Original Paper

Accuracy of a Chatbot (Ada) in the Diagnosis of Mental Disorders: Comparative Case Study With Lay and Expert Users

Stefanie Maria Jungmann[1], PhD; Timo Klan[1], PhD; Sebastian Kuhn[2], MD; Florian Jungmann[2], MD

[1]Department of Psychology, Johannes Gutenberg-University Mainz, Mainz, Germany

[2]University Medical Center, Johannes Gutenberg-University Mainz, Mainz, Germany

Corresponding Author:
Stefanie Maria Jungmann, PhD
Department of Psychology, Johannes Gutenberg-University Mainz
Wallstraße 3
Mainz, 55122
Germany
Phone: 49 61313939201
Email: jungmann@uni-mainz.de

Abstract

Background: Health apps for the screening and diagnosis of mental disorders have emerged in recent years on various levels (eg, patients, practitioners, and public health system). However, the diagnostic quality of these apps has not been (sufficiently) tested so far.

Objective: The objective of this pilot study was to investigate the diagnostic quality of a health app for a broad spectrum of mental disorders and its dependency on expert knowledge.

Methods: Two psychotherapists, two psychology students, and two laypersons each read 20 case vignettes with a broad spectrum of mental disorders. They used a health app (Ada—Your Health Guide) to get a diagnosis by entering the symptoms. Interrater reliabilities were computed between the diagnoses of the case vignettes and the results of the app for each user group.

Results: Overall, there was a moderate diagnostic agreement (kappa=0.64) between the results of the app and the case vignettes for mental disorders in adulthood and a low diagnostic agreement (kappa=0.40) for mental disorders in childhood and adolescence. When psychotherapists applied the app, there was a good diagnostic agreement (kappa=0.78) regarding mental disorders in adulthood. The diagnostic agreement was moderate (kappa=0.55/0.60) for students and laypersons. For mental disorders in childhood and adolescence, a moderate diagnostic quality was found when psychotherapists (kappa=0.53) and students (kappa=0.41) used the app, whereas the quality was low for laypersons (kappa=0.29). On average, the app required 34 questions to be answered and 7 min to complete.

Conclusions: The health app investigated here can represent an efficient diagnostic screening or help function for mental disorders in adulthood and has the potential to support especially diagnosticians in their work in various ways. The results of this pilot study provide a first indication that the diagnostic accuracy is user dependent and improvements in the app are needed especially for mental disorders in childhood and adolescence.

(JMIR Form Res 2019;3(4):e13863) doi: 10.2196/13863

KEYWORDS

artificial intelligence; eHealth; mental disorders; mHealth; screening; (mobile) app; diagnostic

Introduction

Background

Digital media have become enormously important in the health sector. Up to 80% of the internet users inform themselves on the Web about health [1], and about 60% of patients search for their symptoms on the internet before or after a visit to the doctor [2]. Experts estimate that there are over 380,000 health-related mobile apps worldwide [3].

Health apps play an important role not only in physical diseases but also particularly in mental health conditions and disorders [4-6]. For mental disorders, access to professional diagnosis and treatment is often difficult and delayed (eg, long waits and concerns about psychotherapy). In addition, there is considerable uncertainty in the population about the significance of the

symptoms (eg, at what point feelings and behaviors are pathological). The advantage of health apps is low-threshold, locally and temporally flexible, and cost-efficient access [4]. The services are independent of the medical care situation, can be individually adapted and integrated into everyday life, and increase the self-help potential [5,7]. A systematic literature review showed that especially people who felt stigmatized by their problem or ashamed of it (eg, encopresis and eating disorders) use electronic mental health (e–mental health) [8]. Digital media can also be highly relevant for certain target groups. In mental disorders in childhood, for example, there are more possibilities for nonverbal recording of symptoms, and parents can be supported in coping with problems in everyday life [9]. As young people use new media every day (97% daily internet consumption), and mental health problems at this age are usually experienced as stigmatizing and shameful, the youth are considered particularly accessible to health apps [10]. In addition, health apps are promising for patients with chronic or recurrent phases of illness, which are particularly common in mental disorders.

In recent years, an enormous number of health apps have been developed for mental health conditions and disorders, the number of which is now hardly manageable. The proportion of health apps for mental health is about 29% of all health apps worldwide [11]. Health apps for mental health cover various areas of health promotion, prevention, screening and diagnostics, management, treatment, and aftercare [12]. These apps are usually aimed at consumers, that is, people suffering from symptoms. Recent developments also target professionals and, more recently, the public health care system (eg, pilot function and screening) [13,14].

Given the large number of health apps, the problem arises that they are used extensively but are usually not (sufficiently) evaluated and tested. Several reviews [15-18] have found that health apps for mental health have rarely been tested for their usefulness and effectiveness and often have ethical and legal shortcomings (eg, data privacy and safety). For example, Wisniewski et al [15] found that 15% to 45% of studied apps for anxiety, depression, and schizophrenia made medical claims, although these were rarely evidence-based, and no apps had Food and Drug Administration marketing approval. In addition, only 50% to 85% included a privacy policy [15]. Even if the apps have many benefits as described above, health-related internet use can also have negative or harmful effects on one's emotional state and health behavior, as research shows, for example, on the phenomenon of cyberchondria [19]. Cyberchondria refers to an excessive health-related internet search resulting in an increase in emotional distress and health anxiety (eg, because of ambiguous information or serious disease [19]).

There is a particularly great need for research into apps for the screening or diagnosis of mental disorders [5]. This gap in research contrasts with the importance that diagnostic or screening tools can have, for example, in assigning patients to appropriate medical disciplines and practitioners.

