in the project were clinically stable, experiencing mild to moderate symptoms and minimal symptom fluctuations throughout the trial, which limited our ability to assess audiovisual patterns as a function of symptom severity. It is also possible that predominant negative symptoms in our sample, such as facial blunting and lack of vocal inflection, limited our ability to detect a greater number of signs and symptoms from the BPRS, HAMD, and YMRS. Third, the effects of various medications on physiological changes in voice and facial movements in our sample remain unclear and were not taken into consideration. Further research will be needed to determine the impact of the class and dose of prescribed medications on audiovisual patterns, as well as their potential impact on behavior over the course of the interview. Furthermore, demographic variables differed among the 3 groups. Although sex differences were accounted for in our models, the potential impact of physiological differences stemming from age, race, and ethnicity (though much less likely [61,78]) warrant further exploration. Fourth, the interviewer was not blinded to diagnostic groups, which may have biased the ratings. However, the interviewer was highly trained to utilize rating scales and achieved high interrater reliability prior to study initiation. Fifth, diagnoses were clinically ascertained and extracted from the medical records. Future research should consider implementing more reliable and structured methods for diagnostic assessment, such as a structured clinical interview [79], to ensure the most

accurate diagnoses. Sixth, many top features contribute to each of the best performing models, both independently and combined. Given the very large number of relevant features, we chose to emphasize and illustrate a select few in the manuscript. Corresponding clinical interpretations may, therefore, be dependent on the features highlighted and additional research will be necessary to confirm findings before clinical conclusions can be drawn. Finally, we chose to focus our analysis on acoustic components of speech rather than content as they are less dependent on cultural, socioeconomic, and educational backgrounds. Our group is, however, engaged in ongoing research aimed at the integration of speech content in the analytics framework, which we anticipate will improve our ability to detect additional psychiatric signs and symptoms.

Audiovisual data hold promise for gathering objective, scalable, noninvasive, and easily accessed, indicators of psychiatric illness. Much like an x-ray or blood test is routinely used as adjunctive data to inform clinical care, integrating audiovisual data could change the way mental health clinicians diagnose and monitor patients, enabling faster, more accurate identification of illness and enhancing a personalized approach to medicine. This would be a significant step forward for psychiatry, which is limited by its reliance on largely retrospective, self-reported data.

Acknowledgments

The authors are thankful to the volunteer participants without whose active involvement, the present study would not have been possible. We would also like to thank Rachel Ostrand, PhD, who contributed to the development of the speech prompts utilized and helped setup the audiovisual data equipment.

Authors' Contributions

GC, SH, MB, and JK conceptualized and executed the project. AA designed and performed data analysis with input from GC, and MB, AA, SH, and CA performed data preprocessing. AFA and EA performed participant recruitment and data collection. AA and MB wrote the manuscript, and all authors reviewed and edited.

Conflicts of Interest

AA, GC, and CA disclose that their employer, IBM Research, is the research branch of IBM Corporation.

Multimedia Appendix 1
Voice features.
[DOCX File , 15 KB - mental_v9i1e24699_app1.docx]

Multimedia Appendix 2
Facial action units.
[DOCX File , 13 KB - mental_v9i1e24699_app2.docx]

References

1. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, WHO WMH-ICS Collaborators. WHO world mental health surveys international college student project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]
2. Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980-2013. Int J Epidemiol 2014 Apr;43(2):476-493 [FREE Full text] [doi: 10.1093/ije/dyu038] [Medline: 24648481]
3. Jones PB. Adult mental health disorders and their age at onset. Br J Psychiatry Suppl 2013 Jan;54:s5-10. [doi: 10.1192/bjp.bp.112.119164] [Medline: 23288502]
4. O'Connor K, Muller Neff D, Pitman S. Burnout in mental health professionals: a systematic review and meta-analysis of prevalence and determinants. Eur Psychiatry 2018 Sep 26;53:74-99. [doi: 10.1016/j.eurpsy.2018.06.003] [Medline: 29957371]
5. Rotstein S, Hudaib A, Facey A, Kulkarni J. Psychiatrist burnout: a meta-analysis of Maslach burnout inventory means. Australas Psychiatry 2019 Jun 25;27(3):249-254. [doi: 10.1177/1039856219833800] [Medline: 30907115]
6. Chan MK, Chew QH, Sim K. Burnout and associated factors in psychiatry residents: a systematic review. Int J Med Educ 2019 Jul 30;10:149-160 [FREE Full text] [doi: 10.5116/ijme.5d21.b621] [Medline: 31381505]
7. American PA. Diagnostic and Statistical Manual of Mental Disorders (5th ed). Arlington, VA: American Psychiatric Association; 2013.
8. Gaebel W, Zielasek J, Reed G. Mental and behavioural disorders in the ICD-11: concepts, methodologies, and current status. Psychiatr Pol 2017 Apr 30;51(2):169-195 [FREE Full text] [doi: 10.12740/PP/69660] [Medline: 28581530]
9. Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW. The science of prognosis in psychiatry: a review. JAMA Psychiatry 2018 Dec 01;75(12):1289-1297. [doi: 10.1001/jamapsychiatry.2018.2530] [Medline: 30347013]

10. Levchenko A, Nurgaliev T, Kanapin A, Samsonova A, Gainetdinov RR. Current challenges and possible future developments in personalized psychiatry with an emphasis on psychotic disorders. Heliyon 2020 May;6(5):e03990 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e03990] [Medline: 32462093]
11. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Dec;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
12. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018 May 07;14(1):91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
13. Pampouchidou A. Facial geometry and speech analysis for depression detection. 2017 Presented at: 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; July 11-15; Jeju, Korea. [doi: 10.1109/embc.2017.8037103]
14. Girard JM, Cohn JF. Automated audiovisual depression analysis. Curr Opin Psychol 2015 Aug;4:75-79 [FREE Full text] [doi: 10.1016/j.copsyc.2014.12.010] [Medline: 26295056]
15. Dibeklio H, Hammal Z, Yang Y, Cohn JF. Multimodal detection of depression in clinical interviews. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 2015 Presented at: ACM International Conference on Multimodal Interaction; November 9-13; Seattle, Washington p. 307-310. [doi: 10.1145/2818346.2820776]
16. Renfordt E, Busch H. [New diagnostic strategies in psychiatry by means of video-technique. The use of time-blind video analysis for the evaluation of antidepressant drug trials (author's transl)]. Pharmakopsychiatr Neuropsychopharmakol 1976 Mar 20;9(2):67-75. [doi: 10.1055/s-0028-1094480] [Medline: 790410]
17. Kring AM, Sloan DM. The facial expression coding system (FACES): development, validation, and utility. Psychol Assess 2007 Jun;19(2):210-224. [doi: 10.1037/1040-3590.19.2.210] [Medline: 17563202]
18. Cummins N, Baird A, Schuller BW. Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods 2018 Dec 01;151:41-54. [doi: 10.1016/j.ymeth.2018.07.007] [Medline: 30099083]
19. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol 2020 Feb 31;5(1):96-116 [FREE Full text] [doi: 10.1002/lio2.354] [Medline: 32128436]
20. Scherer S, Stratou M, Mahmoud J, Boberg J, Gratch J. Automatic behavior descriptors for psychological disorder analysis. 2013 Presented at: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition; April 22-26; Shanghai, China p. 1-8. [doi: 10.1109/fg.2013.6553789]
21. Abrami A, Gunzler S, Kilbane C, Ostrand R, Ho B, Cecchi G. Automated computer vision assessment of hypomimia in Parkinson disease: proof-of-principle pilot study. J Med Internet Res 2021 Feb 22;23(2):e21037 [FREE Full text] [doi: 10.2196/21037] [Medline: 33616535]
22. Yang Y, Fairbairn C, Cohn JF. Detecting depression severity from vocal prosody. IEEE Trans Affect Comput 2013;4(2):142-150 [FREE Full text] [doi: 10.1109/T-AFFC.2012.38] [Medline: 26985326]
23. Xu S. Automated verbal and nonverbal speech analysis of interviews of individuals with schizophrenia and depression. 2019 Presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society; July 23-27; Berlin, Germany. [doi: 10.1109/embc.2019.8857071]
24. Minor KS, Bonfils KA, Luther L, Firmin RL, Kukla M, MacLain VR, et al. Lexical analysis in schizophrenia: how emotion and social word use informs our understanding of clinical presentation. J Psychiatr Res 2015 May;64:74-78. [doi: 10.1016/j.jpsychires.2015.02.024] [Medline: 25777474]
25. Cohen AS, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. Curr Opin Psychiatry 2014 May;27(3):203-209 [FREE Full text] [doi: 10.1097/YCO.0000000000000056] [Medline: 24613984]
26. de Boer J, Voppel A, Begemann M, Schnack H, Wijnen F, Sommer I. Clinical use of semantic space models in psychiatry and neurology: a systematic review and meta-analysis. Neurosci Biobehav Rev 2018 Oct;93:85-92. [doi: 10.1016/j.neubiorev.2018.06.008] [Medline: 29890179]
27. Rapcan V, D'Arcy S, Yeap S, Afzal N, Thakore J, Reilly RB. Acoustic and temporal analysis of speech: a potential biomarker for schizophrenia. Med Eng Phys 2010 Nov;32(9):1074-1079. [doi: 10.1016/j.medengphy.2010.07.013] [Medline: 20692864]
28. Vanello N. Speech analysis for mood state characterization in bipolar patients. 2012 Presented at: Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 28-September 1; San Diego, California. [doi: 10.1109/embc.2012.6346375]
29. Pan Z, Gui C, Zhang J, Zhu J, Cui D. Detecting manic state of bipolar disorder based on support vector machine and gaussian mixture model using spontaneous speech. Psychiatry Investig 2018 Jul;15(7):695-700 [FREE Full text] [doi: 10.30773/pi.2017.12.15] [Medline: 29969852]
30. Faurholt-Jepsen M, Busk J, Frost M, Vinberg M, Christensen EM, Winther O, et al. Voice analysis as an objective state marker in bipolar disorder. Transl Psychiatry 2016 Jul 19;6(7):e856-e856 [FREE Full text] [doi: 10.1038/tp.2016.123] [Medline: 27434490]
31. Minor KS, Willits JA, Marggraf MP, Jones MN, Lysaker PH. Measuring disorganized speech in schizophrenia: automated analysis explains variance in cognitive deficits beyond clinician-rated scales. Psychol Med 2018 Apr 25;49(3):440-448. [doi: 10.1017/s0033291718001046]

32. Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry 2018 Feb 19;17(1):67-75 [FREE Full text] [doi: 10.1002/wps.20491] [Medline: 29352548]
33. He L, Cao C. Automated depression analysis using convolutional neural networks from speech. J Biomed Inform 2018 Jul;83:103-111 [FREE Full text] [doi: 10.1016/j.jbi.2018.05.007] [Medline: 29852317]
34. Mota NB, Vasconcelos NAP, Lemos N, Pieretti AC, Kinouchi O, Cecchi GA, et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. PLoS One 2012 Apr 9;7(4):e34928 [FREE Full text] [doi: 10.1371/journal.pone.0034928] [Medline: 22506057]
35. Cohen AS, Fedechko TL, Schwartz EK, Le TP, Foltz PW, Bernstein J, et al. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. J Abnorm Psychol 2019 Mar;128(2):97-105. [doi: 10.1037/abn0000397] [Medline: 30714793]
36. Cohen AS, Cowan T, Le TP, Schwartz EK, Kirkpatrick B, Raugh IM, et al. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. Schizophr Res 2020 Jun;220:141-146. [doi: 10.1016/j.schres.2020.03.043] [Medline: 32247747]
37. Kliper R, Vaizman Y, Weinshall D, Portuguese S. Evidence for depression and schizophrenia in speech prosody. 2010 Presented at: Third ISCA Workshop on Experimental Linguistics; 2010; Greece. [doi: 10.36505/exling-2010/03/0022/000142]
38. Kliper R, Portuguese S, Weinshall D, Serino S, Matic A, Giakoumis D, et al. Prosodic analysis of speech and the underlying mental state. In: Serino S, Matic A, Giakoumis D, Lopez G, Cipresso P, editors. Pervasive Computing Paradigms for Mental Health. MindCare 2015. Cham: Communications in Computer and Information Science, vol 604, Springer; 2016.
39. Perlini C, Marini A, Garzitto M, Isola M, Cerruti S, Marinelli V, et al. Linguistic production and syntactic comprehension in schizophrenia and bipolar disorder. Acta Psychiatr Scand 2012 Nov;126(5):363-376. [doi: 10.1111/j.1600-0447.2012.01864.x] [Medline: 22509998]
40. Tahir Y, Yang Z, Chakraborty D, Thalmann N, Thalmann D, Maniam Y, et al. Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. PLoS One 2019 Apr 9;14(4):e0214314 [FREE Full text] [doi: 10.1371/journal.pone.0214314] [Medline: 30964869]
41. Guidi A, Schoentgen J, Bertschy G, Gentili C, Scilingo E, Vanello N. Features of vocal frequency contour and speech rhythm in bipolar disorder. Biomedical Signal Processing and Control 2017 Aug;37:23-31. [doi: 10.1016/j.bspc.2017.01.017]
42. Guidi A. Analysis of running speech for the characterization of mood state in bipolar patients. 2015 Presented at: AEIT International Annual Conference; October 14-16; Naples, Italy. [doi: 10.1109/aeit.2015.7415275]
43. Zhang J, Pan Z, Gui C, Xue T, Lin Y, Zhu J, et al. Analysis on speech signal features of manic patients. J Psychiatr Res 2018 Mar;98:59-63. [doi: 10.1016/j.jpsychires.2017.12.012] [Medline: 29291581]
44. Hamm J, Kohler CG, Gur RC, Verma R. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. J Neurosci Methods 2011 Sep 15;200(2):237-256 [FREE Full text] [doi: 10.1016/j.jneumeth.2011.06.023] [Medline: 21741407]
45. Kupper Z, Ramseyer F, Hoffmann H, Kalbermatten S, Tschacher W. Video-based quantification of body movement during social interaction indicates the severity of negative symptoms in patients with schizophrenia. Schizophr Res 2010 Aug;121(1-3):90-100. [doi: 10.1016/j.schres.2010.03.032] [Medline: 20434313]
46. Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 2015 Jun;37(6):1113-1133. [doi: 10.1109/TPAMI.2014.2366127] [Medline: 26357337]
47. Gupta T, Haase CM, Strauss GP, Cohen AS, Mittal VA. Alterations in facial expressivity in youth at clinical high-risk for psychosis. J Abnorm Psychol 2019 May;128(4):341-351 [FREE Full text] [doi: 10.1037/abn0000413] [Medline: 30869926]
48. Wang P, Barrett F, Martin E, Milonova M, Gur RE, Gur RC, et al. Automated video-based facial expression analysis of neuropsychiatric disorders. J Neurosci Methods 2008 Feb 15;168(1):224-238 [FREE Full text] [doi: 10.1016/j.jneumeth.2007.09.030] [Medline: 18045693]
49. Schneider F, Heimann H, Himer W, Huss D, Mattes R, Adam B. Computer-based analysis of facial action in schizophrenic and depressed patients. Eur Arch Psychiatry Clin Neurosci 1990;240(2):67-76. [doi: 10.1007/BF02189974] [Medline: 2149651]
50. Pampouchidou A. Video-based depression detection using local Curvelet binary patterns in pairwise orthogonal planes. 2016 Presented at: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 16-20; Orlando, Florida. [doi: 10.1109/embc.2016.7591564]
51. Alghowinem S. Cross-cultural detection of depression from nonverbal behaviour. 2015 Presented at: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition; May 4-8; Ljubljana, Slovenia. [doi: 10.1109/fg.2015.7163113]
52. Pearlson GD. Etiologic, phenomenologic, and endophenotypic overlap of schizophrenia and bipolar disorder. Annu Rev Clin Psychol 2015 Mar 28;11(1):251-281. [doi: 10.1146/annurev-clinpsy-032814-112915] [Medline: 25581236]
53. Yalincetin B, Bora E, Binbay T, Ulas H, Akdede BB, Alptekin K. Formal thought disorder in schizophrenia and bipolar disorder: a systematic review and meta-analysis. Schizophr Res 2017 Jul;185:2-8. [doi: 10.1016/j.schres.2016.12.015] [Medline: 28017494]

54. Vijay S, Pennant L, Ongur D, Baker J, Morency L. Computational study of psychosis symptoms and facial expressions. 2016 Presented at: Computer Human Interaction Workshops; May 7-12; San Jose, California.
55. Shafer A. Meta-analysis of the brief psychiatric rating scale factor structure. Psychol Assess 2005 Sep;17(3):324-335. [doi: 10.1037/1040-3590.17.3.324] [Medline: 16262458]
56. Andreasen NC. The scale for the assessment of negative symptoms (SANS): conceptual and theoretical foundations. Br J Psychiatry Suppl 1989 Nov(7):49-58. [doi: 10.1192/S0007125000291496] [Medline: 2695141]
57. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960 Feb 01;23(1):56-62 [FREE Full text] [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
58. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978 Nov 01;133(5):429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
59. Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. 2010 Presented at: International Conference on Multimedia; October 25-29; Firenze, Italy p. 1459-1462. [doi: 10.1145/1873951.1874246]
60. Schuller B. The interspeech computational paralinguistics challenge: social signals, conflict, emotion, autism. 2013 Presented at: 14th Annual Conference of the International Speech Communication Association; August 25-29; Lyon, France.
61. Baltrusaitis T, Robinson P, Morency LP. OpenFace: an open source facial behavior analysis toolkit. 2016 Presented at: 2016 IEEE Winter Conference on Applications of Computer Vision; March 7-10; Lake Placid, New York p. 1-10. [doi: 10.1109/wacv.2016.7477553]
62. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001 Oct 1;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
63. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. GetMobile 2015 Jun;19(1):29-33. [doi: 10.1145/2786984.2786995]
64. Williamson J, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD. Vocal and facial biomarkers of depression based on motor incoordination and timing. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. 2014 Presented at: 4th International Workshop on Audio/Visual Emotion Challenge; November 7; Orlando, Florida p. 65-72. [doi: 10.1145/2661806.2661809]
65. Ray A, Kumar S, Reddy R, Mukherjee P, Garg R. Multilevel attention network using text, audio and video for depression prediction. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. 2019 Presented at: 9th International on Audio/Visual Emotion Challenge and Workshop; 21 October; Nice, France p. 81-88. [doi: 10.1145/3347320.3357697]
66. Dibeklioglu H, Hammal Z, Cohn JF. Dynamic multimodal measurement of depression severity using deep autoencoding. IEEE J Biomed Health Inform 2018 Mar;22(2):525-536 [FREE Full text] [doi: 10.1109/JBHI.2017.2676878] [Medline: 28278485]
67. Abel KM, Drake R, Goldstein JM. Sex differences in schizophrenia. Int Rev Psychiatry 2010;22(5):417-428. [doi: 10.3109/09540261.2010.515205] [Medline: 21047156]
68. Mendrek A, Mancini-Marïe A. Sex/gender differences in the brain and cognition in schizophrenia. Neurosci Biobehav Rev 2016 Aug;67:57-78. [doi: 10.1016/j.neubiorev.2015.10.013] [Medline: 26743859]
69. Ragazan DC, Eberhard J, Berge J. Sex-specific associations between bipolar disorder pharmacological maintenance therapies and inpatient rehospitalizations: a 9-year swedish national registry study. Front Psychiatry 2020;11:598946 [FREE Full text] [doi: 10.3389/fpsyt.2020.598946] [Medline: 33262715]
70. Mitchell RHB, Hower H, Birmaher B, Strober M, Merranko J, Rooks B, et al. Sex differences in the longitudinal course and outcome of bipolar disorder in youth. J Clin Psychiatry 2020 Oct 27;81(6) [FREE Full text] [doi: 10.4088/JCP.19m13159] [Medline: 33113597]
71. Vail AK. Visual attention in schizophrenia eye contact and gaze aversion during clinical interactions. 2017 Presented at: Seventh International Conference on Affective Computing and Intelligent Interaction; October 23-26; San Antonio, Texas p. 490-497. [doi: 10.1109/acii.2017.8273644]
72. Baker JT, Pennant L, Baltrušaitis T, Vijay S, Liebson ES, Ongur D, et al. Toward expert systems in mental health assessment: a computational approach to the face and voice in dyadic patient-doctor interactions. iproc 2016 Dec 30;2(1):e44 [FREE Full text] [doi: 10.2196/iproc.6136]
73. Thombs BD, Roseman M, Kloda LA. Depression screening and mental health outcomes in children and adolescents: a systematic review protocol. Syst Rev 2012 Nov 24;1(1):58 [FREE Full text] [doi: 10.1186/2046-4053-1-58] [Medline: 23176742]
74. Roseman M, Kloda LA, Saadat N, Riehm KE, Ickowicz A, Baltzer F, et al. Accuracy of depression screening tools to detect major depression in children and adolescents: a systematic review. Can J Psychiatry 2016 Dec 09;61(12):746-757 [FREE Full text] [doi: 10.1177/0706743716651833] [Medline: 27310247]
75. Addington J, Stowkowy J, Weiser M. Screening tools for clinical high risk for psychosis. Early Interv Psychiatry 2015 Oct 23;9(5):345-356. [doi: 10.1111/eip.12193] [Medline: 25345316]

76. Mulvaney-Day N, Marshall T, Downey Piscopo K, Korsen N, Lynch S, Karnell LH, et al. Screening for behavioral health conditions in primary care settings: a systematic review of the literature. J Gen Intern Med 2018 Mar 25;33(3):335-346 [FREE Full text] [doi: 10.1007/s11606-017-4181-0] [Medline: 28948432]
77. Gross JJ, Levenson RW. Emotion elicitation using films. Cogn Emot 1995 Jan;9(1):87-108. [doi: 10.1080/02699939508408966]
78. Vorperian HK, Kent RD. Vowel acoustic space development in children: a synthesis of acoustic and anatomic data. J Speech Lang Hear Res 2007 Dec;50(6):1510-1545 [FREE Full text] [doi: 10.1044/1092-4388(2007/104)] [Medline: 18055771]
79. First MB. Structured Clinical Interview for the DSM-IV Axis I Disorders: SCID-I/P, Version 2.0. New York: Biometrics Research Dept., New York State Psychiatric Institute; 1997.

