Diagnostic Assessment via Live Telehealth (Phone or Video) Versus Face-to-Face for the Diagnoses of Psychiatric Conditions:

A Systematic Review

Madeleen van der Merwe, MBiostatistics; Tiffany Atkins, BASc.AHN (Hons), GDipEd, MBiostatistics; Anna Mae Scott, PhD; and Paul P. Glasziou, PhD

Abstract

Objective: To determine the validity of telephone or video interviews, compared to face-to-face, for psychiatric diagnosis.

Data Sources: We searched MEDLINE, Embase, and PsycINFO from inception to June 22, 2023, and performed backward and forward citation analysis on all included studies on August 3, 2023.

Study Selection: We included primary studies comparing live telehealth (via telephone or videoconferencing) with face-to-face interviews using the same standardized diagnostic criteria for a mental health condition. Each patient had to undergo both modes of interviewing. Risk of bias was assessed using Quality Assessment of Diagnostic Accuracy Studies-2.

Results: This review included 35 studies. Seven are clinical studies that compare telehealth with face-to-face consultations for initial psychiatric diagnosis; telehealth via video or telephone is a reliable alternative for some specific disorders or for use in some specific populations that were studied. The other 28 studies compared telehealth to face-to-face interviews for the use of mental health standardized diagnostic instruments for a broad range of conditions, including depression, bipolar disorder, posttraumatic stress disorder, social anxiety disorder, and autism spectrum disorder, demonstrating good agreement and reliability. Telehealth holds promise for psychiatric assessments, especially when in-person evaluations are not feasible.

Conclusions: From the limited studies primarily conducted before the expansion of telehealth during the COVID-19 pandemic, several small studies suggest that telehealth's psychiatric diagnoses or assessments of various psychiatric conditions seem to be a viable option and should be considered for certain patients during situations, settings, or environments. More research is needed, as telehealth has become more broadly utilized.

J Clin Psychiatry 2024;85(4):24r15296

Author affiliations are listed at the end of this article.

An important question for psychiatric diagnosis is the validity of live telehealth (by telephone or by videoconferencing), in comparison with face-to-face (in-person) interviews. We systematically reviewed the evidence to address this question.

In previous research, telehealth's effectiveness in managing mental health problems has been found to be similar to face-to-face care. A systematic review of 14 randomized controlled trials found that for adults with posttraumatic stress disorder (PTSD), there was no meaningful difference in PTSD or depression scores between video consultation and face-to-face delivery of care.[1] For patients with depression, a meta-analysis of 9 randomized trials comparing telehealth (mostly using video consultation) to face-to-face care found no difference in clinical improvement.[2] A meta-analysis of trials in patients with anxiety and related disorders found no difference between telehealth and face-to-face treatments.[3] For patients with insomnia, trials of cognitive behavioral therapy for insomnia showed no significant difference,[4] and, finally, 12 trials of psychotherapy for miscellaneous mental health conditions, including bulimia nervosa and tic disorders, found that telehealth and face-to-face therapies were comparable across all outcomes.[5]

However, the applicability is that these systematic reviews mostly involved patients with an already known diagnosis. Whether telehealth is as effective as

Clinical Points

- Given considerable new research, an updated review of the accuracy of telehealth mental health diagnostic assessments compared to face-to-face assessments was warranted.
- Telehealth assessment and diagnosis of a variety of psychiatric conditions may be a practical and valid alternative to in-person assessments and may improve timeliness and access for both patients and clinicians.

face-to-face for diagnosing is less clear. A 2014 systematic review of diagnostic assessment studies comparing telehealth and face-to-face diagnoses for psychiatric conditions found and analyzed 16 relevant studies.[6] It concluded that "there is insufficient evidence that diagnostic telephone interviews for the diagnosis of psychiatric disorders are valid, although results for depression and anxiety disorders seem promising."[6] The review authors noted that researchers had assumed that telephone interviews were only suitable for gathering factual data and not for more sensitive issues. While telephone interviews were more cost-effective, the absence of visual cues and differences in patient responses raised concern. The authors noted that telephone interviews generally show more compliance, evasiveness ("I don't know" answers, or no response at all), and more extreme responses compared with face-to-face interviews. Telephone interviews may be less suitable for people who are hearing impaired, mistrustful, older, or very ill. A systematic review comparing telephone and face-to-face interviews for depression showed good comparability for the 2 methods, but the study quality was generally low.[6]