Health Apps for Screening and Diagnosis of Mental Disorders

Regarding diagnostics using e–mental health, a distinction is made between the collection of objective and subjective data [5,20]. Objective data (mostly psychophysiological measures or behavioral activity) are recorded via sensors in or connected to the mobile phone or so-called wearables. For example, Valenza et al [21] showed that heart rate variability predicted mood swings in patients with a bipolar spectrum disorder. So far, there are few empirical findings on the use of wearables in mental disorders; only about 1.5% of studies on wearables deal with mental health [22]. A recent systematic review showed that objective data were promising in predicting moods and mood changes, but much more empirical evidence was needed to reliably evaluate potentials and risks [20].

There are countless health apps that assess subjective data, such as apps used for assessments (eg, Web-based questionnaires) or tracking (eg, monitoring mood or medication via diaries) [23]. Regarding self-report instruments that were adapted into a mobile phone app, there are few evaluated Web-based questionnaires on depression and posttraumatic stress disorder that showed a psychometric quality comparable with the paper-pencil version [23-25]. Some apps, such as Moodpath [26], include questions based on the operationalized diagnostic criteria of the International Classification of Diseases (ICD), tenth revision [27]. In Moodpath, users are asked different questions 3 times a day for 14 days according to the diagnostic criteria for depressive disorders. On the basis of the indicated symptom patterns, an algorithm determines possible depression (screening) and makes an assessment of severity. The results of diagnostic apps are often based on algorithms or artificial intelligence (AI), which means that computers can simulate complex human cognitions and actions.

Regarding mental tracking, a few apps on mood and affective disorders have been empirically investigated. For example, Hung et al [28] found in patients with depression that daily data on depression, anxiety, and sleep quality in a mobile phone app were significantly related to clinician-administered depression assessment at baseline. For bipolar affective disorder, a mobile phone app identified lower physical (location changes recorded via global positioning system) and social (outgoing messages) activities as significant predictors for increased depressive symptoms and lower physical but increased social activity for increased manic symptoms [29].

In contrast to apps for physical diseases (eg, Ada—Your Health Guide [30] and IBM Watson Health [31]), apps for mental health focus almost exclusively on a single symptom or single mental disorder, rather than on a broader spectrum. However, especially for the purpose of screening, it seems interesting and necessary at all 3 levels (eg, individual, practitioner, and public health system) that a single app asks for a variety of symptoms and mental disorders and provides information about the range of psychopathology. Only a few apps for mental health, such as WhatsMyM3 [32] (anxiety, depression, bipolar affective disorder, and posttraumatic stress) and T2 Mood Tracker [33] (anxiety, depression, head injury, and posttraumatic stress), assess multiple mental health conditions. However, these are

usually limited to anxiety-depressive symptoms and have so far been little evaluated [23]. Therefore, in this study we used a medicine app that covers a wide range of physical and mental health conditions.

The aim of this pilot study was to test for the first time the diagnostic agreement of a medicine app and case vignettes over a broad spectrum of mental disorders. We expected at least moderate diagnostic agreement (ie, interrater reliability Cohen kappa≥0.41; hypothesis 1). As health apps are used both as a self-assessment at the consumer level and a diagnostic support system by experts and practitioners [34,35], we examined the diagnostic quality, depending on the user's level of expert knowledge (ie, 3 user groups: psychotherapists, psychology students, and laypersons). Given the less advanced state of development of diagnostic health apps for mental health than for physical diseases [5,36], we hypothesized that diagnostic accuracy for mental disorders is dependent on expert knowledge (eg, symptom checker includes fewer psychiatric terms, and alternative terms need to be entered; hypothesis 2).

Methods

Design

A health app (Ada—Your Health Guide [30]) was used to diagnose 20 case vignettes from well-known textbooks of psychiatry and clinical psychology [37-40] by 3 groups: psychological psychotherapists, psychology students, and persons from the general population without previous professional knowledge of mental disorders (laypersons). Figure 1 illustrates the design and method.

Participants

Table 1 shows the sociodemographic characteristics of the participants.

Figure 1. Method and procedure of the study. ADHD: attention-deficit hyperactivity disorder.

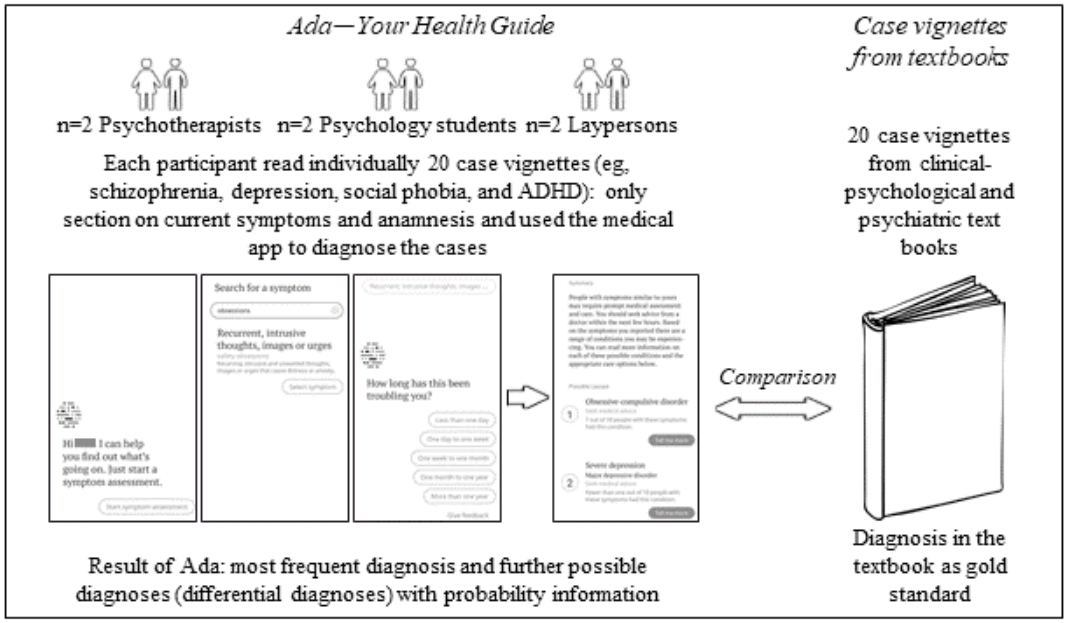

Table 1. Sociodemographic characteristics of the participants, subdivided into psychotherapists, psychology students, and laypersons.