Abbreviations

AUROC: area under the receiver operating characteristic curve
BPRS: Brief Psychiatric Rating Scale
HAMD: Hamilton Depression Rating Scale
SANS: Scale for the Assessment of Negative Symptoms
YMRS: Young Mania Rating Scale

Edited by J Torous; submitted 01.10.20; peer-reviewed by A Hudon, D Hidalgo-Mazzei, D Fulford, A Wright; comments to author 14.11.20; revised version received 29.04.21; accepted 01.12.21; published 24.01.22.

Please cite as:
Birnbaum ML, Abrami A, Heisig S, Ali A, Arenare E, Agurto C, Lu N, Kane JM, Cecchi G
Acoustic and Facial Features From Clinical Interviews for Machine Learning–Based Psychiatric Diagnosis: Algorithm Development
JMIR Ment Health 2022;9(1):e24699
URL: https://mental.jmir.org/2022/1/e24699
doi:10.2196/24699
PMID:35072648

Original Paper

Diagnostic Performance of an App-Based Symptom Checker in Mental Disorders: Comparative Study in Psychotherapy Outpatients

Severin Hennemann[1], PhD; Sebastian Kuhn[2], MD; Michael Witthöft[1], PhD; Stefanie M Jungmann[1], PhD

[1]Department of Clinical Psychology, Psychotherapy and Experimental Psychopathology, University of Mainz, Mainz, Germany

[2]Department of Digital Medicine, Medical Faculty OWL, Bielefeld University, Bielefeld, Germany

Corresponding Author:
Severin Hennemann, PhD
Department of Clinical Psychology, Psychotherapy and Experimental Psychopathology
University of Mainz
Wallstr 3
Mainz, 55122
Germany
Phone: 49 61313939215
Email: s.hennemann@uni-mainz.de

Abstract

Background: Digital technologies have become a common starting point for health-related information-seeking. Web- or app-based symptom checkers aim to provide rapid and accurate condition suggestions and triage advice but have not yet been investigated for mental disorders in routine health care settings.

Objective: This study aims to test the diagnostic performance of a widely available symptom checker in the context of formal diagnosis of mental disorders when compared with therapists' diagnoses based on structured clinical interviews.

Methods: Adult patients from an outpatient psychotherapy clinic used the app-based symptom checker *Ada–check your health* (ADA; Ada Health GmbH) at intake. Accuracy was assessed as the agreement of the first and 1 of the first 5 condition suggestions of ADA with at least one of the interview-based therapist diagnoses. In addition, sensitivity, specificity, and interrater reliabilities (Gwet first-order agreement coefficient [AC1]) were calculated for the 3 most prevalent disorder categories. Self-reported usability (assessed using the System Usability Scale) and acceptance of ADA (assessed using an adapted feedback questionnaire) were evaluated.

Results: A total of 49 patients (30/49, 61% women; mean age 33.41, SD 12.79 years) were included in this study. Across all patients, the interview-based diagnoses matched ADA's first condition suggestion in 51% (25/49; 95% CI 37.5-64.4) of cases and 1 of the first 5 condition suggestions in 69% (34/49; 95% CI 55.4-80.6) of cases. Within the main disorder categories, the accuracy of ADA's first condition suggestion was 0.82 for somatoform and associated disorders, 0.65 for affective disorders, and 0.53 for anxiety disorders. Interrater reliabilities ranged from low (AC1=0.15 for anxiety disorders) to good (AC1=0.76 for somatoform and associated disorders). The usability of ADA was rated as high in the System Usability Scale (mean 81.51, SD 11.82, score range 0-100). Approximately 71% (35/49) of participants would have preferred a face-to-face over an app-based diagnostic.

Conclusions: Overall, our findings suggest that a widely available symptom checker used in the formal diagnosis of mental disorders could provide clinicians with a list of condition suggestions with moderate-to-good accuracy. However, diagnostic performance was heterogeneous between disorder categories and included low interrater reliability. Although symptom checkers have some potential to complement the diagnostic process as a screening tool, the diagnostic performance should be tested in larger samples and in comparison with further diagnostic instruments.

(JMIR Ment Health 2022;9(1):e32832) doi:10.2196/32832

KEYWORDS

mHealth; symptom checker; diagnostics; mental disorders; psychotherapy; mobile phone

Introduction

Background

Digital technologies represent an increasingly important source of health information. Approximately 6 out of 10 European adults use the internet to seek health information [1]. Meanwhile, internet search engines can be considered a common starting point for self-diagnosis, which can have a significant effect on health care decisions and outcomes. The popularity of web-based health information seeking arises from the ease of access and immediacy of a plethora of health resources in various formats (eg, encyclopedias, blogs, social media, video channels, health apps, and telemedicine). Diagnosis websites could promote early diagnosis and help-seeking, which in turn may lead to earlier treatment and thus prevent chronic courses.

Mental health topics are among the most popular search queries [1], and it is estimated that approximately one-third of all health apps worldwide target mental health issues [2]. The use of these digital health resources may have various structural and individual reasons. For example, individuals who feel stigmatized or ashamed by mental health issues (eg, obsessive-compulsive symptoms and sexual dysfunctions) could benefit from anonymity and low-threshold information [3,4]. Interpersonal communication problems, often associated with severe mental disorders, can become barriers to traditional help-seeking and may also turn patients toward digital resources. In addition, there is considerable uncertainty in the population regarding the significance and pathological threshold of mental health issues [5]. Access to adequate treatment and diagnosis is often complicated and delayed (eg, concerns about psychological treatment, long waits, and restricted availability of psychotherapy in rural areas) [6,7].

Although digital health resources can ideally increase access to health care and empower patients to engage in health behavior [8], the information provided is mostly unregulated and can also contain confusing or unsubstantiated facts and recommendations [9]. This could promote incorrect self-diagnosis and problematic health decisions [10]. A study by Grohol et al [11] on the quality of web-based mental health information revealed that 67.5% of 440 investigated websites contained information of good or better quality. However, the quality of information varied between disorders, and readability was rated as difficult. For anxiety disorders, another study found only a poor-to-moderate quality of internet-based information [12]. In addition, many websites also showed a lack of or inadequate information regarding a rough classification of symptoms and possible health care professionals or services to contact [13]. Similarly, studies from the mobile health app database project rated the overall information quality of apps for various mental disorders (eg, depression and posttraumatic stress disorder) as poor to mediocre and found that only a fraction had been evaluated scientifically [14,15].

Selecting, interpreting, and using web-based health information requires sufficient eHealth literacy [16]; however, this can be unevenly distributed across age, socioeconomic, or educational groups, which has been termed “digital divide” [17]. Thus, a substantial proportion of internet users may experience difficulties in web-based health information seeking, and individuals with chronic health problems who may have a particular need for information and support are seemingly less likely to obtain helpful information [18]. Users typically rate the internet “higher as a source to use than a source to trust” [19], particularly when compared with personal medical information (eg, from health professionals). In addition, digital health information may lead to increased illness anxiety [20], which in turn increases unnecessary health care use and costs [21,22]. In this regard, health professionals are also facing new challenges (eg, biased expectations and less trust in medical advice) with internet-informed patients [23].

Symptom Checkers for Condition Suggestion and Triage Advice

An emerging alternative to internet search engines is the so-called symptom checkers, which aim to provide rapid and differentiated condition suggestions and assistance with the urgency of care advice. Symptom checkers typically use dynamically structured interviews or multiple-choice questions and, as a result, provide one or more condition suggestions, usually ranked by their likelihood (eg, 7 *out of 10 persons with these symptoms have been diagnosed with this condition*). The mostly algorithm-based programs typically operate with chatbots to simulate a dialogue-like human interaction [24]. Symptom checkers can also be used as a diagnostic support system for health professionals [25]. General diagnostic and triage advice of specific symptom checkers has been studied for a broad range of general and specialized health problems [26], for example, ophthalmologic [27] or viral diseases [28,29].

Research indicates that, although symptom checkers seem to be easy to use and well-accepted by most users [30,31], the diagnostic performance varies significantly between different symptom checkers and has been interpreted as low to moderate at best [32,33]. Semigran et al [34] investigated the diagnostic accuracy of 23 symptom checkers using 45 standardized case vignettes of various health conditions that would require emergent care (eg, appendicitis and heart attack) or nonemergent care (eg, back pain), or where self-care would be appropriate (eg, bronchitis). Across symptom checkers, the correct diagnosis was listed first in only 34% of cases, with considerable performance variation between symptom checkers (5%-50%). A similar average performance rate was found for a broader set of 200 clinical vignettes in a recent study that compared the condition suggestion accuracy of 8 popular symptom checkers (*Ada–check your health* [ADA], Babylon, Buoy, K Health, Mediktor, Symptomate, WebMD, and Your.MD) with diagnoses obtained from general practitioners for various health conditions, including some mental health issues [35]. The investigated symptom checkers showed a highly variable diagnostic coverage, from 99% (ADA) to 51.5% (Buoy). Significant differences in condition suggestion accuracy were observed between symptom checkers, with accuracy for the first listed condition suggestion ranging from 19% (Symptomate) to 48.5% (ADA) with an average of 26.1%. The symptom checkers listed the correct diagnosis in the top 5 condition suggestions in 40.8% of cases, whereas the best accuracy was reported for ADA (77.5%). However, these findings should be interpreted cautiously as most authors were employees of Ada Health

GmbH. Most recently, a study by Ceney et al [33] yielded comparable average performance rates (51%, range 22.2%-84%) for the top 5 condition suggestions of 12 symptom checkers based on case vignettes.

In contrast to patients' rather positive perspectives on the usability and utility of symptom checkers, health professionals seem to be more skeptical [25], and symptom checkers have had an inferior performance compared with professional diagnoses in previous studies [32]. According to a review by Semigran et al [36], 84.3% of physicians' top 3 diagnoses matched those of clinical vignettes compared with 51.2% of symptom checkers (P<.001). Generally, diagnostic performance seems to converge when the number of diagnostic suggestions taken into account is increased. For example, ADA reached a similar diagnostic accuracy to general practitioners (77.5% vs 82.8%) when considering the range of the top 5 diagnostic suggestions in the study by Gilbert et al [34]. In another study, the Babylon Diagnostic and Triage System reached comparable diagnostic sensitivity (80%) with physicians (83.9%) [37]. However, various methodological concerns regarding this study have been raised, such as sensitivity to outliers [38]. In a Spanish study, 622 patients at a tertiary care university hospital emergency department responded to the questions of the symptom checker Mediktor. The physicians' diagnoses matched 1 of the first 3 diagnoses of Mediktor in 75.4% of cases and the first diagnosis in 42.9% of cases. Again, as this study was conducted by committed future company members of the investigated symptom checker at the time of publication, findings should be interpreted cautiously.

Although previous studies mostly cover a range of physical conditions (which most symptom checkers were primarily designed to detect), the usability and diagnostic performance in mental disorders have not been investigated sufficiently. A recent pilot study by Jungmann et al [39] investigated the performance and dependency on expert knowledge of the symptom checker ADA in diagnosing mental disorders in adults and adolescents. Psychotherapists, psychology students, and laypersons entered symptoms from case vignettes into the app. For mental disorders in adulthood, the diagnostic agreement between the textbook diagnoses and the main condition suggestion by the app was moderate (68%) but increased to 85% when ADA's differential diagnoses were taken into account. Diagnostic agreement with case vignettes was higher for psychotherapists (79%) than for psychology students (58%) or laypersons (63%), demonstrating the beneficial effect of expert knowledge.

Objectives

Notably, previous studies on symptom checkers have relied primarily on standardized case vignettes, which are less likely to represent real-world cases with clinical comorbidity and, as such, may overestimate the diagnostic accuracy of symptom checkers. Furthermore, the diagnostic quality at the consumer level (ie, patients rather than health professionals) has been insufficiently studied but is of paramount interest for a robust evaluation of the accuracy of symptom checkers in clinical settings. Therefore, this study aims to evaluate the diagnostic performance of a widely available symptom checker when used by patients compared with diagnoses by psychotherapists using structured clinical interviews.

Methods

Design

This study was designed as an observational, comparative, prospective study in adult outpatients conducted at the psychotherapy outpatient clinic of the University of Mainz (Germany). In the outpatient clinic, >1400 patients are treated per year on average by approximately 160 therapists. The study was conducted in compliance with ethical principles and approved by the ethics committee of the Department of Psychology at the University of Mainz (2019-JGUpsychEK-009, June 28, 2019).

Participants and Recruitment

Participants were recruited consecutively between August 2019 and December 2020 in the outpatient psychotherapy clinic of the University of Mainz. Inclusion criteria were age ≥18 years and sufficient knowledge of the German language. We excluded patients with acute suicidality (assessed by a score of ≥2 on item 9 of the Beck Depression Inventory-II [40]), patients with any self-indicated acute mental or physical state (eg, psychosis or brain injury) that would prevent safe and meaningful use of the app, and patients who did not receive a diagnosis of a mental disorder by therapists in the diagnostic interview. Diagnoses were obtained from 42 experienced therapists. At the time of the study, the therapists were in advanced cognitive behavioral therapy training (≥1.5 years of clinical practice) and had completed a 2-day training course on the use of structural clinical interviews.

Procedure

After having indicated interest in participating in the trial, participants were screened for inclusion with a web-based questionnaire and received detailed information on the study. Eligible participants provided written informed consent to participate. Consequently, the participants were asked to fill out a demographic questionnaire. During their waiting time before their initial appointment at the outpatient clinic, the participants were then invited to answer the questions of the symptom checker on a 10-inch tablet. The patients were instructed to focus on the current most disturbing mental health symptoms. Patients and therapists were not informed about the condition suggestions by the app until the completion of the diagnostic interviews so that the subsequent diagnostic process would not be influenced. For this purpose, the patients were instructed to stop using the symptom checker before the condition suggestions were displayed. The therapists were informed about the study and routinely performed the German version of the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (SCID) [41], during the initial therapy sessions, which can be considered a *gold standard* of the diagnosis of mental disorders in research along with individually selected self-report instruments. The therapists were asked to report their diagnoses back to the study team and were then unblinded and informed about the symptom checker's condition suggestions, which they

discussed with the patient to allow for professional clarification of ambiguous or contradictory results. For compensation, the patients could participate in a raffle of gift certificates (5 × €20 [US $22.91]), and the therapists were reimbursed with €5 (US $5.73) per case.

Instruments

App-Based Symptom Checker

The symptom checker ADA (Ada Health GmbH) is a Conformité Européenne–certified medical device assisting in the screening of medical conditions. For this purpose, ADA is available at the consumer level as a self-assessment app [42], whereas a prototype diagnostic decision support system for health professionals has been developed as well [43]. This particular app was selected for various reasons: (1) the diagnostic coverage is wide [35], including mental disorders, and ADA has shown acceptable diagnostic performance in this diagnostic spectrum recently [39]; (2) it is free of charge and widely available (>10 million users and 7 languages) for Android- and iOS-running devices [42]; (3) it provides probabilities for a list of differential condition suggestions; (4) in comparison with other symptom checkers, it has performed more accurately in formal diagnosis [34,35]; and (5) it has proven to be well-accepted and easy to use in a large sample of primary care patients [30].

ADA is based on a dynamic medical database, which is updated through research findings and app entries [44]. Using artificial intelligence, a chatbot asks questions in various formats (eg, open questions with text-based answers and discrete items) about current symptoms. Standard questions include age, gender, smoker status, presence of pregnancy, high blood pressure, and diabetes. As a result, ≥1 condition suggestion is determined to best match the pattern of symptoms entered. The user is presented with a probability of possible diagnoses (eg, 6 out of 10 people with these symptoms have a social anxiety disorder), including a list of other less probable condition suggestions (see [45] for an example process). Finally, the app offers information on the urgency of medical help-seeking (eg, urgent care needed). In this study, version 3.1.2 of ADA was used.

Usability

The usability of the symptom checker was assessed using the 10-item, unidimensional System Usability Scale (SUS) [46], a widely used, reliable scale [47]. The items (eg, *I find the app easy to use*) are rated on a 5-point Likert scale (0=*strongly disagree* to 4=*strongly agree*). Reliability was acceptable in this study (McDonald ω=0.72). Furthermore, an adapted version of a 15-item questionnaire, which was previously used to investigate the usability of a computerized standardized clinical interview [48], was implemented. For the purpose of this study, 12 items were selected, which could be answered on a 4-point Likert scale (1=*strongly disagree* to 4=*strongly agree*). Reliability was acceptable in this study (ω=0.74). Both questionnaires were completed as paper and pencil versions after completion of the symptom checker.

Additional Measures

Further items covered demographic characteristics (age, gender, mother tongue, relationship status, and educational level), clinical characteristics (symptom duration, history of mental disorder diagnoses, and psychotherapeutic treatments), previous experience with ADA (yes or no), and frequency of web-based health information seeking (*Do you use the Internet to inform yourself about symptoms of your mental health problems?* with answers from 0=*never* to 3=*always*). The time required to complete the diagnostic process in the app and the number of questions asked until completion were assessed.

Statistical Analyses

All text diagnoses were recoded into International Classification of Diseases, 10th Revision (ICD-10), codes (as a universal medical coding system) by a trained clinical psychologist not otherwise involved in the study and cross-checked by another clinical psychologist at the Masters level (97.1% agreement). Disagreements between the raters were resolved by including a third licensed therapist (first author).

The condition suggestions were compared with the therapists' diagnoses at the level of 4-digit codes in the ICD-10 (eg, F40.1, social phobia). Following the procedure by Jungmann et al [39], if the fourth digit represented a more detailed specification (eg, F32.2, major depressive disorder, single episode, severe without psychotic features), the 3-digit code match was counted for the following disorders: depressive disorder, bipolar affective disorder, obsessive-compulsive disorder, conduct disorder, or schizophrenia. For the diagnosis of agoraphobia with panic disorder (F40.01), both the condition suggestions *agoraphobia* and *panic disorder* were counted as accurate. The condition suggestion *Burnout* was coded as a depressive disorder. As condition suggestions to our knowledge did not include recurrent depressive episodes (F33.X), these diagnoses were treated as equal to the nonrecurrent category (F32.X). Furthermore, the terms *abuse* and *addiction* were judged to agree as the app did not distinguish between abuse and addiction to our knowledge. Functional somatic syndromes (eg, fibromyalgia and irritable bowel syndrome) were associated with somatoform disorders (F45) [49]. Analyses of the agreement were assessed for both the total sample and disorder categories (first 2 ICD-10 digits, eg, affective disorders and anxiety disorders). We noted whether the symptom checker's first condition suggestion or any of the first 5 of the symptom checker's condition suggestions (including *less probable condition suggestions* if not >5 in total) matched any of the interview-based diagnoses to assess diagnostic accuracy. For example, we counted a correct diagnosis listed first if a patient was diagnosed with agoraphobia with panic disorder (F40.01) and specific phobia (F40.2) by therapists using the SCID and ADA's top 1 condition suggestion was *panic disorder (7 out of 10)*. Accuracy was calculated as the percentage of agreement along with the 95% CI for binomial distributions with the Agresti-Coull method [50]. For the 3 most prevalent disorder categories in our sample (according to the interview-based diagnoses), we calculated accuracy based on contingency tables as the sum of true positives and true negatives divided by the total number of cases [51], as well as sensitivity and specificity. In addition, the Gwet first-order

agreement coefficient (AC1) [52] was calculated to assess interrater reliability. The AC1 is less prone to overcorrection for chance agreement and less sensitive to low base rates compared with other coefficients such as the Cohen κ [52,53]. Values <0.20 indicate poor strength of agreement, 0.21-0.40 indicate fair strength of agreement, 0.41-0.60 indicate moderate strength of agreement, 0.61-0.80 indicate good strength of agreement, and >0.81 indicate very good strength of agreement [54].

Scores on the SUS were calculated by subtracting 1 from the raw scores of odd-numbered items and, for the even-numbered items, by subtracting the raw score from 5 and multiplying the sum of these adjusted scores by 2.5 [55] (score range 0-100). According to Bangor et al [56], scores >70 are considered acceptable, and ≥85.5 is considered excellent. Scores for the feedback questionnaire were analyzed at the item level. Missing values in both usability questionnaires were infrequent (maximum of 2/49, 4% per variable) and were replaced with multiple imputations using a Markov chain Monte Carlo algorithm with 5 imputations per missing one. The imputed data sets were merged to obtain 1 data set. Associations between completion time of ADA and patient characteristics were explored using bivariate correlations. The AC1 was calculated using AgreeStat version 2011.3 (Advanced Analytics). All other analyses were performed using SPSS (version 27; IBM Corp) and α=.05 as a level of significance.

Results

Study Flow

Over the 1.5-year recruitment period, 159 persons were screened for inclusion, of which 104 (65.4%) did not meet the inclusion criteria or did not provide informed consent. Of the remaining 55 study participants, 6 (11%) had no interview-based diagnoses available because of early discontinuation of treatment; thus, complete data were available for 49 (89%) study participants. Table 1 shows the demographic and clinical characteristics of the participants. On average, the participants were 33.41 (SD 12.79) years old, and 61% (30/49) were women. Approximately 22% (11/49) of participants reported using the internet *often* or *always* for health information search. The mean symptom duration was 8.25 (SD 8.22) years, and 39% (19/45) of participants with available data reported past diagnoses of mental disorders.

Table 1. Demographic and clinical characteristics of the participants (N=49).

Variable	Values
Age (years), mean (SD, range)	33.41 (12.79, 18-66)
Gender, n (%)	
Female	30 (61)
Male	19 (39)
Level of education, n (%)	
Primary level	3 (6)
Intermediate level	28 (57)
Higher level	17 (35)
Other degrees	1 (2)
Family status, n (%)	
Single	33 (67)
Married or permanent partnership	15 (31)
Divorced, living apart, or widowed	1 (2)
Mother tongue, n (%)	
German	46 (94)
Language other than German	3 (6)
Duration of symptoms (years), mean (SD)	8.25 (8.22)
History of mental disorders,[a] n (%)	
Affective disorders	10 (22)
Anxiety disorders	9 (20)
Other disorders	6 (13)
No history of mental disorders	30 (67)
Past psychotherapy (yes), n (%)	25 (51)
Web-based health information seeking, n (%)	
Never	8 (16)
Rarely	30 (61)
Often	10 (20)
Always	1 (2)

[a]n=45. Multiple answers possible.