As many additional studies on mental health diagnostic assessments have been published since Muskens et al,[6] in 2014 we systematically reviewed the evidence addressing the question of how valid live telehealth (by telephone or by videoconferencing) interviews are, in comparison with face-to-face interviews, for psychiatric diagnosis. Building on the methods of the 2014 review,[6] the present systematic review reviewed and synthesized the evidence about (1) sensitivity and specificity of telehealth interviews using face-to-face interviews as the gold standard and (2) agreement between telehealth and face-to-face interviews.

METHODS

We conducted a systematic review of the available research examining the value of telehealth interviews compared to face-to-face interviews in providing a psychiatric diagnosis. This systematic review is reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement,[7] and the review protocol was developed prospectively.

Inclusion Criteria

Participants. Studies on mental health problems (eg, depression, anxiety, phobias, and psychosis) were included. Studies about assessment of severity only, or compared to a known diagnosis, were excluded. Studies of suspected mental health problems were eligible.

Intervention and comparator. Eligible interventions: live telehealth interview (eg, by telephone or videoconference) compared to face-to-face interview.

Outcomes. Primary outcomes included accuracy of diagnosis, namely, validity (eg, sensitivity and specificity) and/or agreement (eg, κ statistics or intraclass correlation coefficients [ICCs]).

Setting. We included studies conducted in the community; studies of hospital inpatients or institutions were excluded.

Study design. Primary studies that compared telehealth and face-to-face interviews using the same standardized diagnostic criteria or processes were included. Each patient had to undergo both modes of interviewing.

Inclusion and Exclusion Criteria

Studies were included if they considered the comparison between telehealth and face-to-face interviewing as a criterion validity issue, with face-to-face interviewing as the gold standard and the agreement between the 2 methods. We had not prespecified this in the protocol, but we included studies where the time between the diagnoses by telehealth and face-to-face was up to 3 months; studies where the interval exceeded 3 months were excluded because the diagnosis may be more likely to become invalid with greater time periods. We included all possible mental health conditions. We excluded case-control designs, where patients with a known diagnosis are compared with a group of volunteers, as these generally overestimate accuracy and do not represent real clinical consultations.

We excluded studies with (1) interviews outside the field of mental health, (2) nonstandardized psychiatric interviews, (3) nondiagnostic interviews, (4) different diagnostic interviews by telephone than face-to-face, (5) different respondents for the 2 interview methods, and (6) interviews using interactive voice response. We excluded those assessing lifetime illness as the timeframe was considered too wide for diagnostic assessment agreement. We also excluded any neurocognitive or dementia conditions as these often went outside the realm of psychiatric disorders.

Search Strategy to Identify Studies

We searched Medline (via PubMed), Embase (via Elsevier), and PsycINFO (via Ovid), from inception until

June 22, 2023. Search strings for each database are provided in Supplementary Appendix 1, in the Supplementary Materials (available from the authors on request). The search was designed by an information specialist. All publication types and languages were included in the search, and we performed a backward and forward citation analysis on all included studies on August 3, 2023.

Study selection and screening. References were screened independently by 2 reviewers (P.P.G. and T.A.). After title and abstract screening, full texts were retrieved for the potentially includable articles. Two authors (P.P.G. and T.A.) independently screened the full texts. Discrepancies were resolved by consensus or by referring to a third author. The selection process was recorded in sufficient detail to complete a PRISMA flow diagram.

Data extraction. We used a data extraction form that was piloted in 2 studies. Data extraction was conducted independently by 2 authors (T.A. and M.vdM.). Discrepancies were resolved by consensus or referring to a third author. The following data for study characteristics and outcomes were extracted:

- Study characteristics: study authors, year, country, type of study (design), and setting.
- The interviews and interviewers: background and training of the interviewers, duration of the interview, and instruments used.
- Participant characteristics: number of participants, age, gender, and diagnoses.
- Relevant outcomes: primary outcomes included accuracy of diagnosis, namely, validity (eg, sensitivity and specificity) and/or agreement (eg, κ statistics or ICCs).