Characteristics	Psychotherapists (n=2)	Psychology students (n=2)	Laypersons (n=2)	Statistics	
				P value	Effect size
Age (years), mean SD	40.0 (11.3)	22.0 (4.2)	40.0 (8.5)	.19	η^2_p=0.67
Sex, female, n (%)	1 (50)	1 (50)	1 (50)	>.99	Φ=0
Occupation, mean (SD) or with or without vocational training	14.0 (7.1) years professional experience[a]	4.0 (4.2) number of semesters	n=1 data scientist; n=1 without vocational training	—[b]	—[b]

[a]One of both is additionally a child and adolescent psychotherapist (first author).

[b]No comparative statistics possible due to different occupations.

Instruments

Health App for Diagnosis

Ada—Your Health Guide [30] is a Conformité Européenne–certified (ensures safe products within the European economic area) health app for the screening and diagnostic support of health conditions, primarily for physical diseases but increasingly also for mental health conditions and disorders. This app can be used both at the consumer level as a self-assessment app and by experts and practitioners as a diagnostic decision support system [35]. On the basis of AI, the

chatbot asks for existing complaints adaptively and analogously to a medical or psychotherapeutic anamnesis interview. The Ada chatbot [30] is based on a medical database with constantly updated research findings. As a result, the diagnosis is determined that best matches the pattern of symptoms entered. The user is given a probability of a possible diagnosis and, as differential diagnoses, other less probable diagnoses (eg, 8/10 people with the symptoms described suffer from a depressive disorder). Patients or relatives (eg, parents and caregivers) receive an assessment of the urgency of seeking medical advice. Ada—Your Health Guide [35] was selected to investigate the above research questions for the following reasons (also in comparison with other symptom checkers [41,42]): (1) in addition to somatic symptoms, the app considers a wide spectrum of mental health conditions; (2) the app provides probabilities of possible and differential diagnoses (indications of comorbidities); (3) the app is widespread (>5 million users in >130 countries), publicly available, and free; (4) available in different languages (including English and German), and (5) in comparison with other symptom checkers (eg, Your.MD [43] and Babylon Health [44]), Ada provided more accurate diagnoses [42,45].

Knowledge of Mental Disorders

The user's knowledge of mental disorders was assessed on a 5-point Likert scale (1=not at all to 5=very good).

Procedure

After the informed consent, participants were instructed to carefully read the case vignette and then use the app to determine a diagnosis. A total of 20 case vignettes from psychiatry and clinical psychology textbooks were used, with 12 cases from adulthood [37,39] and 8 cases from childhood and adolescence [38,40]. All participants worked on one case after the other. The case vignettes were selected in such a way that a broad spectrum of mental disorders could be examined (see a list of mental disorders in Multimedia Appendix 1). The case vignettes included the initial symptoms before treatment (reason to seek treatment) and the anamnestic information, without naming or citing the diagnosis. The participants worked on the case with the health app on a tablet. The study duration was 3 to 6 hours per participant, divided into 2 to 3 individual sessions (most of the time was spent reading the 20 case vignettes). The participants (except the psychotherapists) received financial compensation (€10/hour) or a course credit (students).

Data Analysis

The main outcome was the agreement between the main diagnosis of the case vignette in the textbook and the result given by the app (the most probable diagnosis). Consistent labeling of the mental disorders was considered when assessing agreement. As an exception, the terms abuse and addiction were judged to agree, as the app did not distinguish between abuse and addiction to our knowledge. The diagnoses were compared at the level of 4-digit codes in the ICD (eg, anxiety disorders such as social anxiety and agoraphobia or personality disorders such as borderline personality disorder). If the 4th digit represents a more detailed specification (eg, obsessive-compulsive disorder: predominantly obsessive-compulsive behavior and thoughts or severity of the depressive episode), the 3-digit code match was counted (for the following disorders: depressive disorder, bipolar affective disorder, obsessive-compulsive disorder, conduct disorder, or schizophrenia). To consider the function and purpose of the screening and diagnostic app (eg, further diagnostic procedures required), no distinction was made between the subtypes of dementia (eg, Alzheimer and vascular dementia) and that of urinary incontinence (eg, stress incontinence and enuresis diurnal or nocturnal). The list of diagnoses in the textbooks and the results from the app can be found in Multimedia Appendix 1. The statistical outcomes were calculated as the percentage of agreement and the Cohen kappa coefficient (interrater reliability) for controlling random agreements. According to Landis and Koch [46], kappa values between 0.41 and 0.60 can be rated as moderate, between 0.61 and 0.80 as good, and >0.81 as very good. The agreement was checked if the secondary or differential diagnosis given by the app was also included (eg, bipolar disorder in the textbook but as a differential diagnosis in the app). Cohen *d* was calculated as effect size for group differences and partial eta-square for variance analyses. All statistical analyses were conducted using SPSS version 23 (IBM SPSS) [47], with an alpha level of .05. Following the study by Field [48], the Ryan, Einot, Gabriel, and Welsch Q procedure was used in post hoc tests to control the alpha error (same sample size; the Gabriel procedure was used when the sample sizes were different).