Diagnostic Agreement

On average, 2.06 (SD 0.99) diagnoses by the therapist and 3.44 (SD 1.06) condition suggestions by ADA were recorded per patient. Approximately 67% (33/49) of patients received >1 diagnosis. The most prevalent diagnostic categories in our sample (101 therapist diagnoses for 49 cases) were affective disorders (F30-F39; 34/101, 33.7%), anxiety disorders (F40-F41; 27/101, 26.7%), and somatoform and associated disorders (including F45; 9/101, 8.9%). Multimedia Appendix 1 contains a detailed list of interview-based diagnoses and ADA's condition suggestions.

In 51% (25/49; 95% CI 37.5-64.4) of cases, ADA's first condition suggestion was in accordance with any of the therapists' diagnoses, and it was in the top 5 condition suggestions in 69% (34/49; 95% CI 55.4-80.6) of cases. When considering the frequency of comorbid diagnoses, on average, ADA was able to detect <1 (mean 0.80, SD 0.64) of the mean 2.06 (SD 0.99) therapist diagnoses per patient.

Table 2 displays the performance statistics of the symptom checker's condition suggestions for the 3 most common disorder categories. The highest accuracy was observed in somatoform and associated disorders (0.76 to 0.82), and the lowest was observed in anxiety disorders (0.45 to 0.53). Sensitivity was highest for affective disorders (0.65 to 0.71) and lowest for somatoform and associated disorders (0.22 to 0.29). Interrater reliabilities (AC1) ranged from low strengths of agreement for anxiety disorders (−0.09 to 0.15) to moderate-to-good strengths of agreement for somatoform and associated disorders (0.65 to 0.76) according to proposed benchmarking thresholds [54].

Table 2. Performance statistics of *Ada–check your health* (ADA) for disorder categories.

Performance statistics	Correct condition suggestion by ADA					
	Listed first			Listed in top 5		
	Affective disorders	Anxiety disorders	Somatoform + associated disorders	Affective disorders	Anxiety disorders	Somatoform + associated disorders
Accuracy (95% CI)	0.65 (0.51 to 0.77)	0.53 (0.39 to 0.66)	0.82 (0.68 to 0.90)	0.63 (0.49 to 0.75)	0.45 (0.32 to 0.59)	0.76 (0.62 to 0.86)
Sensitivity	0.65	0.21	0.22	0.71	0.43	0.33
Specificity	0.67	0.84	0.95	0.50	0.46	0.85
AC1[a] (95% CI)	0.32 (0.46 to 0.60)	0.15 (−0.16 to 0.47)	0.76 (0.59 to 0.93)	0.31 (0.26 to 0.60)	−0.09 (−0.39 to 0.20)	0.65 (0.44 to 0.86)

[a]AC1: Gwet first-order agreement coefficient.

Separately, we examined the diagnostic accuracy of ADA for the level of severity of mild or moderate and severe depression (without cases with partially or fully remitted recurrent depression) as indicated by the therapists' diagnoses. ADA listed the correct (severity) condition suggestion first in 44% (10/23; 95% CI 25.6-63.2) of cases and in the top 5 condition suggestions in 61% (14/23; 95% CI 40.7-77.9) of cases.

Usability

None of the participants indicated having used ADA before. The average completion time of ADA was 7.90 (SD 3.39) minutes, and an average of 31.90 (SD 8.11) questions were asked. Completion time was significantly positively associated with age (r=0.40; P=.004) and illness duration (r=0.41; P=.004) but not with frequency of web-based health information seeking (r=−0.10; P=.497) or level of education (r=0.03; P=.85) and did not differ with gender (t_{47}=0.53; P=.60). On average, the participants rated the usability on the SUS as high (mean 81.51, SD 11.82), with significantly lower values in male compared with female participants (mean difference −8.61, SE 3.28; t_{47}=−2.63; P=.009). Usability was significantly negatively associated with age (r=−0.41; P=.003) but not with illness duration (P=.86), frequency of web-based health information seeking (P=.53), or level of education (P=.57).

Table 3 shows the item statistics for the feedback questionnaire [48]. Approximately 88% (43/49) of participants were satisfied with how they answered ADA's questions, 61% (30/49) found that ADA's questions were clear to them, and 71% (35/49) would have preferred a face-to-face interview.

Table 3. Item descriptions for the feedback questionnaire (adapted from Hoyer et al [48]).

Item number[a]	Item	Agreement,[b] n (%)
1	Sometimes I could not follow the app's instructions.	11 (22)
2	I enjoyed answering the questions.	34 (69)
5	Throughout the questioning, my concentration was good.	46 (94)
6	The questions were clear to me.	30 (61)
7	Now and then I wanted to quit the questioning.	1 (2)
8	The questioning was a pleasant experience for me.	37 (76)
9	During the questioning, my endurance was steady.	47 (96)
10	I'm satisfied with how I answered the questions.	43 (88)
12	I did not understand how the questions were related to my problems.	2 (4)
13	Anything related to apps makes me feel uncomfortable or anxious.	3 (6)
14	I would have preferred a normal face-to-face interview from patient to therapist.	35 (71)
15	I think it was good that the questioning was done in such an exact and detailed manner.	40 (82)

[a]Number of original items. Items 3, 4, and 11 were excluded from this study.

[b]Aggregated frequency of answers (4) *completely agree* and (3) *agree*.

Discussion

Principal Findings

To our knowledge, this comparative study is the first to independently investigate the diagnostic accuracy of a popular symptom checker (ADA) as a screening tool for mental disorders compared with validated formal diagnoses in real-world patients. Our results show that, in approximately half of all investigated cases (25/49, 51%), ADA's first listed condition suggestion was correctly aligned with any of the interview-based expert diagnoses. This transdiagnostic accuracy was higher than the

average rates of symptom checkers from previous comparative studies (26%-36%) that used case vignettes of various health conditions [34,36,57]. Furthermore, the accuracy observed in our study is close to the performance rate of ADA (48.5%) across a broad spectrum of medical conditions in the study by Gilbert et al [34] but lower than in another recent comparative study (72%) [35]. When compared with a study by Barriga et al [58], who investigated the accuracy of another symptom checker (Mediktor) in real patients in an emergency care unit, the accuracy for the first listed condition suggestions was in a comparable range (51% vs 42.9%). In two-thirds (34/49, 67%) of cases, 1 in 5 condition suggestions aligned with any of the interview-based diagnoses, which is somewhat below the range of performance rates of ADA in previous studies using case vignettes (77%-84%) [34,35] or patients seeking emergency care (91.3%) [58]. However, our findings can only be compared with the accuracy from previous studies to a limited extent. These studies included only 1 potentially correct diagnosis per case as opposed to multiple diagnoses per case in our study.

The transdiagnostic accuracy of ADA could be considered lower when compared with sensitivities of self-report screenings for mental disorders that range between 0.72 and 0.90 according to previous studies [59-62]. However, the different measures of agreement must be considered here. Interestingly, the transdiagnostic performance of ADA when used by patients is comparable with that of studies in which medical experts used ADA to enter information based on case vignettes [34]. This is in contrast to previous findings by Jungmann et al [39], who demonstrated lower performance rates of ADA in laypeople compared with health professionals with regard to correctly identifying mental disorders from case vignettes of adults and adolescents. However, our study was designed differently as we did not use standardized vignettes, and therapist diagnoses were not checked by independent raters. An interesting future study design would be to directly compare the expert and consumer-level use of symptom checkers and explore differences in diagnostic performance. However, we provide preliminary evidence that no expert knowledge or user experience may be needed to yield performance rates comparable with those of health professionals using symptom checkers. As our participants were all novices in the use of ADA, we could not test the potential beneficial effect of familiarity on diagnostic accuracy. Future studies could, for example, include a test run where participants enter information from a standardized vignette to familiarize themselves with the symptom checker.

Within the most prevalent subcategories of mental disorders in our sample, we observed considerable differences in performance statistics. For somatoform and associated disorders, accuracy, specificity, and interrater reliabilities were highest and could be considered acceptable. This may resemble the accuracy of ADA, particularly in detecting somatic medical conditions, which has been the focus of previous studies [34,35]. Beyond this, the unifying classification of functional somatic syndromes (eg, irritable bowel syndrome and fibromyalgia) as somatoform disorders is subject to ongoing controversial debate [49,63]. However, the base rate (<10%) was lowest across disorder categories, which in turn may have inflated specificity and interrater reliability. For affective and anxiety disorders, performance was lower than one would expect given that these disorder categories have a high prevalence in the general as well as clinical populations [64,65] and when compared with higher sensitivities of self-report screenings, particularly those observed for anxiety disorders [66-68]. However, with regard to the small sample size, and as the diagnostic coding scheme [39] could be considered relatively liberal for some disorders, replication in a larger sample and with more fine-grained diagnostic coding seems warranted to obtain a more robust estimation of diagnostic performance.

Furthermore, the participants rated the usability of ADA as high, which is in line with data from a previous study in primary care patients [30]. However, self-selection of study participation could have positively biased usability ratings. Concerning acceptability, almost three-fourths of our participants (35/49, 71%) preferred face-to-face diagnostics by a health professional over the symptom checker, which is comparable with preference ratings from the German general population [18]. This could be critical regarding the *reshaping* of diagnostic practice as acceptance represents a crucial premise for the implementation of health resources [69]. As symptom checkers are more likely to complement rather than substitute diagnostic processes, it would be interesting to also investigate patients' and health professionals' views on the combination of traditional and digital diagnostic procedures, for example, whether symptom checkers would be preferred as a first or second opinion in differential diagnoses or as assistance in clinical decision-making. In this regard, we did not confront the patients or therapists directly with the condition suggestions to not influence the diagnostic process. However, for clinical implementation, it would be interesting to study how symptom checkers used early in the patient journey preempt the diagnostic process and medical decisions. Further studies could also investigate the trust of users in the diagnostic and triage suggestions of symptom checkers compared with other sources of health information (eg, the internet and health professionals).

Strengths and Limitations

Concerning the interpretation of our results, several limitations must be considered. Generally, the therapists' diagnoses were based on additional information beyond the diagnostic interview (eg, anamnesis, medical records, and questionnaires) that was not available to the symptom checker, which represents a much more extensive process in terms of time and content, whereas, in using the symptom checker, the patients could decide what and how many different symptom complexes they entered. Although this ensured a user-oriented research focus, findings on diagnostic accuracy must thus be interpreted against the informational disbalance between the 2 rating sources. In this regard, it should also be noted that we compared ADA's *differential* condition suggestions for 1 symptomatology with final diagnoses by therapists (and not vice versa with their differential diagnoses). Thus, it seems reasonable to remind clinicians that expect symptom checkers to be a universal screening tool that these are designed to provide condition suggestions for 1 symptomatology at a time and, given their current intended purpose, are not suited to replace a broad diagnostic screening (eg, via validated questionnaires or

interviews). Furthermore, as digital resources may change over time, particularly when considering learning algorithms, current accuracy rates may do so as well. As previous studies have shown considerable differences between symptom checkers' diagnostic accuracy [33,35], future studies could compare various symptom checkers for the formal diagnosis of mental disorders. On this matter, evidence indicates that the use of algorithms over other methods, the inclusion of demographic information [57], or more rigorous questioning [35] could explain the differences between symptom checkers' diagnostic performances.

In addition, as this study had a pilot character and pandemic restrictions further impeded recruitment, we included a rather small sample when compared with previous studies with patients [58]. Large-scale, multicenter studies are warranted for more robust estimates of diagnostic performance, including a more fine-grained analysis of unprocessed diagnoses. The diagnostic spectrum of our participants was somewhat limited (Multimedia Appendix 1), with substance abuse disorders, eating disorders, or posttraumatic stress disorders being underrepresented. However, the most common mental disorders were frequent in our sample and resembled prevalence rates in medical settings [70]. In contrast to previous comparative studies [34], we did not include >1 diagnostic rater or assess the correctness of interview-based diagnoses. Previous studies have demonstrated a large variation in interrater reliabilities of diagnoses based on SCIDs that can range from substantial to even low agreement [71-73], which may challenge the validity of this as a *gold standard* in diagnosis [74].

Although the therapists who participated in this study were in advanced clinical training, including diagnostic training and regular supervision, and thus were experienced in performing diagnostic procedures, we did not assess the level of (diagnostic) experience or check the therapists' or symptom checker's diagnoses independently. In addition, newer versions of diagnostic systems (eg, the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, and the ICD-11) and corresponding clinical interviews should be considered as comparators in further research. Generally, one could also criticize the exclusive categorical diagnostic approach of this study, which has been challenged recently by a strictly empirical and dimensional understanding and taxonomy of psychopathology such as the Hierarchical Taxonomy of Psychopathology [75], and dimensional self-report instruments would be a logical comparator for future studies.

However, our study constitutes a robust test of the diagnostic accuracy of ADA in comparison with formal clinical diagnostics, which is pivotal for clinical implementation. We considered some major limitations of previous studies [32] given that we collected real-world patient data, which comes closer to the current intended laypeople-oriented application of symptom checkers. In contrast to standardized vignettes, which have been the default method in previous studies, our data were thus not limited to single-diagnosis cases and included consistent comorbidities. In addition, we were able to recruit a diverse sample, which covered various age groups as well as intensities of health-related internet use. Eventually, we performed an independent scientific evaluation of a commercially available product, which seems important given the plethora of health apps that have not been scientifically reviewed [14,15].

Clinical Implications

Our findings offer various clinical implications. At the public health level, symptom checkers have some potential to reduce underdiagnosis and undertreatment of mental disorders [76] and may ideally contribute to reducing chronicity and treatment delay as they represent a low-threshold, multilingual diagnostic instrument. For their possible role in formal diagnosis, the level of diagnostic and triage accuracy is the most important indicator. However, for individuals with mental health problems, the exact differentiation (eg, the severity of major depression and type of anxiety disorder) could be less important than informing on the broader diagnostic category and providing triage advice. Here, evidence shows that, although most symptom checkers seem to provide safe triage advice [33], they are somewhat more risk-averse [57] than health professionals, which could increase health care use and costs. Then again, when compared with entering symptoms into a web-based search engine, symptom checkers are likely to be a superior tool for diagnostic assistance. However, both sources can have a similar risk of adverse emotional or behavioral consequences according to a recent study by Jungmann et al [20]. For example, similar to a search engine, a symptom checker can increase health anxiety and negative affect after searching for causes of symptoms (eg, shortness of breath). In addition, symptom checkers could make the diagnostic process less intuitive and controllable, and vulnerable patient groups, less educated people, or older people are probably less likely to take advantage of this resource at the public health level, thus increasing the "digital divide" [77,78].

As argued by Semigran et al [33], if symptom checkers are regarded as a potential replacement for professional diagnostics (ie, beyond their current intended purpose), they are likely an inferior alternative. Although the average diagnostic performance of symptom checkers can be considered generally low when compared with diagnostic standards (eg, expert diagnoses and validated diagnostic instruments), some symptom checkers show more promising performance rates, including the symptom checker studied here [34,35]. Nevertheless, the progressive dissemination of smart screening instruments may contribute to shared decision-making and promote patients' understanding of and engagement in health decisions. As such, digital health resources have already become an important factor in the therapist-patient relationship [79] as more patients use digital resources for diagnostic and treatment purposes.

Although symptom checkers or even automated (eg, avatar-based) diagnostic systems [80] may reduce clinician time, they still rely on the active engagement of users. The advancement of passive mobile sensing through smartphones or wearables (eg, mobility pattern, facial expression, and speech analysis [81,82]) may allow for in situ, fine-grained digital phenotyping even without this active user input. Although this may reduce the diagnostic effort, at the same time, the perceived control over the diagnostic process could be limited. Thus, both active and passive diagnostic approaches will have to demonstrate their quality and acceptability in routine care.

Besides their potential as a waiting room screening tool, the most typical use case would be to study users in their home environment. This would also allow for a better understanding of adequate medical help-seeking, which seems to be positively associated with the triage advice of symptom checkers [83].

Finally, future research should address the effect of symptom checkers on other meaningful outcomes, such as stigmatization, attitudes toward psychotherapy, health-related self-efficacy, or the association with treatment success, which would advance the understanding of the clinical impact of these tools on mental health care.

Conclusions

Overall, our findings indicate that the diagnostic performance of a widely available symptom checker in detecting mental disorders in real patients is close to the range of performances from previous case vignette studies that covered a broad spectrum of medical conditions. From a formal diagnostic standpoint, ADA could provide clinicians with a list of condition suggestions with moderate-to-good accuracy, whereas diagnostic performances were inconsistent between disorder categories and also included low interrater reliabilities. The symptom checker was rated as user-friendly overall but was less preferred than face-to-face diagnostics. The value of symptom checkers for diagnostic screening needs to be tested on larger samples and in comparison with further diagnostic resources such as established self-report screenings.

Acknowledgments

The authors would like to thank Lea Gronemeier and Stephie Grüßner for their invaluable support in the recruitment of participants and data collection, as well as Luise Loga and Sylvan Germer for their help in organizing the data. This research was supported by internal funding (reimbursement for participants). No third-party financial support was received.

Authors' Contributions

SH, SMJ, and MW designed the study. SH conducted the study and analyzed and interpreted the data. SH wrote the draft of this manuscript. SK, MW, and SMJ provided valuable revisions. All authors contributed to further writing of the manuscript and approved the final version.

Conflicts of Interest

None declared. The authors have no relation whatsoever to Ada Health GmbH or other commercial interests.

Multimedia Appendix 1
Interview-based expert diagnoses and condition suggestions by the symptom checker app (*Ada–check your health*).
[XLSX File (Microsoft Excel File), 15 KB - mental_v9i1e32832_app1.xlsx]

References

1. Europeans becoming enthusiastic users of online health information. European Commission. 2014. URL: https://digital-strategy.ec.europa.eu/en/news/europeans-becoming-enthusiastic-users-online-health-information [accessed 2021-01-14]
2. Anthes E. Mental health: there's an app for that. Nature 2016;532(7597):20-23. [doi: 10.1038/532020a] [Medline: 27078548]
3. Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M. Why are health care interventions delivered over the internet? A systematic review of the published literature. J Med Internet Res 2006;8(2):e10 [FREE Full text] [doi: 10.2196/jmir.8.2.e10] [Medline: 16867965]
4. Berger M, Wagner TH, Baker LC. Internet use and stigmatized illness. Soc Sci Med 2005;61(8):1821-1827. [doi: 10.1016/j.socscimed.2005.03.025] [Medline: 16029778]
5. Erritty P, Wydell TN. Are lay people good at recognising the symptoms of schizophrenia? PLoS One 2013;8(1):e52913. [doi: 10.1371/journal.pone.0052913] [Medline: 23301001]
6. Patel V, Maj M, Flisher AJ, De Silva MJ, Koschorke M, Prince M, WPA Zonal and Member Society Representatives. Reducing the treatment gap for mental disorders: a WPA survey. World Psychiatry 2010;9(3):169-176 [FREE Full text] [doi: 10.1002/j.2051-5545.2010.tb00305.x] [Medline: 20975864]
7. Wang PS, Angermeyer M, Borges G, Bruffaerts R, Chiu WT, Girolamo GDE, et al. Delay and failure in treatment seeking after first onset of mental disorders in the World Health Organization's World Mental Health Survey Initiative. World Psychiatry 2007;6(3):177-185 [FREE Full text] [Medline: 18188443]
8. Chiauzzi E, DasMahapatra P, Cochin E, Bunce M, Khoury R, Dave P. Factors in patient empowerment: a survey of an online patient research network. Patient 2016;9(6):511-523. [doi: 10.1007/s40271-016-0171-2] [Medline: 27155887]
9. Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA 2002;287(20):2691-2700. [doi: 10.1001/jama.287.20.2691] [Medline: 12020305]