Assessment of the Risk of Bias

Two authors (T.A. and M.vdM.) rated the risk of bias independently. Discrepancies were resolved by consensus or, if needed, referring to a third author (P.P.G).

The risk of bias was assessed using the revised Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool for individual studies.[8] The QUADAS-2 consists of 4 key domains to rate. These are Domain 1 "patient selection," Domain 2 "index test," Domain 3 "reference standard," and Domain 4 "flow and timing."

Although we had not prespecified this in the protocol, we also provided an overall risk of bias rating for each study. The overall risk of bias for a particular study was the highest risk of bias rated for any domain.

Measurement of effect and data syntheses. For the outcome assessment of the selected studies, we examined the validity (sensitivity and specificity) and reliability (percentage agreement, the ICC, and κ values [k]). Sensitivity is the proportion of true positives that are correctly identified by the interview. Specificity is the proportion of true negatives that are correctly identified by the interview. In general, the higher the sensitivity, the lower the specificity, and vice versa.

Percentage agreement is the extent to which the classification from the telephone and face-to-face interviews agrees with each other. Kappa is a measure of reliability in which the agreement between 2 observers is calculated with a correction for chance agreement: a κ value of 0 means that any apparent agreement can be attributed to chance, and a κ value of 1 means perfect agreement. The interpretation of the various ranges of the κ value (k) is outlined by Landis and Koch in 1977.[9]

The ICC is a measure of reliability or interrater agreement. These values range from 0 (no agreement) to 1 (perfect agreement). In 1994, Cicchetti categorized these values with values from 0.75 to 1.0 being considered excellent, 0.60 to 0.74 good, 0.40 to 0.59 fair, and anything less than 0.40 poor.[10]

We had intended to meta-analyze the validity and agreement measures, but this was precluded by the paucity of data reporting the same outcome. As meta-analyses were not possible, we did not measure the heterogeneity among the included studies using the I^2 statistic. The unit of analysis was individual patients. We did not contact investigators or study sponsors to provide missing data.

Assessment of publication biases. We did not assess publication bias/small studies effect because appropriate methods have not been developed for this type of review.

Subgroup and sensitivity analysis. Data were insufficient to undertake prespecified subgroup analyses by type of interview (telephone or video) and by diagnostic categories. We did not prespecify any sensitivity analyses, and none were conducted.