Results

Knowledge of Mental Disorders

Self-rated knowledge of mental disorders varied significantly depending on the group (ie, psychotherapists, students, or laypersons)—$F_{2,3}$=18.50; P=.02; partial eta-square=0.93. Post hoc analyses indicated that laypersons (mean 1.50, SD 0.71) reported significantly lower knowledge than students (mean 3.50, SD 0.71; P=.04) and psychotherapists (mean 5.00, SD 0; P=.01), with the last 2 groups having a marginally significant difference from each other (P=.08).

Percentage Agreement and Interrater Reliability

For mental disorders in adulthood, we found for the 72 case records (6 users×12 mental disorders), a percentage agreement of 68% and an interrater reliability according to Cohen kappa 0.64 between the textbook diagnosis and the result produced by the app. Taking into account the differential diagnoses, we found a percentage agreement of 85% and Cohen kappa 0.82. For mental disorders in childhood and adolescence, 48 case records (6 users×8 mental disorders) showed a percentage agreement of 42% (including differential diagnoses: 56%) and a Cohen kappa 0.40 (including differential diagnoses: kappa=0.52).

Table 2 shows the mean number (n), percentage (%), and Cohen kappa coefficients, differentiated among the 3 different user groups (ie, psychotherapists, students, and laypersons).

For mental disorders in adulthood, the Cohen kappa values were 0.78 (95% CI 0.60-0.95) for psychotherapists, 0.55 (95% CI 0.35-0.76) for students, and 0.60 (95% CI 0.39-0.80) for

laypersons. Regarding case vignettes from childhood and adolescence, Cohen kappa values were numerically higher for psychotherapists (kappa=0.53, 95% CI 0.28-0.77) than for students (kappa=0.41, 95% CI 0.18-0.63) and laypersons (kappa=0.29, 95% CI 0.08-0.49).

Multimedia Appendix 1 lists the 20 mental disorders of the case vignettes as well as the main diagnoses in Ada Health and examples of differential diagnoses. The app mostly identified the main diagnosis (67% [8/12] of cases for adulthood and 44% [3.5/8] of cases for childhood and adolescence); it reported the differential diagnoses in an additional 17% (2/12) of cases for adulthood and 13% (1/8) of cases for childhood and adolescence. If the differential diagnoses are included, all diagnoses except undifferentiated somatization disorder, separation anxiety, and selective mutism in childhood were correctly detected.

Table 2. Mean number, percentage, and Cohen kappa coefficients for agreement between the textbook diagnosis and the result from Ada Health.

Case reports	Main diagnosis in Ada Health						Additional consideration of differential diagnoses in Ada Health					
	Psychotherapists		Students		Laypersons		Psychotherapists		Students		Laypersons	
	n (%)	kappa	n (%)	kappa	n (%)	kappa	n (%)	kappa	n (%)	kappa	n (%)	kappa
Adulthood (n_{max}=12)	9.5 (79)	0.78	7 (58)	0.55	7.5 (63)	0.60	11 (92)	0.91	10.5 (88)	0.87	8.5 (71)	0.69
Childhood and adolescence (n_{max}=8)	4.5 (56)	0.53	3.5 (44)	0.41	2.5 (31)	0.29	4.5 (56)	0.59	5 (63)	0.52	4 (40)	0.45

Number of Questions and Duration

To find a solution, the app had to ask an average of 34 questions per case (mean 33.78, SD 8.73) about the type and duration of the symptoms. There was no significant difference between the groups ($F_{2,117}$=1.89; P=.16; partial eta-square=0.03). The average time to complete was 409 seconds (SD 141.23). The groups differed in the average time for completion ($F_{2,96}$=9.93; P<.001; partial eta-square=0.17). Psychotherapists (mean 457.28, SD 138.61) and students (mean 415.82, SD 143.11), who did not differ from each other (P=.40), showed a significantly longer time for completion than the laypersons (the time recorded for only 1 layperson; mean 299.45, SD 141.23; P<.001).

Discussion

Principal Findings

In this pilot study, we tested whether a health app (Ada—Your Health Guide [30]) could detect mental disorders in children, adolescents, and adults. A total of 3 groups of users (ie, psychotherapists, psychology students, and laypersons) used the app to diagnose 20 case vignettes. Across all users, the agreement between the textbook diagnoses and the app was moderate (kappa=0.64) for mental disorders in adulthood and low (kappa=0.40) for that in childhood and adolescence. Adding differential diagnoses, good (kappa=0.82) and moderate (kappa=0.52) values, respectively, were obtained for interrater reliability.

When psychotherapists applied the app, there was a good agreement (kappa=0.78) between the results of the app and the diagnoses in the textbook on mental disorders in adulthood. This value is comparable with interrater reliabilities between 2 psychologists for diagnoses assessed with structured clinical interviews (kappa=0.71 for Axis I disorders and kappa=0.84 for personality disorders [49]). The diagnostic agreement was moderate (kappa=0.55/0.60) when students and laypersons used the app. The addition of differential diagnoses showed a good to very good interrater reliability (kappa=0.69-0.91). In 17% of the cases, the app did not give the diagnosis as the main diagnosis but as a differential diagnosis. Although the app assessed a different diagnosis as more likely, the main diagnosis of the case report was considered in some cases as a differential diagnosis.

For mental disorders in childhood and adolescence, a moderate diagnostic quality was found when psychotherapists (kappa=0.53) and students (kappa=0.41) used the app, whereas the quality was low for laypersons (kappa=0.29). In contrast to mental disorders in adulthood, the addition of differential diagnoses improved the diagnostic quality in childhood and adolescence to a lesser extent.

Taken together, only for mental disorders in adulthood, and when psychotherapists used the app, did Ada—Your Health Guide show good diagnostic quality. The app can serve as an indication of a mental health problem in the range of moderate agreement (adult mental disorders: students and laypersons; child and adolescent mental disorders: psychotherapists, students). With an average app time of 7 min, the app can be an efficient tool for the initial evaluation and screening of mental health problems and disorders. So, this pilot study indicates that expert knowledge tends to lead to better diagnostic quality when using the health app.