10. Weaver III JB, Thompson NJ, Weaver SS, Hopkins GL. Healthcare non-adherence decisions and internet health information. Comput Hum Behav 2009;25(6):1373-1380. [doi: 10.1016/j.chb.2009.05.011]
11. Grohol JM, Slimowicz J, Granda R. The quality of mental health information commonly searched for on the Internet. Cyberpsychol Behav Soc Netw 2014;17(4):216-221. [doi: 10.1089/cyber.2013.0258] [Medline: 24237287]
12. Ipser JC, Dewing S, Stein DJ. A systematic review of the quality of information on the treatment of anxiety disorders on the internet. Curr Psychiatry Rep 2007;9(4):303-309. [doi: 10.1007/s11920-007-0037-3] [Medline: 17880862]
13. North F, Ward WJ, Varkey P, Tulledge-Scheitel SM. Should you search the Internet for information about your acute symptom? Telemed J E Health 2012;18(3):213-218. [doi: 10.1089/tmj.2011.0127] [Medline: 22364307]
14. Terhorst Y, Rathner EM, Baumeister H, Sander L. «Hilfe aus dem App-Store?»: eine systematische Übersichtsarbeit und evaluation von apps zur anwendung bei depressionen. Verhaltenstherapie 2018;28(2):101-112. [doi: 10.1159/000481692]
15. Sander LB, Schorndanner J, Terhorst Y, Spanhel K, Pryss R, Baumeister H, et al. 'Help for trauma from the app stores?' A systematic review and standardised rating of apps for post-traumatic stress disorder (PTSD). Eur J Psychotraumatol 2020;11(1):1701788 [FREE Full text] [doi: 10.1080/20008198.2019.1701788] [Medline: 32002136]
16. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
17. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
18. Baumann E, Czerwinski F, Rosset M, Seelig M, Suhr R. Wie informieren sich die Menschen in Deutschland zum Thema Gesundheit? Erkenntnisse aus der ersten Welle von HINTS Germany. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020;63(9):1151-1160. [doi: 10.1007/s00103-020-03192-x] [Medline: 32666180]
19. Powell J, Clarke A. Internet information-seeking in mental health: population survey. Br J Psychiatry 2006;189:273-277 [FREE Full text] [doi: 10.1192/bjp.bp.105.017319] [Medline: 16946364]
20. Jungmann SM, Brand S, Kolb J, Witthöft M. Do Dr. Google and health apps have (comparable) side effects? An experimental study. Clin Psychol Sci 2020;8(2):306-317. [doi: 10.1177/2167702619894904]
21. Tyrer P, Cooper S, Tyrer H, Wang D, Bassett P. Increase in the prevalence of health anxiety in medical clinics: possible cyberchondria. Int J Soc Psychiatry 2019;65(7-8):566-569. [doi: 10.1177/0020764019866231] [Medline: 31379243]
22. Eastin MS, Guinsler NM. Worried and wired: effects of health anxiety on information-seeking and health care utilization behaviors. Cyberpsychol Behav 2006;9(4):494-498. [doi: 10.1089/cpb.2006.9.494] [Medline: 16901253]
23. Wangler J, Jansky M. General practitioners' challenges and strategies in dealing with Internet-related health anxieties—results of a qualitative study among primary care physicians in Germany. Wien Med Wochenschr 2020;170(13-14):329-339. [doi: 10.1007/s10354-020-00777-8] [Medline: 32767159]
24. Luxton DD. Artificial intelligence in psychological practice: current and future applications and implications. Prof Psychol Res Pr 2014;45(5):332-339. [doi: 10.1037/a0034559]
25. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019;21(4):e12887. [doi: 10.2196/12887] [Medline: 30950796]
26. Millenson ML, Baldwin JL, Zipperer L, Singh H. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. Diagnosis (Berl) 2018;5(3):95-105. [doi: 10.1515/dx-2018-0009] [Medline: 30032130]
27. Shen C, Nguyen M, Gregor A, Isaza G, Beattie A. Accuracy of a popular online symptom checker for ophthalmic diagnoses. JAMA Ophthalmol 2019;137(6):690-692. [doi: 10.1001/jamaophthalmol.2019.0571] [Medline: 30973602]
28. Munsch N, Martin A, Gruarin S, Nateqi J, Abdarahmane I, Weingartner-Ortner R, et al. Diagnostic accuracy of web-based COVID-19 symptom checkers: comparison study. J Med Internet Res 2020;22(10):e21299 [FREE Full text] [doi: 10.2196/21299] [Medline: 33001828]
29. Berry AC, Cash BD, Wang B, Mulekar MS, Van Haneghan AB, Yuquimpo K, et al. Online symptom checker diagnostic and triage accuracy for HIV and hepatitis C. Epidemiol Infect 2019;147:e104. [doi: 10.1017/s0950268819000268] [Medline: 30869052]
30. Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial intelligence-based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. JMIR Hum Factors 2020;7(3):e19713 [FREE Full text] [doi: 10.2196/19713] [Medline: 32540836]
31. Meyer AN, Giardina TD, Spitzmueller C, Shahid U, Scott TM, Singh H. Patient perspectives on the usefulness of an artificial intelligence–assisted symptom checker: cross-sectional survey study. J Med Internet Res 2020;22(1):e14679. [doi: 10.2196/14679] [Medline: 32012052]
32. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019;9(8):e027743. [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]
33. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One 2021;16(7):e0254088. [doi: 10.1371/journal.pone.0254088] [Medline: 34265845]
34. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480. [doi: 10.1136/bmj.h3480] [Medline: 26157077]

35. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open 2020;10(12):e040269. [doi: 10.1136/bmjopen-2020-040269] [Medline: 33328258]
36. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of physician and computer diagnostic accuracy. JAMA Intern Med 2016;176(12):1860-1861. [doi: 10.1001/jamainternmed.2016.6001] [Medline: 27723877]
37. Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, Sangar D, et al. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. Front Artif Intell 2020;3:543405. [doi: 10.3389/frai.2020.543405] [Medline: 33733203]
38. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018;392(10161):2263-2264. [doi: 10.1016/S0140-6736(18)32819-8] [Medline: 30413281]
39. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res 2019;3(4):e13863 [FREE Full text] [doi: 10.2196/13863] [Medline: 31663858]
40. Beck AT, Steer RA, Brown G. Manual for the beck depression inventory-II. San Antonio: Psychological Corporation; 1996.
41. Wittchen HU, Wunderlich U, Gruschwitz S, Zaudig M. SKID I: Strukturiertes Klinisches Interview für DSM-IV. Göttingen: Hogrefe; 1997:1-99.
42. Take care of yourself with Ada. Ada Health GmbH. 2021. URL: https://ada.com/app/ [accessed 2021-07-21]
43. Timiliotis J, Blümke B, Serfözö PD, Gilbert S, Ondresik M, Türk E, et al. A novel diagnostic decision support system for medical professionals: prospective feasibility study. JMIR Form Res. Preprint posted online on January 12, 2022. [FREE Full text] [doi: 10.2196/29943]
44. Hoffmann H. Ada health: our approach to assess Ada's diagnostic performance. Ada. URL: https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180925/Documents/3_Henry%20Hoffmann.pdf [accessed 2021-07-16]
45. Runny nose? - Ada your health companion #tellAda. Ada Health. 2019. URL: https://www.youtube.com/watch?v=cv75UIz8nUU [accessed 2021-11-29]
46. Brooke J. SUS - A quick and dirty usability scale. Jens Oliver Meiert. 1986. URL: https://hell.meiert.org/core/pdf/sus.pdf [accessed 2021-08-23]
47. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
48. Hoyer J, Ruhl U, Scholz D, Wittchen HU. Patients' feedback after computer-assisted diagnostic interviews for mental disorders. Psychother Res 2006;16(3):357-363. [doi: 10.1080/10503300500485540]
49. Fink P, Schröder A. One single diagnosis, bodily distress syndrome, succeeded to capture 10 diagnostic categories of functional somatic syndromes and somatoform disorders. J Psychosom Res 2010;68(5):415-426. [doi: 10.1016/j.jpsychores.2010.02.004] [Medline: 20403500]
50. Dean N, Pagano M. Evaluating confidence interval methods for binomial proportions in clustered surveys. J Surv Stat Methodol 2015;3(4):484-503. [doi: 10.1093/jssam/smv024]
51. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
52. Gwet K. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. Statistical Methods for Inter-Rater Reliability Assessment. 2002. URL: https://www.agreestat.com/papers/inter_rater_reliability_dependency.pdf [accessed 2021-11-29]
53. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013;13:61 [FREE Full text] [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
54. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991:1-624.
55. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Interact 2018;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
56. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123 [FREE Full text]
57. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Med J Aust 2020;212(11):514-519. [doi: 10.5694/mja2.50600] [Medline: 32391611]
58. Moreno Barriga E, Pueyo Ferrer I, Sánchez Sánchez M, Martín Baranera M, Masip Utset J. Experiencia de mediktor®: un nuevo evaluador de síntomas basado en inteligencia artificial para pacientes atendidos en el servicio de urgencias. Emergencias 2017;29(6):391-396 [FREE Full text] [Medline: 29188913]
59. Wittchen HU, Höfler M, Gander F, Pfister H, Storz S, Üstün B, et al. Screening for mental disorders: performance of the composite international diagnostic – screener (CID–S). Int J Method Psychiat Res 1999;8(2):59-70. [doi: 10.1002/mpr.57]
60. Schmitz N, Hartkamp N, Kiuse J, Franke GH, Reister G, Tress W. The symptom check-list-90-R (SCL-90-R): a German validation study. Qual Life Res 2000;9(2):185-193. [doi: 10.1023/a:1008931926181] [Medline: 10983482]
61. Zimmerman M, Mattia JI. A self-report scale to help make psychiatric diagnoses: the psychiatric diagnostic screening questionnaire. Arch Gen Psychiatry 2001;58(8):787-794. [doi: 10.1001/archpsyc.58.8.787] [Medline: 11483146]

62. Donker T, van Straten A, Marks I, Cuijpers P. A brief Web-based screening questionnaire for common mental disorders: development and validation. J Med Internet Res 2009;11(3):e19 [FREE Full text] [doi: 10.2196/jmir.1134] [Medline: 19632977]
63. Wessely S, Nimnuan C, Sharpe M. Functional somatic syndromes: one or many? Lancet 1999;354(9182):936-939. [doi: 10.1016/s0140-6736(98)08320-2] [Medline: 10489969]
64. Olesen J, Gustavsson A, Svensson M, Wittchen HU, Jönsson B, CDBE2010 study group, European Brain Council. The economic cost of brain disorders in Europe. Eur J Neurol 2012;19(1):155-162. [doi: 10.1111/j.1468-1331.2011.03590.x] [Medline: 22175760]
65. Wang J, Wu X, Lai W, Long E, Zhang X, Li W, et al. Prevalence of depression and depressive symptoms among outpatients: a systematic review and meta-analysis. BMJ Open 2017;7(8):e017173 [FREE Full text] [doi: 10.1136/bmjopen-2017-017173] [Medline: 28838903]
66. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. Gen Hosp Psychiatry 2016;39:24-31. [doi: 10.1016/j.genhosppsych.2015.11.005] [Medline: 26719105]
67. Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): a systematic review with meta-analysis. PLoS One 2016;11(5):e0155431 [FREE Full text] [doi: 10.1371/journal.pone.0155431] [Medline: 27182821]
68. von Glischinski M, von Brachel R, Hirschfeld G. How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). Qual Life Res 2019;28(5):1111-1118. [doi: 10.1007/s11136-018-2050-x] [Medline: 30456716]
69. Philippi P, Baumeister H, Apolinário-Hagen J, Ebert DD, Hennemann S, Kott L, et al. Acceptance towards digital health interventions – model validation and further development of the unified theory of acceptance and use of technology. Internet Interv 2021;26:100459. [doi: 10.1016/j.invent.2021.100459] [Medline: 34603973]
70. Ansseau M, Dierick M, Buntinkx F, Cnockaert P, De Smedt J, Van Den Haute M, et al. High prevalence of mental disorders in primary care. J Affect Disord 2004;78(1):49-55. [doi: 10.1016/s0165-0327(02)00219-7] [Medline: 14672796]
71. Lobbestael J, Leurgans M, Arntz A. Inter-rater reliability of the structured clinical interview for DSM-IV axis I disorders (SCID I) and axis II disorders (SCID II). Clin Psychol Psychother 2011;18(1):75-79. [doi: 10.1002/cpp.693] [Medline: 20309842]
72. Cheniaux E, Landeira-Fernandez J, Versiani M. The diagnoses of schizophrenia, schizoaffective disorder, bipolar disorder and unipolar depression: interrater reliability and congruence between DSM-IV and ICD-10. Psychopathology 2009;42(5):293-298. [doi: 10.1159/000228838] [Medline: 19609099]
73. Andreas S, Theisen P, Mestel R, Koch U, Schulz H. Validity of routine clinical DSM-IV diagnoses (Axis I/II) in inpatients with mental disorders. Psychiatry Res 2009;170(2-3):252-255. [doi: 10.1016/j.psychres.2008.09.009] [Medline: 19896721]
74. Lilienfeld SO, Sauvigné KC, Lynn SJ, Cautin RL, Latzman RD, Waldman ID. Fifty psychological and psychiatric terms to avoid: a list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. Front Psychol 2015;6:1100 [FREE Full text] [doi: 10.3389/fpsyg.2015.01100] [Medline: 26284019]
75. Kotov R, Krueger RF, Watson D. A paradigm shift in psychiatric classification: the hierarchical taxonomy of psychopathology (HiTOP). World Psychiatry 2018;17(1):24-25 [FREE Full text] [doi: 10.1002/wps.20478] [Medline: 29352543]
76. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 2017 Dec;210(2):119-124 [FREE Full text] [doi: 10.1192/bjp.bp.116.188078] [Medline: 27908899]
77. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult internet users. J Med Internet Res 2016;18(7):e192 [FREE Full text] [doi: 10.2196/jmir.5413] [Medline: 27432783]
78. Cornejo Müller A, Wachtler B, Lampert T. Digital Divide – Soziale Unterschiede in der Nutzung digitaler Gesundheitsangebote. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020;63(2):185-191. [doi: 10.1007/s00103-019-03081-y] [Medline: 31915863]
79. Tan SS, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017;19(1):e9 [FREE Full text] [doi: 10.2196/jmir.5729] [Medline: 28104579]
80. Rizzo AA, Lucas G, Gratch J, Stratou G, Morency LP, Shilling R, et al. Clinical interviewing by a virtual human agent with automatic behavior analysis. In: Proceedings of the 11th international conference on disability, virtual reality and associated technologies. 2016 Presented at: ICDVRAT'16; September 22-26, 2016; Los Angeles p. 57-64.
81. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
82. Garcia-Ceja E, Riegler M, Nordgreen T, Jakobsen P, Oedegaard KJ, Tørresen J. Mental health monitoring with multimodal sensing and machine learning: a survey. Pervasive Mob Comput 2018;51:1-26. [doi: 10.1016/j.pmcj.2018.09.003]
83. Winn AN, Somai M, Fergestrom N, Crotty BH. Association of use of online symptom checkers with patients' plans for seeking care. JAMA Netw Open 2019;2(12):e1918561. [doi: 10.1001/jamanetworkopen.2019.18561] [Medline: 31880791]

Abbreviations

AC1: Gwet first-order agreement coefficient
ADA: *Ada–check your health*
ICD: International Classification of Diseases
SCID: Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
SUS: System Usability Scale

Edited by J Torous; submitted 24.08.21; peer-reviewed by S Gilbert, Y Terhorst, N Munsch, A Palanica; comments to author 01.10.21; accepted 09.11.21; published 31.01.22.

Please cite as:
Hennemann S, Kuhn S, Witthöft M, Jungmann SM
Diagnostic Performance of an App-Based Symptom Checker in Mental Disorders: Comparative Study in Psychotherapy Outpatients
JMIR Ment Health 2022;9(1):e32832
URL: https://mental.jmir.org/2022/1/e32832
doi:10.2196/32832
PMID:35099395

Original Paper

Effectiveness, User Engagement and Experience, and Safety of a Mobile App (Lumi Nova) Delivering Exposure-Based Cognitive Behavioral Therapy Strategies to Manage Anxiety in Children via Immersive Gaming Technology: Preliminary Evaluation Study

Joanna Lockwood[1], PhD; Laura Williams[1], MSc; Jennifer L Martin[1], PhD; Manjul Rathee[2], MA; Claire Hill[3], DClinPsy, PhD

[1]National Institute of Health Research MindTech MedTech Co-operative, School of Medicine, University of Nottingham, Nottingham, United Kingdom
[2]BFB Labs Ltd, London, United Kingdom
[3]School of Psychology & Clinical Language Sciences, University of Reading, Reading, United Kingdom

Corresponding Author:
Joanna Lockwood, PhD
National Institute of Health Research MindTech MedTech Co-operative
School of Medicine
University of Nottingham
Institute of Mental Health, Jubilee Campus
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 115 8231294
Email: joanna.lockwood@nottingham.ac.uk

Abstract

Background: Childhood anxiety disorders are a prevalent mental health problem that can be treated effectively with cognitive behavioral therapy, in which exposure is a key component; however, access to treatment is poor. Mobile-based apps on smartphones or tablets may facilitate the delivery of evidence-based therapy for child anxiety, thereby overcoming the access and engagement barriers of traditional treatment. Apps that deliver therapeutic content via immersive gaming technology could offer an effective, highly engaging, and flexible treatment proposition.

Objective: In this paper, we aim to describe a preliminary multi-method evaluation of Lumi Nova, a mobile app intervention targeting mild to moderate anxiety problems in children aged 7-12 years using exposure therapy delivered via an immersive game. The primary objective is to evaluate the effectiveness, user engagement and experience, and safety of the beta version of Lumi Nova.

Methods: Lumi Nova was co-designed with children, parents, teachers, clinicians, game industry experts, and academic partnerships. In total, 120 community-based children with mild to moderate anxiety and their guardians were enrolled to participate in an 8-week pilot study. The outcome measures captured the app's effectiveness (anxiety symptoms, child-identified goal-based outcomes, and functional impairment), user engagement (game play data and ease-of-use ratings), and safety (mood ratings and adverse events). The outcome measures before and after the intervention were available for 30 children (age: mean 9.8, SD 1.7 years; girls: 18/30, 60%; White: 24/30, 80%). Additional game play data were automatically generated for 67 children (age: mean 9.6, SD 1.53 years; girls: 35/67, 52%; White: 42/67, 63%). Postintervention open-response data from 53% (16/30) of guardians relating to the primary objectives were also examined.

Results: Playing Lumi Nova was effective in reducing anxiety symptom severity over the 8-week period of game play (t_{29}=2.79; P=.009; Cohen d=0.35) and making progress toward treatment goals (z=2.43; P=.02), but there were no improvements in relation to functional impairment. Children found it easy to play the game and engaged safely with therapeutic content. However, the positive effects were small, and there were limitations to the game play data.

Conclusions: This preliminary study provides initial evidence that an immersive mobile game app may safely benefit children experiencing mild to moderate anxiety. It also demonstrates the value of the rigorous evaluation of digital interventions during the development process to rapidly improve readiness for full market launch.

(JMIR Ment Health 2022;9(1):e29008) doi:10.2196/29008

KEYWORDS

anxiety; children; exposure therapy; cognitive behavioral therapy; immersive gaming; digital intervention; app; smartphone; mobile phone

Introduction

Background

Anxiety disorders are among the most common and impairing mental health difficulties experienced in childhood and are characterized by excessive fear, worry, and negative beliefs that can result in distress and functional impairments in social, academic, and family life [1,2]. Anxiety disorders typically begin in childhood [3], often co-occurring with other anxiety disorders and depressive, behavioral, and neurodevelopmental disorders [4]. When not treated successfully, children who experience high levels of anxiety can continue to have problems over their life course and are at increased risk for other persistent, long-term adverse outcomes [2,4-6]. Recent national survey data from the United Kingdom indicate that emotional difficulties (anxiety and low mood) have increased by nearly 50% in young people over the 2004-2017 period [1]. Into what was already a concerning situation, the COVID-19 pandemic has contributed significant disruption and uncertainty to young lives, and early findings point to heightening anxiety in primary-age children and those with pre-existing vulnerabilities during the first stages of lockdown [7]. Early identification and access to effective treatment is critical.

Evidence-Based Treatment and Associated Challenges

Substantial clinical evidence suggests that anxiety in children can be effectively treated using psychological approaches [8]. Cognitive behavioral therapy (CBT) demonstrates consistent superiority in randomized controlled trials over no therapy for mild to moderate childhood anxiety, and consequently is the first-line recommended treatment for children and young people [9,10]. CBT for anxiety involves psychoeducation, identifying and challenging anxious thoughts, facing feared objects and situations through graded exposure, and problem-solving techniques. Treatment format, delivery in shortened form, or comorbidity does not appear to substantially alter CBT efficacy [9,11]. However, around a third of children and adolescents retain their primary anxiety disorder following a course of CBT treatment, suggesting that alternative or more targeted approaches are warranted [11]. Most children who could benefit from an intervention do not access formal support. Around 60% of children with anxiety disorders do not seek professional help, with only a small minority receiving support from specialist mental health services (15.2%), and less than 3% receiving CBT [12,13]. Barriers include overstretched mental health services with lengthy waitlists, as well as attitudinal issues around stigma, negative beliefs or lack of awareness about mental health services, and preferences for self-help over clinical support [14-16]. Poor adherence to treatment and high dropout rates (23%-60%) are also a threat to treatment benefits and suggest that the interventions may be lacking appeal for young people [17,18].

Importantly, although expert consensus and dismantling studies indicate that exposure-based elements of CBT are active components that are effective in treating anxiety disorders, exposure-based CBT is infrequently included in interventions for children [19,20]. This underutilization may relate to high costs and time constraints, as well as a lack of therapist training and confidence or negative beliefs about the approach [21,22]. In particular, anxious children may lack intrinsic motivation to comply with exposure elements of therapy, given that they are unlikely to have initiated help-seeking in the first place and they might be naturally hesitant to face anxiety-provoking situations [22]. Novel treatment platforms for therapy delivery that (1) appeal to children, (2) are accessible to children, and (3) optimize exposure-based treatment in ways that are acceptable to children may help address some of the barriers to successful treatment.

Maximizing Access and Benefit Through Mobile Apps

Digital mental health interventions (including web-based or computer-based programs) that draw on CBT-based techniques are effective in reducing anxiety symptom severity in children and young people [23-26]. However, the evidence base is limited, particularly for younger children [25], the uptake and adherence to treatment for young people is often low or variable, dropout rates can be high, and there are little systematic data on levels of engagement. Outside of controlled clinical trials, real-world uptake and adherence with web-based digital interventions for mood disorders are similarly variable [27], which makes it difficult to establish the translation of impact to natural settings. The evaluation of digital interventions is largely limited to web-based or computer-based programs that were developed several years ago and, crucially, did not draw on a co-design approach. Recent guidance has called for increased participatory approaches that actively engage stakeholders throughout the development cycle of digital interventions to ensure that innovations fit needs, are acceptable, and are used [28]. Transparent reporting of the contribution of co-design and user-centered processes is necessary to benchmark their role in the development of new innovations [29-31].

Outcomes may be optimized for children when the capabilities of mobile technologies (smartphones and tablets) are fully leveraged. Many children are comfortable and familiar with processing information and engaging with content via mobile devices. The levels of digital independence for children are increasing, with around 50% of those aged 8-11 years using a smartphone and 72% using a tablet [32]. Interventions delivered remotely via mobile devices (mobile health [mHealth]) may bring the advantage of increased appeal and access for young people, potentially extending to those less likely to access support in traditional mental health settings [33]. However, although mHealth interventions may hold promise [34], few child-focused interventions have been subject to empirical evaluation [35-37] or are supported only in relation to feasibility,

but not efficacy, usability, or safety [36,38,39]. Given the increasing ubiquity of apps for childhood mental health, robust evaluation studies are a research priority. Recent guidance has called for granular evaluation of use and engagement indicators in mHealth apps, including multidimensional objective and subjective engagement measures, and to understand how apps impact treatment outcomes [40,41].

Immersive Games for Anxiety

The application of game design elements is heralded as a strategy to increase engagement and adherence with mHealth interventions, offering an intrinsically motivating option for therapeutic delivery for children, particularly where content supports user preferences for being interactive, personable, and relatable [30,42]. Although empirical evidence to support game-based mental health interventions for childhood anxiety is lacking [29,33], increased user engagement and improved outcomes have been attributed to the integration of gamification techniques and interactive features in smartphone-delivered CBT [43]. The gamification elements that scaffold learning may help to make complex models of therapy (such as CBT) more understandable for children [44]. The structured stepped approach in exposure therapy is also suited to a game format in which progression and reward systems lend themselves to graduated challenges and motivation. Digital innovation offers the potential to deliver exposure-based therapy through immersive technologies (eg, providing the user with an experience of being able to view and interact with simulated objects and environments such as 360-degree photography and virtual and augmented reality). This innovation may help overcome some of the practical and cost barriers to delivering exposure therapy in real-world settings. Limited data support the viability of using computer games and video-based platforms to support the delivery of CBT-based therapeutic processes, including exposure tasks for childhood mental health problems [44-46], and studies have shown that exposure-based game mechanics provide an effective therapeutic action mechanism [47]. However, robust outcome evidence is sparse, and high-end immersive game-based apps that deliver structured exposure-based treatment at a self-help level for children remain underexplored.

Study Objectives

This preliminary study evaluates the effectiveness, user engagement and experience, and safety of a novel app for smartphones and tablets (Lumi Nova), which uses immersive gaming technology to deliver exposure therapy for children aged 7-12 years with mild to moderate anxiety difficulties. Specifically, the primary objective is to evaluate the following: (1) whether exposure therapy delivered via Lumi Nova is associated with a reduction in guardian-reported anxiety symptoms and functional impairment in children and progression toward the child-identified goals related to anxiety; (2) user engagement, ease of use, and experience of Lumi Nova; and (3) whether Lumi Nova is safe to use (ie, is not associated with harm or unintended negative consequences). Our expectation is that playing Lumi Nova would be associated with lower anxiety symptom severity and interference after the intervention and positive progression toward treatment goals. No further hypotheses have been offered for this exploratory study.

Methods

Study Design

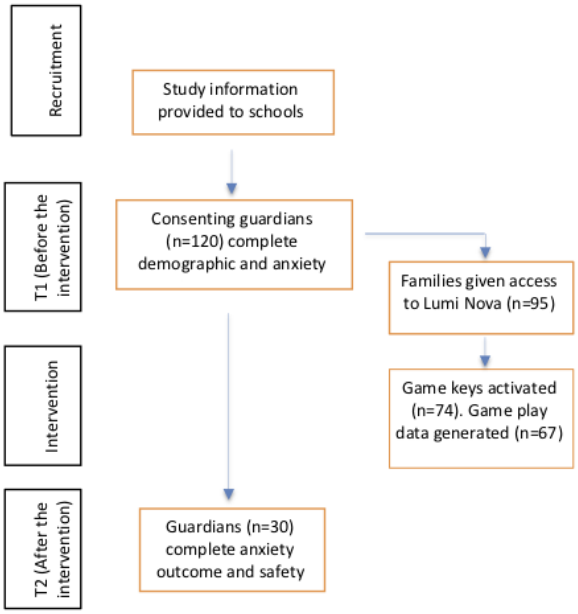
Multiple quantitative and qualitative methods were used. A pre-post design was used to compare the guardian-rated outcome measures captured via survey before (T1) and after the intervention (T2). In addition, game play data were collected over the course of the intervention, and player ratings and guardian open survey responses were collected after the intervention (T2). Data were collected during a 10-week intervention period between January and March 2020 with game play data generated over approximately 8 weeks of play. The study was approved by the Faculty of Medicine and Health Sciences Research Ethics Committee, University of Nottingham (Reference: 452-1911; December 19, 2019).

Participants

A total of 120 English-speaking children aged 7-12 years and their guardians completed T1 anxiety measures. Children were identified by school-based staff in 12 participating schools as experiencing difficulties with anxiety and not concurrently receiving psychological treatment. The participating schools were 9 primary schools and 3 secondary schools in the South East England identified through a partnership with the local council Personal, Social, Health and Economic education curriculum and Healthy School Lead and supported by a Children and Young People's Mental Health and Wellbeing Steering Group. The mean eligibility for free school meals across these schools (a proxy for socioeconomic status) was 18.1% (SD 7.3%), indicating that the school sample from which children were drawn was broadly representative of nationally reported proportions (15.8%) across all primary school types (Department for Education, January 2019). Most children (88/120, 73.3%) had not sought or received previous treatment for anxiety before starting the pilot intervention through the Children and Adolescent Mental Health Service (CAMHS), a general practitioner or nurse (92/120, 76.7%), or a psychologist or counselor (94/120, 78.3%).

Of the 120 participants with complete anxiety-related outcome measures at T1, follow-up measures at T2 were available for 30 (25%) children aged 6-13 years (mean 9.8, SD 1.7 years); 2 (1.7%) children were marginally outside the target age range of 7-12 years (aged 6.97 and 13.0 years) at the point of entering the study and were retained in the analysis. Of the 120 guardians from the T1 sample, 95 (79.2%) completed an additional anxiety measure survey following an automated SMS text message prompt to be provided with a game key, and 74 (61.7%) guardians activated the game key and downloaded Lumi Nova. Subsequent game play data were recorded for 67 (71%) out of 95 participants. Among the 30 participants with complete T1-T2 anxiety-related outcome measures, game play data were recorded for 25 (83%). Details of the study recruitment and attrition are shown in Figure 1.

Figure 1. Overview of the recruitment and study process.



Analyses were conducted on the two subsamples for whom there was complete data: the T1-T2 complete outcome measure subsample (n=30) and the game play analytics subsample (n=67). The demographic characteristics and outcome variables of these samples are presented in Table 1. Children for whom there were complete outcome measures at T1 and T2 did not differ statistically on demographic variables or outcome measures (based on 2-tailed independent samples *t* tests and chi-square tests) before the intervention in comparison to the 90 children lost to follow-up at T2: gender (*P*=.32), ethnicity (*P*=.12), disability (*P*=.76), free school meal status (*P*=.22), predominant language (*P*=.32), other anxiety treatment (*P*=.06), Revised Child Anxiety and Depression Scale–Parent version (RCADS-P; *P*=.052), Child Anxiety Impact Scale–Parent version (CAIS-P; *P*=.33); however, they were significantly more anxious (*P*=.04; Spence Child Anxiety Scale–Parent version [SCAS-P-8]). Children who played Lumi Nova for whom we had complete outcome measures (25/67, 37%) did not statistically differ from those who played the game but did not provide outcome measure data (42/67, 63%) on demographic variables or outcome measures before the intervention. Regarding clinical characteristics, before the intervention, 40% (12/30) of the T1-T2 subsample and 23% (15/67) of the game play subsample scored within a clinical range for anxiety disorders.

Table 1. Demographic data and clinical characteristics for study subsamples.

Demographic details	T1-T2 subsample (n=30)	Game play subsample (n=67)
Age[a] (years), mean (SD)	9.81 (1.70)	9.6 (1.53)
Gender, n (%)		
Male	12 (40)	31 (46)
Female	18 (60)	35 (52)
Free school meals, n (%)		
Yes	10 (33)	17 (25)
Disability, n (%)		
No	29 (97)	64 (96)
Ethnicity, n (%)		
Asian or Asian British	1 (3)	2 (3)
Black or African or Caribbean or Black British	3 (10)	8 (12)
Mixed or multiple ethnicities	1 (3)	4 (6)
Other ethnic groups	1 (3)	1 (1)
White	24 (80)	42 (63)
Predominant language, n (%)		
English	30 (100)	57 (85)
Treatment history[b], n (%)		
Other anxiety treatment		
No	25 (83)	53 (79)
Yes	2 (7)	3 (4)
Do not know	1 (3)	1 (1)
CAMHS[c] contact for anxiety		
No	23 (77)	46 (69)
Yes	6 (20)	10 (15)
Do not know	1 (3)	1 (1)
GP[d] or nurse contact for anxiety		
No	25 (83)	48 (72)
Yes	5 (16)	8 (12)
Do not know	0 (0)	1 (1)
Clinical characteristics (n=59), mean (SD)		
SCAS-P-8[e]	8.33 (4.56)	7.83 (3.71)
RCADS-P[f,g] (total anxiety)	30.30 (16.92)	28.97 (14.45)
CAIS-P[h] (total)	20.57 (15.40)	18.39 (13.25)
Clinical thresholds[i,j,k] (n=56), n (%)		
At clinical cutoff	8 (40)	13 (23)
At borderline cutoff	1 (5)	4 (7)
Within normal range	11 (55)	39 (70)

[a]Game play subsample age was based on 59 responses.

[b]Treatment history was based on the previous 3 months.

[c]CAMHS: Children and Adolescent Mental Health Service.

[d]GP: general practitioner.

[e]SCAS-P-8: Spence Child Anxiety Scale–Parent version.

[f]RCADS-P: Revised Child Anxiety and Depression Scale–Parent version.

[g]Clinical characteristics were based on 58 responses for the Revised Child Anxiety and Depression Scale–Parent version.

[h]CAIS-P: Child Anxiety Impact Scale–Parent version.

[i]Clinical thresholds describe the top 2% of scores of unreferred children of the same age and the top 7% for borderline clinical threshold.

[j]Clinical cutoffs were based on 56 participants who met the age range for standardized Revised Child Anxiety and Depression Scale–Parent version *t* scores (*t* scores are calculated from raw scores to enable comparison of anxiety scores to population-level data).

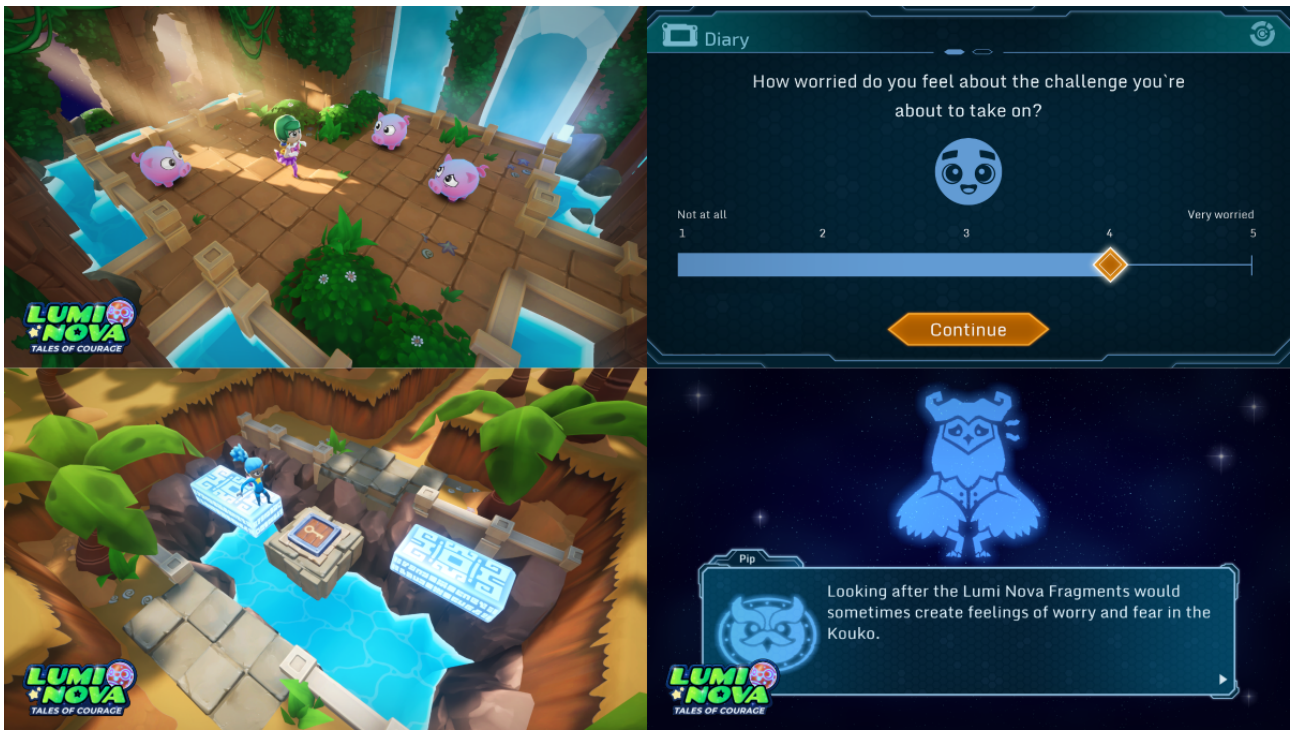
[k]For the T1-T2 subsample, the clinical cutoffs were based on 20 participants who met the age range for standardized Revised Child Anxiety and Depression Scale–Parent version *t* scores.

Intervention Development and Therapeutic Approach

Lumi Nova combines evidence-based therapeutic content (exposure therapy) and psychoeducational content within an immersive game designed to provide timely support to children aged 7-12 years, who are facing difficulties with anxiety. The app uses a diverse range of techniques, including storytelling, photographs, videos, 360° videos, and game mechanics with a progressive narrative, rewards, customization of avatars, and unlocking new levels to deliver an immersive experience to users. The development and design of Lumi Nova resulted from a robust coproduced and collaborative user-centered design process that involved children, parents, teachers, clinical practitioners, academics, and game industry experts to build the game concept, design, and clinical model parameters. In the initial phase of development, the aim was to develop a prototype game that delivered exposure therapy in a way that would be engaging, effective, and viable for children. The development phase involved multiple and multi-school site cocreation and user-testing sessions, and early prototype testing sessions with key stakeholders over a period of 5 months.

The game narrative is an intergalactic role-playing adventure in which players assume the role of a treasure hunter on a quest to save the galaxy and explore the universe, helping characters on various planets while training to overcome real-world fears (Figure 2). The game is played independently and is downloadable to a mobile or tablet (Android [Google Inc], and iOS [Apple Inc]) and does not require additional hardware or software. Guardians are the parents or carers, or other adults with parental responsibility. Guardian involvement is encouraged through automated SMS text message technology triggered by the child's progress in the game and is necessary for goal-setting and the supervision of out-of-game challenges.

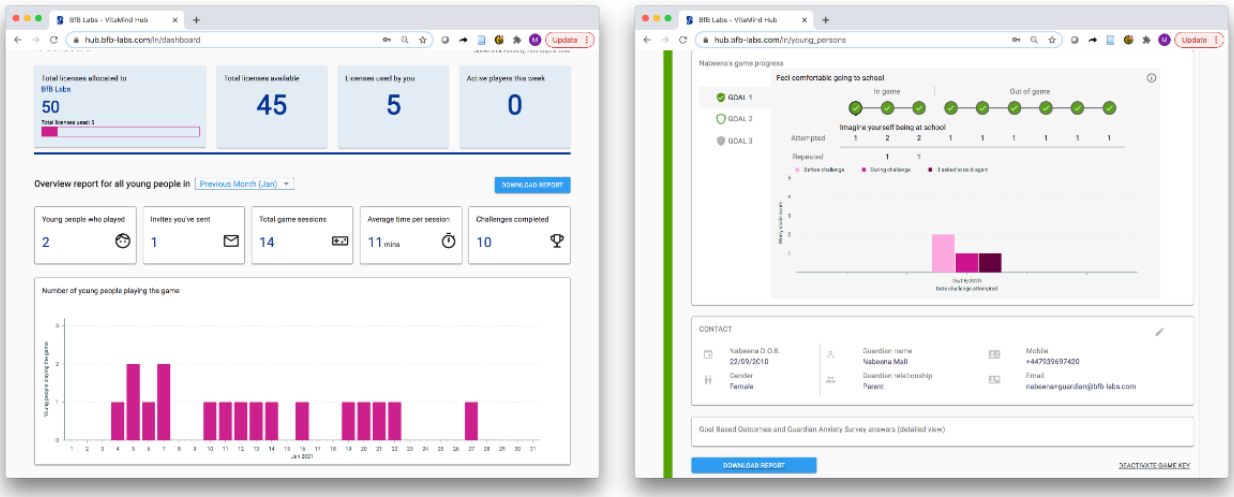
Figure 2. Example screenshots from Lumi Nova game play.



The mechanics of the intervention facilitates players to set anxiety-related goals and build a graded ladder of exposure steps (challenges) and to undertake these steps recording their before, after, and future exposure reflection in response to clinical psychologist determined prompts (eg, "What do you think might happen during this challenge?" "How worried did you feel during the challenge?" "How worried would you feel if you have to do it again?"). This approach is underpinned by strategies for optimizing learning during exposure, based on inhibitory learning perspectives. Negative expectancies associated with a perceived aversive outcome are countered by emphasizing the mismatch between what is expected to occur and what actually occurs [48]. In total, 14 common anxiety-related goals are available for selection during game play, which are related to social anxiety, separation anxiety, and specific phobias. Anxiety goals and exposure steps were

determined in consultation with clinicians, parents, and children during coproduction workshops. These include exposure steps completed within the game, and within a real-world setting (in vivo) with guardian support, to combine multiple opportunities and varied contexts for exposure practice in line with recommended practice [48]. Players must complete each exposure step to progress through the game and achieve their goals. The game also provides embedded psychoeducational information about anxiety and exposure therapy. There was no suggested amount of time for game play per session. However, players can only play for up to 40 minutes per day after the first session, which included a tutorial. This time limit was judged as providing sufficient time to engage beneficially with Lumi Nova but was short enough to address parental concerns around too much screen time [49]. Access to Lumi Nova is provided through a secure web-based platform, VitaMind Hub (BfB Labs Ltd), which is a point of access for professionals and tracks player progress with the game (Figure 3). Progress data included the goals the child was working on, child worry scales before and after each challenge step, child-reported progress toward reaching their goal (ie, goal-based outcomes [GBOs]), and scores on a brief guardian-reported anxiety measure (ie, SCAS-P-8) before the intervention and after the completion of each goal (see the *Measures* section). Progress data were accessible to the authorized health and social care or education professional providing the child with access to Lumi Nova, thereby helping to better inform care and support. Guardians had access to the Lumi Nova webpage, which carried additional psychoeducational information about anxiety.

Figure 3. Example screenshots of VitaMind Hub. Progress data are accessible to authorized professionals to facilitate active remote monitoring and care decisions.



Measures

Demographic Information

Guardian-completed survey items captured demographic information (age, gender, ethnicity, primary language spoken at home, and eligibility for free school meals) and clinical history (such as previous treatment for anxiety, contact with CAMHS, or a general practitioner or nurse because of anxiety in the previous 3 months) for their child.

Anxiety Outcomes

Brief SCAS-P-8

The parent-rated brief SCAS-P-8 [12] was used to assess child anxiety symptoms at T1 and T2. The SCAS-P-8 contains 8 items from the original 38-item SCAS [50], which assesses Diagnostic and Statistical Manual of Mental Disorders-fifth edition–related anxiety disorders (generalized anxiety, separation anxiety, social anxiety, panic, and agoraphobia) and is appropriate for use with children aged 7-12 years. Items are scored on a 4-point scale (never, sometimes, often, and always) and summed to derive a total score. Robust psychometric properties have been shown for the SCAS-P-8 [12], and the internal consistency was good (Cronbach α=.88) in this sample.

RCADS-P Questionnaire

The RCADS-P [51] was used to assess child anxiety and low mood at T1 and T2. The RCADS-P is a 47-item parent report scale comprising 5 subscales that assess symptoms of anxiety diagnoses (separation anxiety disorder, social anxiety disorder, generalized anxiety disorder, panic disorder, and obsessive-compulsive disorder) and one subscale that assesses symptoms of low mood (major depressive disorder). Items are scored on a 4-point scale ranging from 0 to 3 (never, sometimes, often, and always). A total anxiety score (summed anxiety subscale scores; 37 items) and a subscale raw score for major depressive disorder were generated. The RCADS has robust psychometric properties in children and young people [51], and internal consistency for the subscales was good (Cronbach α: range .84-.90) in this sample.

CAIS-P Questionnaire

The CAIS-P [52,53] was used to measure the functional impairment of anxiety in children at T1 and T2. The CAIS-P is a 27-item questionnaire that assesses the extent to which anxiety impacts the functioning of children within school, social, and home and family contexts. Two items that were not relevant to preadolescent children (going on a date and having a boyfriend or girlfriend) were excluded. The items are scored on a 4-point

scale (score 0 to 3; not at all, just a little, pretty much, and very much) and summed to produce 4 subscales (school, social, home and family, and global) and a total impairment score. The total impairment score was used in this study. The CAIS-P has demonstrated good psychometric properties [52,53], and internal consistency was good (Cronbach α: range .82-.88) in this sample.

GBO Tool

The GBO tool [54] was used to measure child-rated progress toward an individual therapeutic goal. Children, supported by a guardian, were asked to select up to 3 goals from 14 common anxiety-related goals prepopulated in Lumi Nova. The available goals were identified by academic clinical partners in relation to common childhood anxieties (eg, for separation anxiety related to being away from a parent or caregiver, "Be able to sleep on their own" was classified as an appropriate goal and for social anxiety, "Be comfortable speaking in front of a group" was deemed an appropriate goal). Children then undertook up to 10 exposure *challenge* steps (in-game and out-of-game challenges with guardian support) to gradually work toward 1 selected goal. Progress toward goal achievement was tracked on a 10-point Likert scale with end points ranging from 0 (no progress toward goal) to 10 (goal reached). GBO scores were collected at T1 and then weekly until a final T2 score was obtained. GBOs are routinely used for outcome monitoring within CAMHS settings and can provide a useful subjective assessment of intervention impact (goal achievement) to support standardized symptom assessment tools.

User Engagement and Ease of Use

Anonymized game play data, automatically generated during game play and uploaded to the hub when connected to Wi-Fi, captured game play information, for example, the frequency (total number) of play sessions per player, and duration of play (number of days playing). One question ("How easy is Lumi Nova to play?") was adapted from the Program Content and Usability questionnaire [55] and assessed child-rated ease of use after the intervention. Scores were rated on a Likert scale ranging from 1 (very easy) to 5 (very hard).

Safety

The safety of Lumi Nova was assessed using three indices: (1) change in the major depressive disorder subscale of the RCADS-P (see *Anxiety Outcomes* section) across the intervention; (2) guardian-reported change (positive or negative) in their child at T2, which they attributed to playing Lumi Nova; and (3) guardian-reported adverse events over the duration of the intervention.

Open-Response Questions (Optional)

Optional open-response questions for guardians (within the guardian-rated survey at T2) solicited thoughts about the following: (1) guardian perceived changes (positive or negative) associated with playing Lumi Nova, (2) general comment regarding accessing or playing the game, and (3) additional comments. Responses pertinent to the study objectives, that is, those describing (1) effectiveness, (2) user engagement and experience, and (3) safety were summarized.