When comparing mental disorders in adulthood and childhood and adolescence, the app shows deficits for mental disorders in children and adolescents. For example, the app could not detect separation anxiety in childhood or selective mutism in any operation. On the one hand, this may be because of deficits in the app, on the other, mental disorders in childhood and adolescence are more often characterized by less specific symptom descriptions—children and adolescents show fewer specific symptoms and, from a developmental perspective, more frequent temporary subclinical symptoms [50]. This may also have led to confusion with the concrete naming and focusing of symptoms in childhood and adolescence. Examples include case reports on attention-deficit hyperactivity disorder (ADHD)

and separation anxiety. In the ADHD case vignette, fears are mentioned first (eg, *would see ghosts*) [40]. In the case of separation anxiety, the initial focus is on describing the problematic relationship of the parents. In both cases, the hallmarks of the disorders are reported later and relatively profoundly. In addition, the app [30] may not include relevant terms and psychopathological characteristics, such as school fear and selective mutism. There is a clear need to catch up here. Especially in the case of enuresis, the results generated by the app, such as *mixed incontinence* or *stress incontinence*, made it clear that these were primarily terms pertaining to adults. As Ada—Your Health Guide [30] is based on a medical database with updated research findings, these deficits in the detection of mental disorders can also be because research activity in children and adolescents is significantly lower than that in adults. In the case of disorders with somatic symptoms (eg, undifferentiated somatization disorder), the diagnosis was more difficult because of the delimitation of psychological and physical symptoms. The overall interrater reliability in this study is lower than in studies that use structured clinical interviews [49].

It is important to consider the aims of screening and diagnostic apps. Health apps (eg, Ada—Your Health Guide [30]) do not aim to replace doctors or psychotherapists. Psychopathological symptoms can only be adequately understood and classified by a detailed anamnesis, the consideration of the temporal course, and the correct assessment of inclusion and exclusion criteria. For example, a severe, recurrent depressive disorder or multiple comorbidities worsen prognosis and require treatment (eg, combined treatment with psychotropic drugs) different from more circumscribed cases, such as a mild and single depressive episode. To our knowledge, there is currently no diagnostic app that captures this complexity (especially several comorbidities). Furthermore, the benefits of personal interaction should not be underestimated, as some behavioral abnormalities become apparent especially in direct contact (eg, hyperactivity or personality disorders), and unintended or intentional bias tendencies (eg, social desirability) can be more easily identified. Therefore, we consider the clarification of problems and diagnostics by experts to be of immense importance. The evaluated diagnostic apps should rather be regarded as low-cost, low-threshold, and time-efficient support in the diagnosis of mental disorders in adulthood [5]. There is great potential for the application of AI-supported diagnostics at the level of the consumer or patient, the experts, and the health care system, for example the following [14]:

- *Consumers and patients:* for example, screening of symptoms, combined with possible emotional relief for the affected person (eg, diagnosis as an explanation or treatment option) and a recommendation for action (eg, seeking medical advice).
- *Professionals:* for example, support in more efficient exploration and diagnosis (eg, bringing the result of the health app to the initial consultation), consideration and explanation of differential diagnoses, rapid reaction to significant symptoms (eg, suicidal intentions and alcohol consumption), and support in making indication decisions.
- *Macro/health care system:* for example, optimizing the assignment to treatment providers or treatment settings, supporting employees of other occupational groups in the health care system.

Limitations and Research Perspectives

In this study, the health app was only tested on case vignettes, and the user groups had a very small sample size. This limits the transferability of our results to everyday practice (low ecological validity). In addition, in the case of small samples, the performance of individual and outlier values plays a major role [51]. A recent study examined another symptom checker (Babylon Health [44]) that had comparable methodological limitations (case vignettes and small sample [52]). In contrast to this study, we investigated mental disorders for which the apps have so far been little developed, requiring a first pilot study. In addition, we focused on the question of whether the diagnostic quality is dependent on expert knowledge and examined the quality when experts, students, and laypersons used the app.

A next step will be to investigate the diagnostic accuracy of health apps for mental disorders in a direct interaction of practitioner and patient and with a larger sample. Depending on the research question, the design has to be differentiated. If the diagnostic quality is of interest, the agreement of the results of the app applied by the end user or patient could be compared with the current gold standard for the diagnosis of mental disorders, that is, structured or standardized interviews (eg, Diagnostic Interview for Mental Disorders [43]). If investigating the question of how well the health app can support clinicians in diagnosing mental disorders, the comparison of the clinical diagnosis with and without an additional health app should be examined. It should also be noted that the present design could not determine a match for *no diagnosis present* as the case vignettes always included a diagnosis. In a future naturalistic study with patients, this limitation would be removed.

Health apps are considered to be a support system rather than a substitute for doctors and psychotherapists, both by development companies and by doctors [53] and psychotherapists [5,54]. For example, a recent study [54] interviewed 720 general practitioners about future digitization in the health care system. Of them, 68% considered it unlikely that doctors would ever be replaced for diagnostic tasks. Previous findings on the appropriateness of the recommendation for further treatment vary between 33% [41] and 81% [55] agreement regarding the triage performance of the app and doctors or nurses, depending on, for example, the app used, the urgency of the treatment, and the judging person (doctor or nurse).

Combined with future research to test diagnostic accuracy, it would also be interesting to compare the extent to which differences exist when patients do the input themselves. As already mentioned, there is a clear need to catch up in the field of diagnostics in childhood and adolescence using the app tested here. Parents are often uncertain about the significance of existing symptoms, behavioral abnormalities, or developmental deficits. Even if electronic health systems are to be understood as diagnostic indications or screenings and not something that

can replace a doctor or psychotherapist, such a system can provide parents with relevant information and initial instructions for action.

As the app is used particularly at the consumer level, and our pilot study indicated that diagnostic quality was lower among users from the general population and students, an important research perspective is to examine in which areas the weaknesses and deficits lie with nonprofessionals and how these can be addressed in further development. Such development could also be valuable, for example, for use in regions or countries with limited medical and psychotherapeutic care. The professional level would also benefit from a higher reliability of AI-supported diagnosis of mental disorders in childhood and adolescence. The fact that a patient is referred to an appropriate medical or psychotherapeutic specialty, for example, has relevant effects on the patient and the physician and can have considerable health economic effects.

As health apps collect and process highly sensitive health data, data security is of immense importance. Frequent shortcomings of current health apps are inadequate information about the nature and purpose of further processing of the data, missing or excessively complex data privacy statements, and comparatively easy access and manipulation by third parties [6,18,56]. Health apps should increasingly be certified based on defined catalogues of criteria and provided with a seal of quality, although this has rarely been done to date [57]. Overall, challenges remain to improve data security and the standardization of quality assurance, in particular, transparency for users, data protection control, and the handling of big data [14,36,57].

Conclusions

Health-related apps are also widely used for mental health conditions and disorders (in the general population and increasingly by practitioners and the public health system), but little is known about the diagnostic quality of health apps for mental disorders. This pilot study found that the diagnostic agreement between the health app and the diagnosis of the case vignettes for mental disorders was overall low to moderate. The diagnostic quality was shown to be dependent on the user and the type of mental disorder. Only when psychotherapists used the app for mental disorders in adulthood, good diagnostic agreements were found. Therefore, the health app should be used with caution in the general population and should be considered as a first indication of possible mental health conditions. In particular, improvements in the app with regard to mental disorders in childhood and adolescence and further research are needed.

Acknowledgments

The authors would like to thank Louisa Wagner for her support in the recruitment of participants and data collection.

Authors' Contributions

SMJ, SK, and FJ were involved in the study concept and design. SMJ and TK used the app as psychotherapists and were involved in statistical analysis. SMJ, TK, SK, and FJ interpreted and discussed the results. SJ was the main author of the first version of the paper; TK, SK, and FJ completed it, and all authors agreed to the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of mental disorders.
[PDF File (Adobe PDF File), 96 KB-Multimedia Appendix 1]

References

1. Muse K, McManus F, Leung C, Meghreblian B, Williams JM. Cyberchondriasis: fact or fiction? A preliminary examination of the relationship between health anxiety and searching for health information on the Internet. J Anxiety Disord 2012 Jan;26(1):189-196. [doi: 10.1016/j.janxdis.2011.11.005] [Medline: 22137465]
2. Bertelsmann Stiftung. Gesundheitsinfos Wer suchet, der findet – Patienten mit Dr. Google zufrieden. [Health information Who searches, finds - patients are satisfied with Dr. Google.] URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/VV_SpotGes_Gesundheitsinfos_final.pdf [accessed 2019-02-11] [WebCite Cache ID 766oJS9BN]
3. Lucht M, Boeker M, Krame U. University of Freiburg. 2015. Gesundheits-und Versorgungs-Apps Hintergründe zu deren Entwicklung und Einsatz. [Health and care apps: Backgrounds to their development and use.] URL: https://www.uniklinik-freiburg.de/fileadmin/mediapool/09_zentren/studienzentrum/pdf/Studien/150331_TK-Gesamtbericht_Gesundheits-und_Versorgungs-Apps.pdf [accessed 2019-02-11] [WebCite Cache ID 766lwArcr]
4. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 1;3(1):e7 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]