Procedure

All guardians provided informed consent, and the children provided verbal assent before participating in the study. The guardians were asked to complete the demographic and anxiety outcome questionnaires (SCAS-P-8, RCADS-P, and CAIS-P) at T1 using a web-based survey platform. Subsequently, authorized school staff with access to the VitaMind Hub set up child profiles, which automatically triggered an SMS text message to their guardians with access to Lumi Nova via a game key. Participating families were asked to encourage their children to play Lumi Nova multiple times a week over the course of 8 weeks. At the end of the intervention (T2), guardians were asked to complete the anxiety outcome questionnaires (SCAS-P-8, RCADS-P, and CAIS-P).

Analytic Strategy

Analyses were performed using SPSS (version 26; IBM Corp). Descriptive statistics were used to summarize sample data; 2-tailed paired sample *t* tests were computed to demonstrate changes in outcome measures before and after the intervention; Wilcoxon signed-rank tests evaluated median difference in goal progression; frequency distributions were computed for ease-of-use scores; and descriptive statistics were used to summarize the game play (duration and frequency of game play sessions), adoption, and completion of exposure challenges in game and in vivo. Simple content analysis summarized and systematized open-response data in accordance with the following study domains identified a priori: effectiveness, user engagement and experience, and safety [56].

Results

Anxiety Symptoms and Interference

Mean scores relating to symptom severity and interference before (T1) and after intervention (T2) are reported in Table 2. There was a small reduction in mean scores for symptom severity from T1-T2 for RCADS-P total anxiety and SCAS-P-8. This reduction was statistically significant for SCAS-P-8 (P=.009) with a small to moderate effect size and survived correction for multiple analyses. However, no significant difference was reported in RCADS-P total anxiety or in anxiety impairment (CAIS-P).

Table 2. Mean change in primary outcome measures for the T1-T2 sample.

Measure	T1, mean (SD)	T2, mean (SD)	*P* value
Anxiety symptoms			
SCAS-P-8[a,b] (total)	8.33 (4.56)	7.43 (3.28)	.009
RCADS-P[c] (total anxiety)	30.73 (13.94)	30.30 (16.92)	.20
Functional impairment			
CAIS-P[d,e] (total)	20.57 (15.40)	20.97 (15.49)	.80
Safety			
RCADS-P (MDD[f])	7.07 (4.91)	6.60 (3.94)	.46

[a]SCAS-P-8: Spence Child Anxiety Scale–Parent version.

[b]Only 1 variable (Spence Child Anxiety Scale–Parent version total) was associated with a statistically significant finding (t_{29}=2.79; *P*=.009; Cohen *d*=0.35), which remained after Bonferroni correction at *P*<.01.

[c]RCADS-P: Revised Child Anxiety and Depression Scale–Parent version.

[d]CAIS-P: Child Anxiety Impact Scale–Parent version.

[e]Significance testing was based on Wilcoxon signed-rank tests for the Child Anxiety Impact Scale–Parent version home and social subscales; otherwise, significance was based on paired sample *t* tests.

[f]MDD: major depressive disorder.

Comparison of the first and last ever child-rated GBO in relation to an active goal established if playing Lumi Nova was associated with therapy-aligned improvement as determined by users. In total, 54 (81%) of the 67 players with game play data selected a goal and subsequently recorded a GBO score for exposure challenges. Out-of-game exposure challenges associated with that goal were recorded for 43 (64%) of the 67 players with game play data, and 45 (67%) players rated their progress by completing at least two GBO scores. A Wilcoxon signed-rank test showed that there was a significant difference between the first and last outcome score over the course of the intervention (*z*=2.433; *P*=.02). On average, players indicated that they had moved closer to reaching their goal over a period of game play, that is, the median score of 7 at the last assessment was significantly higher than the median score of 5 at the first assessment.

Of the 30 guardians who completed the follow-up survey at T2, 16 (53%) provided optional open-response comments. The responses were collated and systematized in relation to the primary study objectives: effectiveness, user engagement and experience, and safety (Table 3).

Table 3. Guardian open-response content summarized by research domain (n=16).

Research domain and summarized content	Comments, n (%)
Effectiveness	
Increased confidence and bravery to tackle challenges	6 (38)
Increased appreciation that taking small steps is helpful	3 (19)
Perceived progression in relation to goal choice	2 (13)
Facilitated discussion about anxiety	1 (6)
Beneficial in conjunction with other support	1 (6)
Engagement and experience	
Neutral endorsement of use	5 (31)
Laudatory comments	4 (25)
Barriers to adoption (design and process)	6 (38)
Barriers to adoption (technical barriers)	2 (13)
Increased frustration	1 (6)
Safety and adverse outcomes	
Adverse outcomes	0 (0)

Regarding effectiveness, comments in this domain were all related to positive improvements in anxiety-related outcomes; 6 (20%) of the 30 guardians described witnessing an increase in confidence or bravery in their child and suggested that children were able to recognize fears and successfully challenge their thoughts:

> *When she did the challenge, getting an answer wrong, that gave her a bit of confidence that [a] little mistake doesn't put one in trouble by teachers.* [guardian of a girl, aged 12 years]

For one child, playing Lumi Nova prompted greater discussion around fears and worries. The child's guardian said, "He seems more willing to talk about feeling anxious, he asks questions about anxiety" [guardian of a boy, aged 9 years]. Guardians felt that Lumi Nova had generated new learning in line with core processes of exposure therapy about what happens when an anxiety-provoking situation occurs and was effective in helping children work through a step-by step approach:

> *She liked knowing that she could take small steps towards a recognised fear and liked remembering that she coped with all those steps comfortably.* [guardian of a girl, aged 7 years]

> *He took to the game very well and I think it helped him rationalise one of his fears – staying away from home...I definitely think the game put in some excellent groundwork for him to draw on going forward.* [guardian of a boy, aged 12 years]

In one case, a guardian reported that the game had proved effective in conjunction with existing support: "This, along with weekly play therapy, has helped her anxiety" [guardian of a girl, aged 8 years].

User Engagement and Experience

Table 4 presents frequency data (average number of game play sessions) over the course of the intervention and duration of game play data (average number of days playing) as an indication of player engagement for those with complete game play data (n=67) and players from the T1-T2 subsample with complete game play data (n=25). Results indicate large variability in the number of times children played the game, ranging from just once to 46 individual episodes of game play out of a maximum potential of 56 episodes, with players averaging 11 (SD 9.41) sessions over a median period of 15 days.

Table 4. Average frequency and duration of game play.

	Game play sample (n=67)	T1-T2 sample (n=25)
Frequency (times played)		
Value, mean (SD)	11.22 (9.41)	12.16 (10.45)
Value, median (range)	8 (1-46)	8 (1-46)
Duration[a] (days played)		
Value, mean (SD)	18.37 (14.75)	18.28 (14.60)
Value, median (range)	15 (1-53)	16 (1-53)

[a]Duration of play from the first recorded date to the last date of game play per participant.

In total, 10 (15%) of the 67 players with game play data rated how easy they found playing Lumi Nova on a scale from 1 (very easy) to 5 (very hard); 8 (12%) players provided a positive or neutral evaluation, with most (6/8, 75%) finding the game easy or very easy, and the rest (2/8, 25%) finding the game neither easy nor hard. Furthermore, 3% (2/67) of players reported finding the game very hard to play.

In total, 18 open-response comments related to player engagement and experience of using Lumi Nova in the T2 guardian survey (Table 3), and 5 neutral comments endorsed the adoption of the game. For example, "Downloaded it and played most days for several weeks" [guardian of a girl, aged 7 years]. A total of 12 comments specifically captured interest in playing the game and its appeal to children. Of these 12 comments, 4 (33%) were laudatory. For example, "[My son] played the game approximately 10 times. He enjoyed it very much..." [guardian of a boy, aged 12 years] and "We'll miss Lumi Nova...She wanted the chance to deal with other anxieties" [guardian of a girl, aged 7 years]. Six comments suggested that although the premise of the game or elements within it were appealing, there were barriers to its adoption that related to the target audience (a perception it was pitched too young), restrictions in the game processes (eg, limited choice or low relevance of options or insufficient challenge), or a perception of repetition that children found frustrating:

> *My daughter lost interest in the game and thought it was more aimed at younger children. She has specific worries that weren't covered.* [guardian of a girl, aged 7 years]

> *The feelings bit at the beginning was good, but the tasks following this could be repetitive.* [guardian of a boy, aged 9 years]

Two guardians commented specifically on technical difficulties (subsequently redressed), which affected the player experience (eg, difficulties downloading the game or saving progress). One guardian simply reported that playing Lumi Nova made her daughter (aged 10 years) *frustrated* but provided no additional context.

Safety of Lumi Nova

Playing Lumi Nova was not associated with increased symptoms of low mood over the course of the intervention, that is, the mean RCADS-P major depressive disorder scores did not increase from T1 to T2 (Table 2). At T2, 30 parents provided data regarding any positive or negative changes in their child, which were perceived as connected to playing Lumi Nova. In total, of the 30 parents, 22 (73%) reported no change, and the remainder (n=8, 27%) reported positive associated outcomes. No adverse events were spontaneously reported during the course of the intervention. Therefore, overall, there was no

evidence to suggest harm or unintended negative consequences associated with playing Lumi Nova.

Discussion

Principal Findings

This small-scale preliminary evaluation study examined the effectiveness, user engagement and experience, and safety of Lumi Nova, a mobile app delivering targeted exposure-based CBT strategies for children with mild to moderate difficulties with anxiety. Over an 8-week period of game play, we found that playing Lumi Nova was associated with a reduction in anxiety symptom severity and progress toward treatment goals, and this effectiveness was positively endorsed by guardians. The children engaged with the content and did so safely.

Regarding the app's effectiveness, there was a reduction in the guardian-rated mean anxiety symptom severity (SCAS-P-8) between T1 and T2 with a small to moderate effect. Such findings are consistent with the literature showing moderate effectiveness in computer-based CBT for childhood anxiety [25,26,36] and contribute to emerging findings from tests on the effectiveness of game-based interventions that have reported moderate child- and parent-rated improvements in symptom severity after a short period of game play [45]. This was a small, low-powered study, and it would be important to establish effectiveness in a larger study. As a simple noncomparative evaluation, we cannot directly attribute symptom severity reduction to Lumi Nova, and the use of an active control group in future studies would help establish whether improvements in anxiety symptomatology could be attributable to the app. Nonetheless, in open responses, guardians associated anxiety-related improvements in children to game play, commenting additionally on perceived broader benefits in relation to increased confidence and successful new learning about stepped approaches to tackling fears and worries.

For player-rated effectiveness, children recorded positive movement toward achieving a self-identified therapy-aligned goal (ie, GBO) over the course of the intervention, on average moving up 2 points toward achieving their goal. Clinically, involving children in the setting and tracking of therapeutic goals provides an essential element of agency and personal activation, which may improve treatment outcomes [30]. Lumi Nova enables players to set a target and chart and reflect on their own progress and demonstrates how the mechanics of a mobile app can facilitate personalization and relevance of treatment. Further exploration of the contribution of this functionality to treatment experience and outcomes would be beneficial. It is noteworthy that this positive child-rated progression contrasts with the parent-reported measurements of effectiveness, which did not support perceived functional improvement in symptom impact. Parent and child informants rating child anxiety symptoms in clinical samples have shown variability in their capacity to identify anxiety disorders [57]. It may also be that parents did not pick up on goal progression in the same way as their child did or that the parent-rated outcome measures were not sufficiently sensitive to this progression. In fact, open-response comments from guardians that identify several positive benefits from participation in their child align with child-reported positive progression. The findings underscore the value of multi-informant approaches in the evaluation of treatment gains. The addition of teacher-rated response measures would offer an additional marker to gauge improvement, particularly where functional impairments manifest within a school context are less apparent at home.

In terms of user engagement and experience, evidence was provided from game play data capturing the quantity of play (frequency and duration of sessions) to indicate game adoption and repeated use over the intervention period. On average, children played Lumi Nova 11 times (SD 9.41) over 18 days (SD 14.75). However, these engagement metrics varied considerably among the players. In addition, data were not reported on the duration of each session of game play, which would help establish that the sessions involved meaningful interaction. In addition to objective (game play) markers, there was also modest support from the limited data that children found the game easy to use. Open-response comments reinforced that children played the game on multiple occasions, sometimes with parents, over many weeks and appeared to enjoy doing so.

It is interesting to note that there is little shared understanding or agreement of what constitutes sufficient engagement for mHealth apps [41] and there is a lack of established usability measures for children [39]. No predefined threshold of sufficient engagement to deliver impact has been specified by the developers of Lumi Nova or targeted in this preliminary study. It is recognized that the optimal dose for intervention effectiveness is likely to vary depending on the user characteristics and context [40]. Notably, 54 (81%) of the 67 players for whom there was game play data selected a goal, completed associated in-game exposure steps and reflections, and recorded at least one GBO score; almost two-thirds (43/67, 64% of players) went on to complete related out-of-game challenges. This engagement with the therapeutic mechanics of the game provides an indicator of engagement breadth and depth [45]. Recently, Zhang et al [58] have suggested that greater understanding of beneficial app interaction for digital health interventions is derived from considering *clinically meaningful activity*, that is, the completion of behaviors indicative of meaningful use (learning, goal-setting, and self-tracking), which is not captured by the *quantity* of engagement. We can gauge user progress in Lumi Nova through in-app progression which is broken down into linear steps. This modular approach is modeled on exposure therapy where each session of use translates to clinically meaningful contact when compared with face-to-face delivery. The receipt of a GBO response thus establishes user progress, as a GBO query event is only triggered when a user has successfully completed all previous steps. Altogether, our findings offer a preliminary indication that Lumi Nova provided an experience that engaged and maintained interest and facilitated progression. However, further work employing inferential analyses which explores how children engage with Lumi Nova (eg, the quantity of play and completion of meaningful activities in game and in vivo) relates to improvements in anxiety symptoms and interference would provide an indication of what might constitute effective and sufficient engagement to deliver treatment benefit [40,41,59].

Gamification is seen as a strategy to increase engagement and adherence with digital mental health interventions by delivering therapeutic content in a format with intrinsic appeal for children [30,33]. To date, few game-based digital mental health interventions specifically developed for children have been empirically evaluated. However, the limited literature that has explored 3D computer and immersive video game approaches for the treatment of anxiety has shown that children enjoy and engage with game-based therapeutic approaches [29] and supports game-based tools to supplement the delivery of therapist-led CBT [44,46]. Lumi Nova's application of immersive technology and augmented reality to deliver exposure-based CBT strategies in a standalone mobile app is therefore a novel contribution to an emerging evidence base.

Relatively few apps for anxiety in childhood implemented in *real-world* (nontrial) settings have been empirically evaluated [25,26,35]. Promising findings have supported the clinician-supported delivery of CBT skills via smartphones [43]. Our findings extend our understanding of how digital apps can be used to deliver remote self-help interventions and further support the potential of mobile apps to widen reach and facilitate early access to effective treatments for anxiety [26,34]. Given the poor prognosis of anxiety disorders in children when left untreated and the associated burden on health care [60], exploring the potential of digital tools to facilitate and optimize early access to effective treatment and thus prevent the escalation of symptoms and functional impairment is an important focus. Notably, most children recruited to our study did not seek or receive previous treatment for anxiety before starting the intervention, suggesting that participation offered access to evidence-based treatment to a group with an identified need, but for the most part, hidden from services.

Lumi Nova was developed using a robust co-design framework that involved children, parents, teachers, clinicians, academics, and technical experts in prototype design, development, and evaluation via rapid user-testing. This is a strength of the app and in line with guidance, which has called for increased co-design processes that actively engage the intended users and other stakeholders throughout the development cycle of digital game-based innovations for mental health [25,30]. Nonetheless, challenges remain in creating content that maximizes engagement and adherence across a span of ages, disorders, and abilities, which can offer only limited individualization. The ability to respond quickly and modify is an advantage of agile development processes in digital mental health delivery; consequently, many of the learnings identified by children and guardians during this early evaluation (eg, to improve game progression and rewards and cater to a wider range of game play abilities) have now been incorporated. Traditional intervention approaches that assess effectiveness once development is complete diminish the value that can be gained from evaluation during the development process. Digital intervention development enables an iterative multi-cycle approach to improving interventions, codeveloping with users and other stakeholders, as an explicit part of the development process. Rigorous evaluation at an early development phase (as in this study) can improve readiness for product launch. This approach facilitated the achievement of regulatory status (Medicines & Health Care Products Regulatory Agency) for Lumi Nova and its subsequent full market launch.

Limitations and Future Directions

Although children adopted and engaged with Lumi Nova, and the game play sample was sufficient to demonstrate its use, the evidence of at least one or more sessions of game play was available for only around half of those consenting to play at T1. Analytic information about game play sessions was captured for analysis only when the player's device had internet connectivity, enabling data to be sent to the data hub, which was not always achieved consistently every week, as directed. Therefore, it is possible that our data underrepresent true player interest and the adoption of the game (ie, game play occurred offline). The drop from those with preintervention consent (n=120) to those with guardians activating an access key (n=74) may have resulted from technical difficulties that guardians faced in downloading the beta version of the game as well as the additional requirement on guardians to complete the SCAS-P-8 to generate the game key. Therefore, poorer uptake may index the study burden on guardians rather than the game's appeal among players. It would be interesting to analyze adoption and use in a natural (nonstudy) setting. Close partnerships working with teachers and guardians, including practical support with processes of enrollment in the study and game setup, were provided to maximize engagement in the study; nonetheless, guardian retention was a challenge, and this was consistent with other evaluation studies in digital mental health [31]. In addition, the data collection overlapped with the COVID-19 pandemic and the national lockdown in the United Kingdom, which may have had an impact on the study involvement. No data in this study were captured or analyzed from users of the hub (ie, education professionals). It is not clear therefore how users were engaging with the hub and how its functionality, such as access to real-time evidence of player progress, was adopted to support professional decision-making. Of note, the final T1-T2 sample was not sociodemographically diverse and outcomes for this sample may not reflect those that would be obtained (or the appeal more generally for the game) within a broader cross-section of the population.

Further work to establish the maintenance of treatment gains over the short and long term would be an important next step in establishing the effectiveness of Lumi Nova. A study powered to explore potential moderators of effectiveness, engagement, and experience (eg, age, gender, anxiety presentation, additional comorbidities, and disability) would also help clarify who is likely to benefit from playing Lumi Nova and in what circumstances. Contextual factors associated with home-based engagement, such as the level of parental involvement, could be explored [26]. Evidence has shown that parental involvement may play a role in child treatment adherence in CBT [61]. As a remotely delivered digital self-help tool that requires guardian facilitation and supervision, the role of guardian motivation and encouragement to support child engagement with the game remains unclear. As a future direction, it is important to analyze optimum approaches for integrating evidenced digital interventions within care pathways. Work to examine how Lumi Nova sits within and complements the health care ecosystem could, for example, include exploring its clinical use as an

adjunct to face-to-face treatment, or where treatment is delayed [26]. Limited health economic data have been reported to support the use of digital health interventions [26,29]. Establishing the cost-effectiveness of Lumi Nova would be an important step in clarifying the value proposition of incorporating a commercially available digital self-help intervention within a clinical implementation model.

Conclusions

App-based treatment platforms that deliver therapeutic content via gaming technology may provide an opportunity to offer effective early intervention for childhood anxiety disorders and address documented barriers to successful treatment by delivering an appealing and acceptable option for children experiencing difficulties with anxiety that can be accessed within a home environment. This small-scale evaluation study provides early evidence in support of the effectiveness, safety, and acceptability (user engagement and experience) of Lumi Nova, a coproduced and collaboratively developed self-help app delivering exposure-based CBT strategies via immersive technology. Further evaluation is recommended to support and extend these preliminary findings.

Acknowledgments

The authors thank all the study participants who contributed to this research. JL, LW, and JM acknowledge the financial support of the National Institute of Health Research Nottingham Biomedical Research Centre and National Institute of Health Research MindTech Med Tech Co-operative.

Conflicts of Interest

This paper details a preliminary independent evaluation study completed by MindTech Med Tech Co-operative as part of a collaborative project with BfB Labs Ltd (the creator of Lumi Nova) and the University of Reading. The development of Lumi Nova was funded through a small business research development contract awarded to BfB Labs Ltd by National Health Service England.