5. Lüttke S, Hautzinger M, Fuhr K. E-Health in Diagnostik und Therapie psychischer Störungen: Werden Therapeuten bald überflüssig? [E-Health in diagnosis and therapy of mental disorders: Will therapists soon become superfluous?]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Mar;61(3):263-270. [doi: 10.1007/s00103-017-2684-9] [Medline: 29318339]
6. Magee JC, Adut S, Brazill K, Warnick S. Mobile app tools for identifying and managing mental health disorders in primary care. Curr Treat Options Psychiatry 2018 Sep;5(3):345-362 [FREE Full text] [doi: 10.1007/s40501-018-0154-0] [Medline: 30397577]
7. Löcherer R, Apolináro-Hagen J. Efficacy and acceptance of webbased programs for self-help for facilitation of psycological health and handling of stress during academic studies. Wirksamkeit und Akzeptanz von webbasierten Selbsthilfeprogrammen zur Förderung psychischer Gesundheit und zur Stressbewältigung im Studium: Ein Scoping-Review der aktuellen Forschungsliteratur. [Efficacy and acceptance of webbased programs for self-help for facilitation of psycological health and handling of stress during academic studies.] URL: https://www.e-beratungsjournal.net/ausgabe_0117/Loecherer_Apolinario-Hagen.pdf [accessed 2019-02-11] [WebCite Cache ID 766mIUNWj]
8. 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 Jun 23;8(2):e10 [FREE Full text] [doi: 10.2196/jmir.8.2.e10] [Medline: 16867965]
9. Döpfner M, Schürmann S, Lehmkuhl G. Wackelpeter und Trotzkopf. Hilfen für Eltern bei ADHS-Symptomen, hyperkinetischem und oppositionellem Verhalten. Mit Online-Materialien. 4., überarb. Aufl. [Jello-Pete and stubbornness: Help for Parents with ADHS-symptoms, hyperkinetic and opposing behavior. With online materials. 4th, revised edition.]. Weinheim, Basel: Beltz; 2011.
10. D'Alfonso S, Santesteban-Echarri O, Rice S, Wadley G, Lederman R, Miles C, et al. Artificial intelligence-assisted online social therapy for youth mental health. Front Psychol 2017;8:796 [FREE Full text] [doi: 10.3389/fpsyg.2017.00796] [Medline: 28626431]
11. Anthes E. Mental health: there's an app for that. Nature 2016 Apr 7;532(7597):20-23. [doi: 10.1038/532020a] [Medline: 27078548]
12. Riper H, Andersson G, Christensen H, Cuijpers P, Lange A, Eysenbach G. Theme issue on e-mental health: a growing field in internet research. J Med Internet Res 2010 Dec 19;12(5):e74 [FREE Full text] [doi: 10.2196/jmir.1713] [Medline: 21169177]
13. de Rosis S, Nuti S. Public strategies for improving eHealth integration and long-term sustainability in public health care systems: Findings from an Italian case study. Int J Health Plann Manage 2018 Jan;33(1):e131-e152 [FREE Full text] [doi: 10.1002/hpm.2443] [Medline: 28791771]
14. Kuhn S, MME, Jungmann S, Jungmann F. Deutsches Ärzteblatt International. Künstliche Intelligenz für Ärzte und Patienten: „Googeln“ war gestern. [Artificial intelligence for doctors and patients: 'Googling' was yesterday.] URL: https://www.aerzteblatt.de/archiv/198854 [accessed 2019-02-11] [WebCite Cache ID 766okKhVs]
15. Wisniewski H, Liu G, Henson P, Vaidyam A, Hajratalli NK, Onnela J, et al. Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. Evid Based Ment Health 2019 Feb;22(1):4-9. [doi: 10.1136/ebmental-2018-300069] [Medline: 30635262]
16. Torous J, Roberts LW. The ethical use of mobile health technology in clinical psychiatry. J Nerv Ment Dis 2017 Jan;205(1):4-8. [doi: 10.1097/NMD.0000000000000596] [Medline: 28005647]
17. Radovic A, Vona PL, Santostefano AM, Ciaravino S, Miller E, Stein BD. Smartphone applications for mental health. Cyberpsychol Behav Soc Netw 2016 Jul;19(7):465-470 [FREE Full text] [doi: 10.1089/cyber.2015.0619] [Medline: 27428034]
18. 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]
19. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord 2019 Feb 15;245:270-278. [doi: 10.1016/j.jad.2018.11.037] [Medline: 30419526]
20. Dogan E, Sander C, Wagner X, Hegerl U, Kohls E. Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review. J Med Internet Res 2017 Jul 24;19(7):e262 [FREE Full text] [doi: 10.2196/jmir.7006] [Medline: 28739561]
21. Valenza G, Nardelli M, Lanata' A, Gentili C, Bertschy G, Kosel M, et al. Predicting mood changes in bipolar disorder through heartbeat nonlinear dynamics. IEEE J Biomed Health Inform 2016 Jul;20(4):1034-1043. [doi: 10.1109/JBHI.2016.2554546] [Medline: 28113920]
22. Baig MM, GholamHosseini H, Moqeem AA, Mirza F, Lindén M. A systematic review of wearable patient monitoring systems - current challenges and opportunities for clinical adoption. J Med Syst 2017 Jul;41(7):115. [doi: 10.1007/s10916-017-0760-1] [Medline: 28631139]
23. van Ameringen M, Turna J, Khalesi Z, Pullia K, Patterson B. There is an app for that! The current state of mobile applications (apps) for DSM-5 obsessive-compulsive disorder, posttraumatic stress disorder, anxiety and mood disorders. Depress Anxiety 2017 Jun;34(6):526-539. [doi: 10.1002/da.22657] [Medline: 28569409]

24. Fann JR, Berry DL, Wolpin S, Austin-Seymour M, Bush N, Halpenny B, et al. Depression screening using the Patient Health Questionnaire-9 administered on a touch screen computer. Psychooncology 2009 Jan;18(1):14-22 [FREE Full text] [doi: 10.1002/pon.1368] [Medline: 18457335]
25. Bush NE, Skopp N, Smolenski D, Crumpton R, Fairall J. Behavioral screening measures delivered with a smartphone app: psychometric properties and user preference. J Nerv Ment Dis 2013 Nov;201(11):991-995. [doi: 10.1097/NMD.0000000000000039] [Medline: 24177488]
26. Goering M, Frauendorf F. Moodpath App. 2019. URL: https://mymoodpath.com/en/ [accessed 2019-02-11] [WebCite Cache ID 766oLDIU2]
27. World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnosis Guidelines. Geneva: World Health Organization; 1992.
28. Hung S, Li M, Chen Y, Chiang J, Chen Y, Hung GC. Smartphone-based ecological momentary assessment for Chinese patients with depression: an exploratory study in Taiwan. Asian J Psychiatr 2016 Oct;23:131-136. [doi: 10.1016/j.ajp.2016.08.003] [Medline: 27969071]
29. Beiwinkel T, Kindermann S, Maier A, Kerl C, Moock J, Barbian G, et al. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment Health 2016 Jan 6;3(1):e2 [FREE Full text] [doi: 10.2196/mental.4560] [Medline: 26740354]
30. Ada: Your health companion. URL: https://ada.com/ [accessed 2019-02-11] [WebCite Cache ID 766o3R2hN]
31. IBM. IBM Watson Health. URL: https://www.ibm.com/watson/health/ [accessed 2019-02-11] [WebCite Cache ID 766oiCB1k]
32. Hurowitz G, Post R. Whats My M3 - M3 Information. 2018. URL: https://whatsmym3.com/ [accessed 2019-02-11] [WebCite Cache ID 766oMNPro]
33. Military Health System. MHSRS 2019. URL: https://www.health.mil/Military-Health-Topics/Research-and-Innovation/MHSRS-2018 [accessed 2019-02-11] [WebCite Cache ID 766nvY0HA]
34. Albrecht U, Afshar K, Illiger K, Becker S, Hartz T, Breil B, et al. Expectancy, usage and acceptance by general practitioners and patients: exploratory results from a study in the German outpatient sector. Digit Health 2017;3:2055207617695135 [FREE Full text] [doi: 10.1177/2055207617695135] [Medline: 29942582]
35. Hoffmann H. ITU: Committed to connecting the world. Ada Health Our Approach to Assess Ada's Diagnostic Performance. URL: https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180925/Documents/3_Henry%20Hoffmann.pdf [accessed 2019-05-10]
36. Millenson ML, Baldwin JL, Zipperer L, Singh H. Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. Diagnosis (Berl) 2018 Sep 25;5(3):95-105. [doi: 10.1515/dx-2018-0009] [Medline: 30032130]
37. Stieglitz R, Baumann U, Perrez M. Fallbuch zur klinischen Psychologie und Psychotherapie. [Case Book on Clinical Psychology and Psychotherapy.]. Bern: Huber; 2007.
38. Petermann F. Fallbuch der klinischen Kinderpsychologie. [Casebook of Clinical Child Psychology.]. Göttingen: Hogrefe; 2009.
39. Freyberger H, Dilling H. Fallbuch Psychiatrie: Kasuistiken zum Kapitel V(F) der ICD-10. 2 Auflage. [Case Book Psychiatry: Casuistics on Chapter V(F) of the ICD-10. 2nd edition.] . Bern: Huber; 2014.
40. Poustka F, van Goor-Lambo G. Fallbuch Kinder- und Jugendpsychiatrie. Erfassung und Bewertung belastender Lebensumstände von Kindern nach Kapitel V (F) der ICD-10. [Case Book Child and Adolescent Psychiatry. Recording and evaluation of stressful life circumstances of children according to Chapter V (F) of the ICD-10.]. Göttingen: Hogrefe; 2008.
41. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. Br Med J 2015 Jul 8;351:h3480 [FREE Full text] [doi: 10.1136/bmj.h3480] [Medline: 26157077]
42. WIRED. Can You Really Trust the Medical Apps on Your Phone? URL: https://www.wired.co.uk/article/health-apps-test-ada-yourmd-babylon-accuracy [accessed 2019-05-10]
43. Your.MD - Health Guide and Symptom Checker. URL: https://www.your.md/ [accessed 2019-05-10]
44. Babylon Health. URL: https://www.babylonhealth.com/ [accessed 2019-05-10]
45. Armstrong S. The apps attempting to transfer NHS 111 online. Br Med J 2018 Jan 15;360:k156. [doi: 10.1136/bmj.k156] [Medline: 29335297]
46. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977 Jun;33(2):363-374. [doi: 10.2307/2529786] [Medline: 884196]
47. IBM. 2015. IBM SPSS Statistics. URL: https://www.ibm.com/in-en/products/spss-statistics [accessed 2019-10-03]
48. Field A. Discovering Statistics Using IBM SPSS Statistics. Fourth Edition. Thousand Oaks, California, United States: SAGE; 2013.
49. 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]

50. Merten EC, Cwik JC, Margraf J, Schneider S. Overdiagnosis of mental disorders in children and adolescents (in developed countries). Child Adolesc Psychiatry Ment Health 2017;11:5 [FREE Full text] [doi: 10.1186/s13034-016-0140-5] [Medline: 28105068]
51. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018 Nov 24;392(10161):2263-2264. [doi: 10.1016/S0140-6736(18)32819-8] [Medline: 30413281]
52. Razzaki S, Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, et al. arXiv. 2018. A Comparative Study of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosi. URL: https://marketing-assets.babylonhealth.com/press/BabylonJune2018Paper_Version1.4.2.pdf [accessed 2019-09-27]
53. WIRED. Stop Googling Your Symptoms – The Smartphone Doctor is Here to Help URL: https://www.wired.co.uk/article/ada-smartphone-doctor-nhs-gp-video-appointment [accessed 2019-05-10]
54. Blease C, Bernstein MH, Gaab J, Kaptchuk TJ, Kossowsky J, Mandl KD, et al. Computerization and the future of primary care: a survey of general practitioners in the UK. PLoS One 2018;13(12):e0207418 [FREE Full text] [doi: 10.1371/journal.pone.0207418] [Medline: 30540791]
55. Verzantvoort NC, Teunis T, Verheij TJ, van der Velden AW. Self-triage for acute primary care via a smartphone application: practical, safe and efficient? PLoS One 2018;13(6):e0199284 [FREE Full text] [doi: 10.1371/journal.pone.0199284] [Medline: 29944708]
56. Morera EP, Díez I, Garcia-Zapirain B, López-Coronado M, Arambarri J. Security recommendations for mHealth apps: elaboration of a developer's guide. J Med Syst 2016 Jun;40(6):152. [doi: 10.1007/s10916-016-0513-6] [Medline: 27147515]
57. Albrecht U, Hillebrand U, von Jan U. Relevance of trust marks and CE labels in German-language store descriptions of health apps: analysis. JMIR Mhealth Uhealth 2018 Apr 25;6(4):e10394 [FREE Full text] [doi: 10.2196/10394] [Medline: 29695374]

Abbreviations

ADHD: attention-deficit hyperactivity disorder
AI: artificial intelligence
e–mental health: electronic mental health
ICD: International Classification of Diseases

Edited by G Eysenbach; submitted 28.02.19; peer-reviewed by H Tunnell, G Wadley; comments to author 29.04.19; revised version received 26.05.19; accepted 31.08.19; published 29.10.19

Please cite as:
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
URL: http://formative.jmir.org/2019/4/e13863/
doi: 10.2196/13863
PMID: 31663858