References

1. Sadler K, Vizard T, Ford T, Goodman A, Goodman R, McManus S. Mental Health of Children and Young People in England, 2017: Trends and Characteristics. Leeds, UK: NHS Digital; 2018.
2. Beesdo K, Knappe S, Pine DS. Anxiety and anxiety disorders in children and adolescents: developmental issues and implications for DSM-V. Psychiatr Clin North Am 2009 Sep;32(3):483-524 [FREE Full text] [doi: 10.1016/j.psc.2009.06.002] [Medline: 19716988]
3. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
4. Essau CA, Lewinsohn PM, Lim JX, Ho MR, Rohde P. Incidence, recurrence and comorbidity of anxiety disorders in four major developmental stages. J Affect Disord 2018 Mar 01;228:248-253 [FREE Full text] [doi: 10.1016/j.jad.2017.12.014] [Medline: 29304469]
5. Copeland WE, Angold A, Shanahan L, Costello EJ. Longitudinal patterns of anxiety from childhood to adulthood: the Great Smoky Mountains Study. J Am Acad Child Adolesc Psychiatry 2014 Jan;53(1):21-33 [FREE Full text] [doi: 10.1016/j.jaac.2013.09.017] [Medline: 24342383]
6. Balázs J, Miklósi M, Keresztény A, Hoven CW, Carli V, Wasserman C, et al. Adolescent subthreshold-depression and anxiety: psychopathology, functional impairment and increased suicide risk. J Child Psychol Psychiatry 2013 Jun;54(6):670-677. [doi: 10.1111/jcpp.12016] [Medline: 23330982]
7. Pearcey S. Report 04: changes in children and young people's emotional and behavioural difficulties through lockdown. Co-space Study: Covid-19 Supporting Parents Children and Adolescents During Epidemics. 2020. URL: https://emergingminds.org.uk/wp-content/uploads/2020/06/CoSPACE-Report-4-June-2020.pdf [accessed 2021-03-22]
8. Higa-McMillan CK, Francis SE, Rith-Najarian L, Chorpita BF. Evidence base update: 50 years of research on treatment for child and adolescent anxiety. J Clin Child Adolesc Psychol 2016;45(2):91-113. [doi: 10.1080/15374416.2015.1046177] [Medline: 26087438]
9. James A, James G, Cowdrey FA, Soler A, Choke A. Cognitive behavioural therapy for anxiety disorders in children and adolescents. Cochrane Database Syst Rev 2015 Feb 18(2):CD004690 [FREE Full text] [doi: 10.1002/14651858.CD004690.pub4] [Medline: 25692403]
10. Schwartz C, Barican JL, Yung D, Zheng Y, Waddell C. Six decades of preventing and treating childhood anxiety disorders: a systematic review and meta-analysis to inform policy and practice. Evid Based Ment Health 2019 Aug;22(3):103-110 [FREE Full text] [doi: 10.1136/ebmental-2019-300096] [Medline: 31315926]
11. Seligman LD, Ollendick TH. Cognitive-behavioral therapy for anxiety disorders in youth. Child Adolesc Psychiatr Clin N Am 2011 Apr;20(2):217-238 [FREE Full text] [doi: 10.1016/j.chc.2011.01.003] [Medline: 21440852]

12. Reardon T, Spence SH, Hesse J, Shakir A, Creswell C. Identifying children with anxiety disorders using brief versions of the Spence Children's Anxiety Scale for children, parents, and teachers. Psychol Assess 2018 Oct;30(10):1342-1355 [FREE Full text] [doi: 10.1037/pas0000570] [Medline: 29902050]
13. Reardon T, Harvey K, Creswell C. Seeking and accessing professional support for child anxiety in a community sample. Eur Child Adolesc Psychiatry 2020 May;29(5):649-664 [FREE Full text] [doi: 10.1007/s00787-019-01388-4] [Medline: 31410579]
14. Reardon T, Harvey K, Young B, O'Brien D, Creswell C. Barriers and facilitators to parents seeking and accessing professional support for anxiety disorders in children: qualitative interview study. Eur Child Adolesc Psychiatry 2018 Aug;27(8):1023-1031 [FREE Full text] [doi: 10.1007/s00787-018-1107-2] [Medline: 29372331]
15. Crouch L, Reardon T, Farrington A, Glover F, Creswell C. "Just keep pushing": parents' experiences of accessing child and adolescent mental health services for child anxiety problems. Child Care Health Dev 2019 Jul;45(4):491-499. [doi: 10.1111/cch.12672] [Medline: 30990911]
16. Velasco AA, Cruz IS, Billings J, Jimenez M, Rowe S. What are the barriers, facilitators and interventions targeting help-seeking behaviours for common mental health problems in adolescents? A systematic review. BMC Psychiatry 2020 Jun 11;20(1):293 [FREE Full text] [doi: 10.1186/s12888-020-02659-0] [Medline: 32527236]
17. de Haan AM, Boon AE, de Jong JT, Hoeve M, Vermeiren RR. A meta-analytic review on treatment dropout in child and adolescent outpatient mental health care. Clin Psychol Rev 2013 Jul;33(5):698-711. [doi: 10.1016/j.cpr.2013.04.005] [Medline: 23742782]
18. Lee P, Zehgeer A, Ginsburg GS, McCracken J, Keeton C, Kendall PC, et al. Child and adolescent adherence with cognitive behavioral therapy for anxiety: predictors and associations with outcomes. J Clin Child Adolesc Psychol 2019;48(sup1):215-226 [FREE Full text] [doi: 10.1080/15374416.2017.1310046] [Medline: 28448176]
19. Whiteside SP, Ale CM, Young B, Dammann JE, Tiede MS, Biggs BK. The feasibility of improving CBT for childhood anxiety disorders through a dismantling study. Behav Res Ther 2015 Oct;73:83-89. [doi: 10.1016/j.brat.2015.07.011] [Medline: 26275761]
20. Whiteside SP, Deacon BJ, Benito K, Stewart E. Factors associated with practitioners' use of exposure therapy for childhood anxiety disorders. J Anxiety Disord 2016 May;40:29-36 [FREE Full text] [doi: 10.1016/j.janxdis.2016.04.001] [Medline: 27085463]
21. Deacon BJ, Lickel JJ, Farrell NR, Kemp JJ, Hipol LJ. Therapist perceptions and delivery of interoceptive exposure for panic disorder. J Anxiety Disord 2013 Mar;27(2):259-264. [doi: 10.1016/j.janxdis.2013.02.004] [Medline: 23549110]
22. Gola JA, Beidas RS, Antinoro-Burke D, Kratz HE, Fingerhut R. Ethical considerations in exposure therapy with children. Cogn Behav Pract 2016 May;23(2):184-193 [FREE Full text] [doi: 10.1016/j.cbpra.2015.04.003] [Medline: 27688681]
23. Ebert DD, Zarski A, Christensen H, Stikkelbroek Y, Cuijpers P, Berking M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One 2015;10(3):e0119895 [FREE Full text] [doi: 10.1371/journal.pone.0119895] [Medline: 25786025]
24. Hill C, Creswell C, Vigerland S, Nauta MH, March S, Donovan C, et al. Navigating the development and dissemination of internet cognitive behavioral therapy (iCBT) for anxiety disorders in children and young people: A consensus statement with recommendations from the #iCBTLorentz Workshop Group. Internet Interv 2018 Jun;12:1-10 [FREE Full text] [doi: 10.1016/j.invent.2018.02.002] [Medline: 30135763]
25. Pennant ME, Loucas CE, Whittington C, Creswell C, Fonagy P, Fuggle P, Expert Advisory Group. Computerised therapies for anxiety and depression in children and young people: a systematic review and meta-analysis. Behav Res Ther 2015 Apr;67:1-18. [doi: 10.1016/j.brat.2015.01.009] [Medline: 25727678]
26. Hollis C, Falconer CJ, Martin JL, Whittington C, Stockton S, Glazebrook C, et al. Annual research review: digital health interventions for children and young people with mental health problems - a systematic and meta-review. J Child Psychol Psychiatry 2017 Apr;58(4):474-503. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
27. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
28. Jones RB, Stallard P, Agha SS, Rice S, Werner-Seidler A, Stasiak K, et al. Practitioner review: co-design of digital mental health technologies with children and young people. J Child Psychol Psychiatry 2020 Aug;61(8):928-940 [FREE Full text] [doi: 10.1111/jcpp.13258] [Medline: 32572961]
29. Halldorsson B, Hill C, Waite P, Partridge K, Freeman D, Creswell C. Annual research review: immersive virtual reality and digital applied gaming interventions for the treatment of mental health problems in children and young people: the need for rigorous treatment development and clinical evaluation. J Child Psychol Psychiatry 2021 May;62(5):584-605. [doi: 10.1111/jcpp.13400] [Medline: 33655534]
30. Fleming TM, de Beurs D, Khazaal Y, Gaggioli A, Riva G, Botella C, et al. Maximizing the impact of e-therapy and serious gaming: time for a paradigm shift. Front Psychiatry 2016;7:65 [FREE Full text] [doi: 10.3389/fpsyt.2016.00065] [Medline: 27148094]

31. Bergin AD, Vallejos EP, Davies EB, Daley D, Ford T, Harold G, et al. Preventive digital mental health interventions for children and young people: a review of the design and reporting of research. NPJ Digit Med 2020;3:133 [FREE Full text] [doi: 10.1038/s41746-020-00339-7] [Medline: 33083568]
32. Children and parents: media use and attitudes report 2019. Ofcom. 2020. URL: https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/children-and-parents-media-use-and-attitudes-report-2019 [accessed 2021-03-22]
33. Fleming TM, Bavin L, Stasiak K, Hermansson-Webb E, Merry SN, Cheek C, et al. Serious games and gamification for mental health: current status and promising directions. Front Psychiatry 2016;7:215 [FREE Full text] [doi: 10.3389/fpsyt.2016.00215] [Medline: 28119636]
34. Donker T, Petrie K, Proudfoot J, Clarke J, Birch M, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247 [FREE Full text] [doi: 10.2196/jmir.2791] [Medline: 24240579]
35. Bry LJ, Chou T, Miguel E, Comer JS. Consumer smartphone apps marketed for child and adolescent anxiety: a systematic review and content analysis. Behav Ther 2018 Mar;49(2):249-261 [FREE Full text] [doi: 10.1016/j.beth.2017.07.008] [Medline: 29530263]
36. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176 [FREE Full text] [doi: 10.2196/jmir.7332] [Medline: 28546138]
37. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
38. Whiteside SP. Mobile device-based applications for childhood anxiety disorders. J Child Adolesc Psychopharmacol 2016 Apr;26(3):246-251. [doi: 10.1089/cap.2015.0010] [Medline: 26244903]
39. Stoll RD, Pina AA, Gary K, Amresh A. Usability of a smartphone application to support the prevention and early intervention of anxiety in youth. Cogn Behav Pract 2017 Nov;24(4):393-404 [FREE Full text] [doi: 10.1016/j.cbpra.2016.11.002] [Medline: 29056845]
40. Pham Q, Graham G, Carrion C, Morita PP, Seto E, Stinson JN, et al. A library of analytic indicators to evaluate effective engagement with consumer mhealth apps for chronic conditions: scoping review. JMIR Mhealth Uhealth 2019 Jan 18;7(1):e11941 [FREE Full text] [doi: 10.2196/11941] [Medline: 30664463]
41. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
42. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019;10:759 [FREE Full text] [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
43. Pramana G, Parmanto B, Lomas J, Lindhiem O, Kendall PC, Silk J. Using mobile health gamification to facilitate cognitive behavioral therapy skills practice in child anxiety treatment: open clinical trial. JMIR Serious Games 2018 May 10;6(2):e9 [FREE Full text] [doi: 10.2196/games.8902] [Medline: 29748165]
44. van der Meulen H, McCashin D, O'Reilly G, Coyle D. Using computer games to support mental health interventions: naturalistic deployment study. JMIR Ment Health 2019 May 09;6(5):e12430 [FREE Full text] [doi: 10.2196/12430] [Medline: 31094346]
45. Schoneveld EA, Malmberg M, Lichtwarck-Aschoff A, Verheijen GP, Engels RC, Granic I. A neurofeedback video game (MindLight) to prevent anxiety in children: a randomized controlled trial. Comput Hum Behav 2016 Oct;63:321-333. [doi: 10.1016/j.chb.2016.05.005]
46. Brezinka V. Ricky and the Spider - a video game to support cognitive behavioural treatment of children with obsessive-compulsive disorder. Clin Neuropsychiatry 2013;10(3):6-12 [FREE Full text] [doi: 10.5167/uzh-93917]
47. Wols A, Lichtwarck-Aschoff A, Schoneveld EA, Granic I. In-game play behaviours during an applied video game for anxiety prevention predict successful intervention outcomes. J Psychopathol Behav Assess 2018;40(4):655-668 [FREE Full text] [doi: 10.1007/s10862-018-9684-4] [Medline: 30459485]
48. Craske MG, Treanor M, Conway CC, Zbozinek T, Vervliet B. Maximizing exposure therapy: an inhibitory learning approach. Behav Res Ther 2014 Jul;58:10-23 [FREE Full text] [doi: 10.1016/j.brat.2014.04.006] [Medline: 24864005]
49. Przybylski AK. Electronic gaming and psychosocial adjustment. Pediatrics 2014 Sep;134(3):716-722. [doi: 10.1542/peds.2013-4021] [Medline: 25092934]
50. Spence SH. A measure of anxiety symptoms among children. Behav Res Ther 1998 May;36(5):545-566. [doi: 10.1016/s0005-7967(98)00034-5] [Medline: 9648330]
51. Chorpita BF, Yim L, Moffitt C, Umemoto LA, Francis SE. Assessment of symptoms of DSM-IV anxiety and depression in children: a revised child anxiety and depression scale. Behav Res Ther 2000 Aug;38(8):835-855. [doi: 10.1016/s0005-7967(99)00130-8] [Medline: 10937431]
52. Langley AK, Bergman RL, McCracken J, Piacentini JC. Impairment in childhood anxiety disorders: preliminary examination of the child anxiety impact scale-parent version. J Child Adolesc Psychopharmacol 2004;14(1):105-114. [doi: 10.1089/104454604773840544] [Medline: 15142397]

53. Langley AK, Falk A, Peris T, Wiley JF, Kendall PC, Ginsburg G, et al. The child anxiety impact scale: examining parent- and child-reported impairment in child anxiety disorders. J Clin Child Adolesc Psychol 2014;43(4):579-591 [FREE Full text] [doi: 10.1080/15374416.2013.817311] [Medline: 23915200]
54. Law D, Jacob J. Goals and Goals Based Outcomes (GBOs): Some Useful Information. Third Edition. London, UK: CAMHS Press; 2015.
55. Wozney L, Baxter P, Newton AS. Usability evaluation with mental health professionals and young people to develop an internet-based cognitive-behaviour therapy program for adolescents with anxiety disorders. BMC Pediatr 2015 Dec 16;15:213 [FREE Full text] [doi: 10.1186/s12887-015-0534-1] [Medline: 26675420]
56. Erlingsson C, Brysiewicz P. A hands-on guide to doing content analysis. Afr J Emerg Med 2017 Sep;7(3):93-99 [FREE Full text] [doi: 10.1016/j.afjem.2017.08.001] [Medline: 30456117]
57. Wei C, Hoff A, Villabø MA, Peterman J, Kendall PC, Piacentini J, et al. Assessing anxiety in youth with the multidimensional anxiety scale for children. J Clin Child Adolesc Psychol 2014;43(4):566-578 [FREE Full text] [doi: 10.1080/15374416.2013.814541] [Medline: 23845036]
58. Zhang R, Nicholas J, Knapp AA, Graham AK, Gray E, Kwasny MJ, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J Med Internet Res 2019 Dec 20;21(12):e15644 [FREE Full text] [doi: 10.2196/15644] [Medline: 31859682]
59. Deacon BJ, Farrell NR, Kemp JJ, Dixon LJ, Sy JT, Zhang AR, et al. Assessing therapist reservations about exposure therapy for anxiety disorders: the Therapist Beliefs about Exposure Scale. J Anxiety Disord 2013 Dec;27(8):772-780. [doi: 10.1016/j.janxdis.2013.04.006] [Medline: 23816349]
60. Creswell C, Waite P. Recent developments in the treatment of anxiety disorders in children and adolescents. Evid Based Ment Health 2016 Aug;19(3):65-68. [doi: 10.1136/eb-2016-102353] [Medline: 27402874]
61. Wei C, Kendall PC. Parental involvement: contribution to childhood anxiety and its treatment. Clin Child Fam Psychol Rev 2014 Dec;17(4):319-339. [doi: 10.1007/s10567-014-0170-6] [Medline: 25022818]

Abbreviations

CAIS-P: Child Anxiety Impact Scale–Parent version
CAMHS: Children and Adolescent Mental Health Service
CBT: cognitive behavioral therapy
GBO: goal-based outcome
mHealth: mobile health
RCADS-P: Revised Child Anxiety and Depression Scale–Parent version
SCAS-P-8: Spence Child Anxiety Scale–Parent version

Edited by J Torous; submitted 22.03.21; peer-reviewed by J Hamid, M Potter, S Badawy; comments to author 10.08.21; revised version received 18.10.21; accepted 18.10.21; published 24.01.22.

Please cite as:
Lockwood J, Williams L, Martin JL, Rathee M, Hill C
Effectiveness, User Engagement and Experience, and Safety of a Mobile App (Lumi Nova) Delivering Exposure-Based Cognitive Behavioral Therapy Strategies to Manage Anxiety in Children via Immersive Gaming Technology: Preliminary Evaluation Study
JMIR Ment Health 2022;9(1):e29008
URL: https://mental.jmir.org/2022/1/e29008
doi:10.2196/29008
PMID:35072644

Original Paper

Patient Satisfaction and Recommendations for Delivering a Group-Based Intensive Outpatient Program via Telemental Health During the COVID-19 Pandemic: Cross-sectional Cohort Study

Michelle K Skime[1], MS; Ajeng J Puspitasari[1], PhD; Melanie T Gentry[1], MD; Dagoberto Heredia Jr[1], PhD; Craig N Sawchuk[1], PhD; Wendy R Moore[2], MSN, RN, NE-BC; Monica J Taylor-Desir[1], MD; Kathryn M Schak[1], MD

[1]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, United States

[2]Department of Nursing, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Michelle K Skime, MS
Department of Psychiatry and Psychology
Mayo Clinic
200 First Street SW
Rochester, MN, 55902
United States
Phone: 1 507 255 0501
Email: skime.michelle@mayo.edu

Abstract

Background: Although group-based intensive outpatient programs (IOPs) are a level of care commonly utilized by adults with serious mental illness, few studies have examined the acceptability of group-based IOPs that required rapid transition to a telemental health (TMH) format during the COVID-19 pandemic.

Objective: The aim of this study was to evaluate patient satisfaction and future recommendations for a group-based IOP that was transitioned to a TMH format during the COVID-19 pandemic.

Methods: A 17-item patient satisfaction questionnaire was completed by patients at discharge and covered 3 areas: IOP TMH satisfaction, future recommendations, and video technology challenges. Descriptive and content analyses were conducted for the quantitative and open-ended questions, respectively.

Results: A total of 76 patients completed the program in 2020. A subset of patients (n=40, 53%) responded to the survey at program discharge. The results indicated that the patients were satisfied overall with the TMH program format; 50% (n=20) of the patients preferred the program continue offering the TMH format, and the rest preferred returning to in-person formats after the pandemic. The patients indicated the elements of the program that they found most valuable and provided recommendations for future program improvement.

Conclusions: Overall, adults with serious mental illness reported high satisfaction with the group-based IOP delivered via TMH. Health care systems may want to consider offering both TMH and in-person formats regardless of the state of the pandemic. Patients' feedback on future improvements should be considered to help ensure long-term success.

(JMIR Ment Health 2022;9(1):e30204) doi:10.2196/30204

KEYWORDS

COVID-19; telemental health; teletherapy; telepsychiatry; telemedicine; intensive outpatient; patient satisfaction

Introduction

The COVID-19 pandemic has led to increased demand for mental health services worldwide, and most countries are reporting significant disruptions to the delivery of critical mental health services [1]. Early evidence suggests that symptoms of anxiety, depression, and self-reported stress were common responses to COVID-19 in the general population [2]. Concerns that suicide rates during and after the pandemic might increase have been highlighted [3], though data are still limited on the rates and risk of suicide in the context of the current pandemic. Certain populations, such as those with serious mental illness (SMI), may be particularly vulnerable to the stressors and hardships related to COVID-19. Thus, it is pertinent to ensure

adequate access to behavioral health services during this pandemic, particularly for adults with SMI.

The COVID-19 pandemic has created significant obstacles to the delivery of mental health services, especially for services delivered in a group setting due to the need for social distancing. However, maintaining access to group-based interventions is essential given their efficiency in treatment delivery to a larger population when resources are limited. Telemental health (TMH), defined as the delivery of mental health care services at a distance through the use of information and telecommunications technology, has emerged during the COVID-19 pandemic as an essential platform to ensure continuous mental health care delivery. TMH has been shown to be highly effective and increases access to care [4]. It has been shown to be an effective mode of health care delivery across different patient populations, diagnoses, and settings, including group interventions [5-7]. The COVID-19 Federal Emergency Order temporarily lifted several administrative barriers to TMH, allowing for its expanded use during the pandemic [8]. As a result, TMH services have been increasing substantially in the wake of COVID-19, with the veterans administration reporting a 500% increase in TMH use in the early stages of the pandemic [9]. Initial TMH studies during the pandemic have shown increased utilization and decreased no-show rates [10]. Though TMH has provided essential mental health care during this time, questions remain regarding how different populations accept and respond to TMH interventions. A study of patient satisfaction related to TMH services during the perinatal period showed that a majority of participants indicated that TMH improved their health care access and that the visit was as effective as in-person visits [11]. Understanding patient satisfaction and engagement with TMH interventions is crucial to the sustainability of TMH programs both during and beyond the pandemic.

Understanding patients' perspective on the quality of behavioral health services delivered via telehealth is important to ensure their engagement with treatment and to improve outcomes. Several pre–COVID-19 studies indicated that patients had a positive perception toward telehealth and were satisfied with the delivery format [12]. Although the literature is still limited, studies are also finding high patient satisfaction with telehealth programs developed during the pandemic [13,14]. Emerging research during this pandemic were consistent with previous findings indicating that patients were satisfied with the option to continue behavioral health services via telehealth. Most of this research, however, has focused on individual outpatient behavioral health services. A gap in the literature exists on patient satisfaction for group-based intensive outpatient programs (IOP) delivered via telehealth during the pandemic.

The aim of this study was to evaluate patient satisfaction while exploring future recommendations of a group-based IOP for adults with SMI, which was rapidly transformed to a telehealth format during the COVID-19 pandemic. The results from this study can be used to improve the quality of programming and enhance the delivery of services in the future.

Methods

The protocol for this cross-sectional cohort survey research was approved by the Mayo Clinic Institutional Review Board. Data were collected as part of clinical care at the Adult Transitions Program (ATP), a group-based IOP within the Mayo Clinic Department of Psychiatry and Psychology. This program was intended to treat adults with SMI who were recently discharged from psychiatric hospitalization or were at risk of psychiatric hospitalization if not treated in a more intensive level of outpatient care. Inclusion criteria for the present study were patients who were admitted to ATP, were at least 18 years old, and consented for their clinical data to be used for research purposes. The patients completed the satisfaction survey over the phone with research personnel after they were discharged from the program. The phone call took approximately 15 minutes to complete.

ATP was delivered by a multidisciplinary team that included psychologists, a psychiatrist, nurse practitioners or physician assistants, licensed professional clinical counselors, occupational therapists, and registered nurses. The patients received the program 5 days per week, 3 hours a day, for a 3-week period. The programming was mainly group-based and informed by evidence-based cognitive and behavioral interventions such as Behavioral Activation [15], dialectical behavioral therapy (DBT) [16], and acceptance and commitment therapy [17]. The patients were assigned to 1 of the 3 tracks, with 8 patients in each track. The inclusion criteria for the program were adults aged 18 years and older, who were diagnosed with SMI (eg, mood disorders, anxiety disorders, psychosis, personality disorders, and substance use), who had recent psychiatric hospitalization or were at risk for psychiatric hospitalization, and who reported having access to a mobile or computer device to connect to the video teleconference software (ie, Zoom). The exclusion criteria were cognitive impairment and higher symptom severity that did not require a higher level of care with a psychiatric hospitalization or residential settings.

The patient satisfaction questionnaire was developed through a literature review. Some items were generated based on the acceptability of intervention measure, intervention appropriateness measure, and feasibility of intervention measure by Weiner and colleagues [18]. These original measures have Cronbach alphas from .85 to .91, and test-retest reliability coefficients ranged from 0.73 to 0.88. The research team generated and reviewed the initial items, and the suggested changes included adding and removing certain questions and improving grammatical errors and wording. The research team members took each iteration of the survey to ensure the readability of the content items. The final version of the Patient Satisfaction Questionnaire (Multimedia Appendix 1) included 14 quantitative questions answered on a Likert-type scale from 1 to 5 with the higher numbers indicating higher satisfaction. Three open-ended questions assessed the patients' overall experience with TMH, the most valuable part of the TMH format, and recommendations for future program improvement. In addition, demographic variables were pulled from the electronic health record.

Descriptive statistics were generated to identify the most commonly endorsed items. The open-ended questions were analyzed using summative content analysis [19]. Keywords were identified and quantified to characterize the themes that emerged from the 3 open-ended questions. Two researchers independently read the qualitative responses multiple times to identify the keywords. These keywords were then sorted into categories, and the themes were then quantified using frequency counts. The 2 researchers compared emerging categories for validation purposes.

Results

A total of 76 patients were admitted to the program between March and August of 2020. Of the 76 patients admitted to the program, 40 (53%) completed the survey over the phone with research personnel. The referral source and track attended for those who did and did not complete the survey were similar. The referral source for completers versus noncompleters, respectively, was as follows: inpatient, 42.5% versus 35%; emergency department, 2.5% for both groups; primary care, 30% versus 27.5%; other outpatient programs, 15% versus 32.5%; and other programs, 10% versus 2.5%. The track attended for the completers versus noncompleters, respectively, were as follows: cognitive behavioral therapy morning 25% versus 30%; DBT morning 30% versus 42.5%; and DBT afternoon 45% versus 27.5%. The patients had a mean age of 36.55 (SD 13.43) years. The majority of the patients were female (n=32, 80%) and White (n=33, 82.5%), married (n=14, 35%) or single (n=23, 57.5%), cisgender (n=38, 95%), heterosexual (n=30, 75%), and employed (n=23, 57.5%). The patients had the following psychiatric diagnoses as a primary presenting problem: major depressive disorder (n=29, 72.5%), anxiety disorder (n=2, 5%), borderline personality disorder (n=6, 15%), and suicidal ideation (n=2, 5%). Full baseline characteristics are reported in Table 1.

Table 1. Baseline characteristics of study sample.

Characteristics	Values
Gender, n (%)	
Female	32 (80)
Male	6 (15)
Transgender female or male to female	1 (2.5)
Nonbinary or genderqueer	1 (2.5)
Age (years), mean (SD)	36.55 (13.43)
Race, n (%)	
White	33 (82.5)
Other	6 (15)
African American	1 (2.5)
Ethnicity, n (%)	
Hispanic or Latino	3 (7.5)
Non-Hispanic or Latino	36 (90)
Unknown	1 (2.5)
Marital status, n (%)	
Single	23 (57.5)
Married	14 (35)
Separated	2 (5)
Divorced	1 (2.5)
Employment, n (%)	
Currently employed	23 (57.5)
Not employed	14 (35)
Disabled	3 (7.5)
Financial resource strain, n (%)	
Not hard at all	17 (42.5)
Not very hard	8 (20)
Somewhat hard	10 (25)
Hard	2 (5)
Very hard	1 (2.5)
Not on file	2 (5)
Sexual orientation, n (%)	
Lesbian or gay	2 (5)
Heterosexual	30 (75)
Something else	1 (2.5)
Don’t know	2 (5)
Choose not to disclose	1 (2.5)
Presenting problems, n (%)	
Major depressive disorder	29 (72.5)
Suicidal ideation	2 (5)
Anxiety disorder	2 (5)
Borderline personality disorder	6 (15)
Other	1 (2.5)

Characteristics	Values
Comorbidity, n (%)	
Yes	17 (42.5)
No	23 (57.5)
Track, n (%)	
DBT[a] morning	12 (30)
DBT afternoon	18 (45)
CBT[b] morning	10 (25)
Source of referral, n (%)	
Inpatient	17 (42.5)
Emergency department	1 (2.5)
Primary care	12 (30)
Other outpatient	6 (15)
Other programs	4 (10)
Days completed, mean (SD)	14.4 (1.5)
Program absences (days), mean (SD)	0.7 (1.6)
Program absences (days), n (%)	
None	28 (70)
1-3	10 (25)
4-7	2 (5)

[a]DBT: dialectical behavioral therapy.

[b]CBT: cognitive behavioral therapy.

The complete results for the quantitative portion of the satisfaction survey are presented in Table 2. Overall, the majority of patients reported high satisfaction, comfort, appropriateness, relevance, and compatibility of the TMH format of ATP. Most patients (92.5% [n=37]) reported that they would recommend this service format to a friend or family member. They noted that the TMH format was well organized and executed, user friendly, and not burdensome. We also assessed preference between in-person versus a TMH format. We found a split among the patients where 35% (n=14) preferred to receive an in-person format, 50% (n=20) preferred continuing with a TMH format, and 15% (n=6) were neutral when asked, “Once COVID-19 travel restrictions are lifted, would you still want to continue with video format?” (Table 2).

Table 2. Satisfaction survey results.

Survey items	(1), n (%)	(2), n (%)	(3), n (%)	(4), n (%)	(5), n (%)
How did the care you received over video compare to a regular in-person health care visit?	2 (5)	2 (5)	11 (27.5)	12 (30)	13 (32.5)
How willing are you to use the video visit system in the near future?	1 (2.5)	1 (2.5)	4 (10)	5 (12.50)	29 (72.5)
Would you recommend this service to a friend or family member?	1 (2.5)	0 (0)	2 (5)	5 (12.5)	32 (80)
If you could choose between receiving the service in person versus video visit, which would you prefer?	12 (30)	1 (2.5)	10 (25)	5 (12.5)	12 (30)
To what extent are you satisfied with the video format of the service that you received?	1 (2.5)	2 (5)	2 (5)	15 (37.5)	20 (50)
How well-organized and well-executed was the video format of the service that you received?	1 (2.5)	0 (0)	1 (2.5)	13 (32.5)	25 (62.5)
How comfortable are you with the video format of the service that you received?	1 (2.5)	1 (2.5)	3 (7.5)	12 (30)	23 (57.5)
How user-friendly is the video format of the service that you received?	1 (2.5)	0 (0)	4 (10)	14 (35)	21 (52.5)
How burdensome it is to receive the service via video?[a]	1 (2.5)	2 (5)	3 (7.5)	9 (22.5)	25 (62.5)
How compatible was the video visit with access to devices (eg, cell phone and computer) that you already have?	1 (2.5)	0 (0)	4 (10)	8 (20)	27 (67.5)
How appropriate is it to receive the service via video versus in person?	0 (0)	0 (0)	8 (20)	11 (27.5)	21 (52.5)
How relevant is it to receive the video format versus the in-person format in your current life context?	0 (0)	1 (2.5)	4 (10)	2 (5)	33 (82.5)
Once COVID-19 travel restrictions are lifted, would you still want to continue with video format?	10 (25)	4 (10)	6 (15)	6 (15)	14 (35)
Did you have any difficulty with the telemental health format and video technology?					
Yes	18 (46.15)				
No				21 (53.85)	

[a]Reversed item.

We also assessed to what extent patients experienced technological difficulties with the TMH format. A portion of the patients (46.15% [n=18]) reported experiencing challenges during the program. We analyzed the qualitative open-ended responses and reported that challenges included problems with slow internet connection, the video camera of their devices, logging into the teleconference room, and being inadvertently removed from the session.

We conducted content analyses of the qualitative questions and extracted themes from each question. The frequency counts for the categories within each question are presented in Table 3. Examples of the qualitative feedback are presented in Table 4.

Table 3. Qualitative feedback.

Questions and categories	Values, n (%)
Patients' perceptions of the TMH[a] format	
Positive attitudes toward the format and program	28 (70)
Increased access to treatment	6 (15)
Treatment was effective and beneficial	8 (20)
Increased social support	4 (10)
Preferred in-person format	7 (18)
Technological issues	8 (20)
Negative attitudes towards the format and program	2 (5)
Most valuable part of the TMH format and the program	
Social support	9 (23)
Learning coping skills	5 (13)
The convenience that telemedicine offers	27 (68)
No valuable experience	1 (3)
Recommendations for future improvement	
Improvement on the technology or TMH delivery process	5 (13)
Improvement on therapy materials	3 (8)
Improvement on therapeutic process or delivery	5 (13)
Offering in-person format	1 (3)
No further recommendations	25 (63)

[a]TMH: telemental health.

Table 4. Examples of qualitative responses.

Questions and categories	Sample responses
Patients' perception of the TMH[a] format	
Positive attitudes toward the format and program	• "I thought it was nice… I don't mind the telehealth format. It was a lot organized. Each group was timed very well. I thought it was very pleasant for the most part" • "I was really happy with it. In fact I still use telehealth to communicate with my other providers. This is really good. I am really thankful and grateful for it."
Increased access to treatment	• "I am glad I had the option to continue receiving treatment via telehealth during COVID" • "I think it was really good especially because I live in Michigan so it would be challenging to find a different program."
Treatment was effective and beneficial	• "I thought it was weird starting off but actually it was still just like being in a room full of people. Honestly, I think it saved my life." • "So that is the positive of video format to use the skills immediately in my home environment."
Increased social support	• "it was good to see other people over video" • "It's nice to see everyone while still feeling safe."
Preferred in-person format	• "For me it is easier to do it in person. I think I would get more out of the program if it is in person." • "I very much prefer face to face. It felt more welcoming. With video you can only answer the questions. there couldn't really be a discussion like if we have face to face and sitting in the same room."
Technological issues	• "It was just hard to log on sometimes." • "A few times I was disconnected but that could have been on my end"
Negative attitudes towards the format and program	• "I didn't like it. I don't like video format."
Most valuable part of the TMH format and the program	
Social support	• "Being able to still see other patients in group via Zoom." • "You get to interact with everyone still just like when you are in person."
Learning coping skills	• "It gave me tools to overcome depression and anxiety. It gave you the tools, it just you have to learn and use it." • "You learned so much. It's not like information overload. I'm someone who learns that way. The coping skills and being able to be honest were phenomenal."
The convenience that TMH offers	• "The flexibility that we could do it from anywhere." • "Just being able to continue receiving therapy and not being cut off because of COVID. It is good to have it as an option."
No valuable experience	• "I didn't really value the program because it was in the video format."
Recommendations for future improvement	
Improvement on the technology or TMH delivery process	• "Using more of the Zoom features such as the whiteboard." • "There are ways where you could have people type on the screen, I would actually use that feature more on Zoom."
Improvement on therapy materials	• "I found a few easy things that will make the binder easier, maybe some tabs to find things [easier]" • "Maybe just making sure that we get the binder and number the pages. Or maybe give the blank copy of the materials. Maybe improving the structure of the binder. And maybe to be able to send the powerpoint and all the learning tools."
Improvement on therapeutic process or delivery	• "Maybe allow for more collaboration among the patients. They did that though in DBT group but maybe a bit more." • "The provider should be organized and know what they are teaching and explaining. Other than that they didn't see any real issue."
Offering in-person format	• "I do wish it could be in person."

Questions and categories	Sample responses
No further recommendations	• "No, I like everything about the video format." • "No. I don't think so."

[a]TMH: telemental health.

Regarding the patients' overall perception of the TMH program, they provided both positive feedback and challenges that they encountered. The patients provided overall positive attitudes toward the TMH format. They noted that TMH provided easier access to treatment and that treatment was effective and beneficial to learning skills and coping with their problems. Some individuals also reported that TMH increased social support during the pandemic. These findings are similar to those found by Ackerman et al [11], which showed increased satisfaction with TMH. Others noted challenges of this delivery format, which included experiencing technological issues, with one patient reporting an overall negative experience with the program. Some patients (18% [n=7]) also expressed preferences to receive services in-person rather than via TMH.

We asked the patients to identify the most valuable part of the program. More than half of the patients stated that they found the convenience of TMH as valuable, with others reporting the benefits from social support and the adequate learning skills to cope with their presenting problems.

Most patients did not provide further recommendations to improve the TMH program format. Some suggested improvements on the TMH delivery process, such as using more features on Zoom. Others suggested that the therapeutic delivery process and materials could be improved. One patient suggested that we offer the in-person format again once the pandemic is over.

Discussion

Principal Findings

Prior to the COVID-19 pandemic, very little information existed in the empirical literature on how to rapidly convert group-based IOPs to a TMH format. This study assessed the acceptability of a group-based IOP delivered via TMH during the COVID-19 pandemic. Our data show that patients were satisfied with the TMH ATP, and IOP, with most reporting that they would recommend these services to a friend or family member. When asked to describe their preference, most patients preferred to continue the TMH format during the pandemic and beyond. These results demonstrate that a "hybrid" model of care, which allows for both approaches (depending upon the patient's choice and availability of stable internet services in their area) may be a viable alternative. Common technological difficulties experienced by patients included slow or unstable internet connections, malfunctioning cameras, and log-in difficulties. However, for most patients, these technological difficulties did not negatively affect their experience with the program. TMH services are important in reaching patients that are geographically distanced from mental health facilities. It is important to recognize that the infrastructure for stable internet connections within communities and access to devices that can facilitate this type of treatment play a role in who can access TMH.

Content analyses of qualitative data suggest that the patients were willing to effectively address technological problems in the spirit of accessing convenient, in-home services that reduce the risk of health care-associated infections during the COVID-19 pandemic. Further, patients noted that the TMH format facilitated the acquisition of evidence-based coping skills and engendered a sense of social connection despite ongoing social and physical distancing measures. These findings suggest that TMH IOPs are sustainable and acceptable to adults with SMI. Moreover, mental health systems should consider offering both TMH and traditional in-person services to best meet the needs of patients with diverse preferences, technologic capabilities, and learning needs regardless of the state of the pandemic.

The lack of patient-identified quality improvement recommendations is likely due to the high degree of satisfaction reported by the overall sample. Start-point recommendations offered by respondents included expanding platform features (eg, using the virtual whiteboard), improving the use of program handouts (eg, sending documents virtually) and maintaining the availability of in-person IOPs for those who prefer face-to-face treatment.

Limitations

This study used the data gathered through convenience sampling, which limits the generalizability of our findings to other populations. Although TMH IOPs may be helpful for a large proportion of adults with SMI, not all clinics or programs may be prepared to provide such services. This study was performed at a large clinical and academic center with previous experience with telehealth programming. There was also significant administrative and information technology support available, which limits the generalizability of our findings to other clinics. Additionally, to determine patient satisfaction, we used selected items from established measures of acceptability of interventions, which may have influenced internal consistency. Furthermore, the findings may contain positive bias given that not all patients completed the satisfaction survey. Lastly, our sample lacked a comparison, in-person group, and was limited in terms of racial and ethnic diversity. This sample was also limited to those patients who had sufficient technologic knowledge, skills, and resources (eg, high-speed internet, smartphone, and computer) to engage in the TMH platform. Subsequent research should aim to report TMH IOP outcome data, ideally across a broader range of patient characteristics. Despite these limitations, the findings detailed here reinforce the benefits of delivering TMH IOPs during public health emergencies and contribute to the sparse literature available on real-world program adaptations.

Conclusions

The COVID-19 pandemic led to the rapid adoption of TMH services across mental health systems. Our findings indicate that TMH IOPs are feasible and can be an effective, safe, and convenient treatment framework for adults with SMI. High satisfaction with TMH IOP delivery and content can be achieved without compromising ongoing social and physical distancing measures. Additional research is needed to assess the efficacy of TMH IOPs in treating mental health concerns.

Acknowledgments

This research would not be feasible without the great work of dedicated multidisciplinary treatment team members at the Adult Transitions Program.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Satisfaction survey.
[DOCX File , 16 KB - mental_v9i1e30204_app1.docx]

References

1. The impact of COVID-19 on mental, neurological and substance use services: results of a rapid assessment. World Health Organization. URL: https://www.who.int/publications/i/item/978924012455 [accessed 2022-01-24]
2. Rajkumar RP. COVID-19 and mental health: A review of the existing literature. Asian J Psychiatr 2020 Aug;52:102066 [FREE Full text] [doi: 10.1016/j.ajp.2020.102066] [Medline: 32302935]
3. Gunnell D, Appleby L, Arensman E, Hawton K, John A, Kapur N, COVID-19 Suicide Prevention Research Collaboration. Suicide risk and prevention during the COVID-19 pandemic. Lancet Psychiatry 2020 Jun;7(6):468-471 [FREE Full text] [doi: 10.1016/S2215-0366(20)30171-1] [Medline: 32330430]
4. Hilty DM, Ferrer DC, Parish MB, Johnston B, Callahan EJ, Yellowlees PM. The effectiveness of telemental health: a 2013 review. Telemed J E Health 2013 Jun;19(6):444-454 [FREE Full text] [doi: 10.1089/tmj.2013.0075] [Medline: 23697504]
5. Gentry MT, Lapid MI, Clark MM, Rummans TA. Evidence for telehealth group-based treatment: A systematic review. J Telemed Telecare 2019 Jul;25(6):327-342. [doi: 10.1177/1357633X18775855] [Medline: 29788807]
6. Shore JH. Telepsychiatry: videoconferencing in the delivery of psychiatric care. Am J Psychiatry 2013 Mar;170(3):256-262. [doi: 10.1176/appi.ajp.2012.12081064] [Medline: 23450286]
7. Bashshur RL, Shannon GW, Bashshur N, Yellowlees PM. The Empirical Evidence for Telemedicine Interventions in Mental Disorders. Telemed J E Health 2016 Feb;22(2):87-113 [FREE Full text] [doi: 10.1089/tmj.2015.0206] [Medline: 26624248]
8. Breitinger S, Gentry MT, Hilty DM. Key Opportunities for the COVID-19 Response to Create a Path to Sustainable Telemedicine Services. Mayo Clin Proc 2020 Dec;95(12):2602-2605 [FREE Full text] [doi: 10.1016/j.mayocp.2020.09.034] [Medline: 33276833]
9. Connolly SL, Stolzmann KL, Heyworth L, Weaver KR, Bauer MS, Miller CJ. Rapid Increase in Telemental Health Within the Department of Veterans Affairs During the COVID-19 Pandemic. Telemed J E Health 2021 Apr;27(4):454-458. [doi: 10.1089/tmj.2020.0233] [Medline: 32926664]
10. Mishkind MC, Shore JH, Bishop K, D'Amato K, Brame A, Thomas M, et al. Rapid Conversion to Telemental Health Services in Response to COVID-19: Experiences of Two Outpatient Mental Health Clinics. Telemed J E Health 2021 Jul 28;27(7):778-784. [doi: 10.1089/tmj.2020.0304] [Medline: 33393857]
11. Ackerman M, Greenwald E, Noulas P, Ahn C. Patient Satisfaction with and Use of Telemental Health Services in the Perinatal Period: a Survey Study. Psychiatr Q 2021 Sep;92(3):925-933 [FREE Full text] [doi: 10.1007/s11126-020-09874-8] [Medline: 33389477]
12. Jenkins-Guarnieri MA, Pruitt LD, Luxton DD, Johnson K. Patient Perceptions of Telemental Health: Systematic Review of Direct Comparisons to In-Person Psychotherapeutic Treatments. Telemed J E Health 2015 Aug;21(8):652-660. [doi: 10.1089/tmj.2014.0165] [Medline: 25885491]
13. Ramaswamy A, Yu M, Drangsholt S, Ng E, Culligan PJ, Schlegel PN, et al. Patient Satisfaction With Telemedicine During the COVID-19 Pandemic: Retrospective Cohort Study. J Med Internet Res 2020 Sep 09;22(9):e20786 [FREE Full text] [doi: 10.2196/20786] [Medline: 32810841]
14. Isautier JM, Copp T, Ayre J, Cvejic E, Meyerowitz-Katz G, Batcup C, et al. People's Experiences and Satisfaction With Telehealth During the COVID-19 Pandemic in Australia: Cross-Sectional Survey Study. J Med Internet Res 2020 Dec 10;22(12):e24531 [FREE Full text] [doi: 10.2196/24531] [Medline: 33156806]
15. Kanter JW, Busch AM, Rusch LC. Behavioral Activation: Distinctive Features. New York, US: Routledge; 2009.
16. Linehan MM. DBT Skills Training Manual, Second Edition. New York, US: The Guilford Press; 2014.

17. Hayes SC. Get Out of Your Mind and Into Your Life: The New Acceptance and Commitment Therapy. Oakland, California, US: New Harbinger Publications; 2005.
18. Weiner BJ, Lewis CC, Stanick C, Powell BJ, Dorsey CN, Clary AS, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug 29;12(1):108 [FREE Full text] [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
19. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]

Abbreviations

ATP: adult transitions program
DBT: dialectical behavioral therapy
IOP: intensive outpatient program
SMI: serious mental illness
TMH: telemental health

Edited by J Torous; submitted 05.05.21; peer-reviewed by P Yellowlees, J Chong; comments to author 08.06.21; revised version received 19.11.21; accepted 02.12.21; published 28.01.22.

Please cite as:
Skime MK, Puspitasari AJ, Gentry MT, Heredia Jr D, Sawchuk CN, Moore WR, Taylor-Desir MJ, Schak KM
Patient Satisfaction and Recommendations for Delivering a Group-Based Intensive Outpatient Program via Telemental Health During the COVID-19 Pandemic: Cross-sectional Cohort Study
JMIR Ment Health 2022;9(1):e30204
URL: https://mental.jmir.org/2022/1/e30204
doi:10.2196/30204
PMID:34878999

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org