RESULTS

Results of the Search

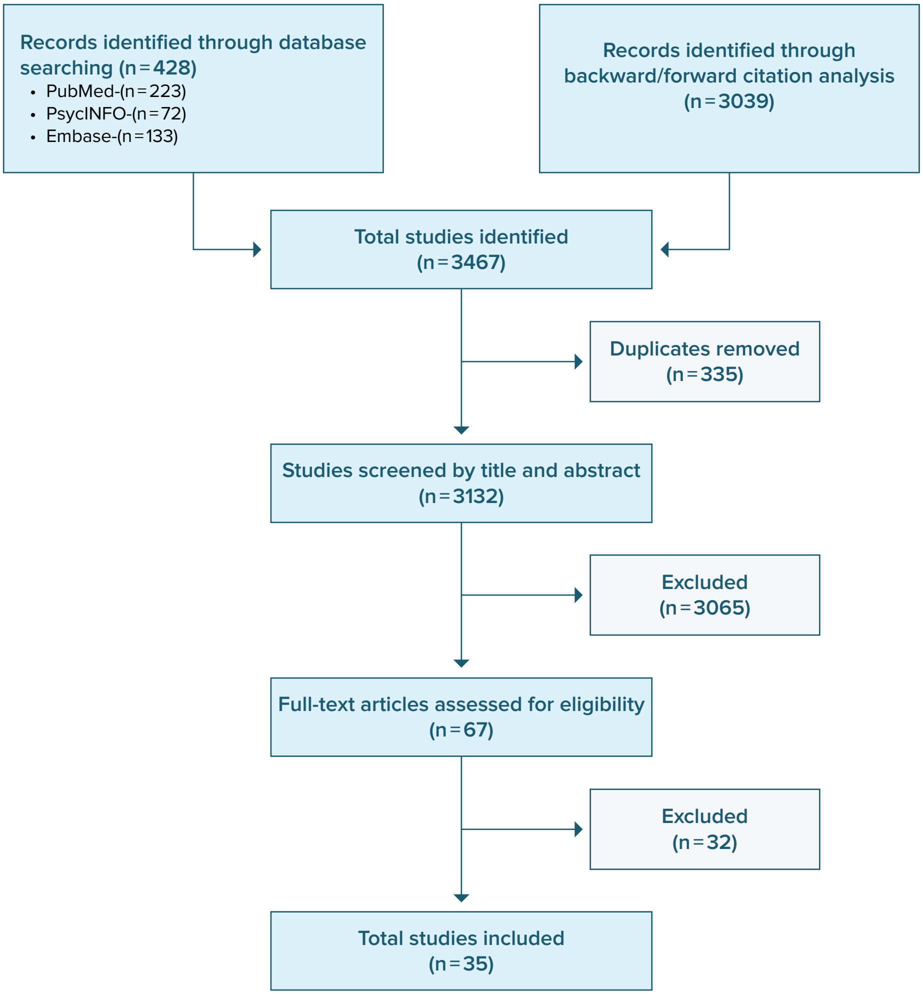
The searches yielded 428 records from database searching and 3,039 records from the backward (cited) and forward (citing) citation analysis, yielding a total of 3,467 references. After deduplication, 3,132 records were screened in title and abstract, and 3,065 were excluded. A total of 67 references were screened in full text, and 32 were excluded (see Supplementary Material, Supplementary Appendix 2 for the full list and reasons for exclusion [Supplementary Materials available from the authors on request]). A total of 35 studies (across 35 references) were included (Figure 1).

Summary of Included Studies (Overall)

Of the 35 identified studies, only 7 compared telehealth with face-to-face consultations for initial clinical psychiatric diagnoses and, therefore, conducted in clinical settings. The remaining 28 studies (18 with specific conditions and 10 with multiple or miscellaneous conditions) were conducted in nonclinical settings and assessed the agreement between telehealth and face-to-

Figure 1.

PRISMA Flow Diagram


Records identified through database searching (n = 428)
• PubMed-(n = 223)
• PsycINFO-(n = 72)
• Embase-(n = 133)
Records identified through backward/forward citation analysis (n = 3039)
Total studies identified (n = 3467)
Duplicates removed (n = 335)
Studies screened by title and abstract (n = 3132)
Excluded (n = 3065)
Full-text articles assessed for eligibility (n = 67)
Excluded (n = 32)
Total studies included (n = 35)

Abbreviation: PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

face interviews or assessed the ICC of structured diagnostic instruments between telehealth and in-person interviews. Of the 35 included studies, only the study by Bistre et al (2022) was conducted after the COVID-19 pandemic.

Clinical Psychiatric Diagnosis in Real Clinical Settings

Seven small studies have compared telehealth with face-to-face consultation for initial psychiatric diagnosis in different clinical settings (see Table 1 for details). In brief, these were listed as follows.

Emergency room assessments. Two studies of emergency presentations concluded that telepsychiatry via video is a reliable and acceptable alternative to face-to-face psychiatric assessments. Bistre and colleagues[11] undertook a prospective study of psychiatric patients presenting to an emergency room in 2020. Patients had both a videoconference assessment and a face-to-face assessment with questionnaires or tests based on the *DSM-5* criteria, and assessors were blinded to each other. While not randomized, the patients had roughly equal frequencies of which mode was done first (20 and 18, respectively). A third opinion was provided by the on-duty psychiatry resident who observed both face-to-face and video assessments. There were high levels of agreement on both the probable diagnosis and the recommended disposition (admission or discharge), with κ values ranging from 0.81 to 0.95. This small but high-quality study concluded that telepsychiatry via video is a reliable and acceptable alternative to face-to-face psychiatric assessments on emergency room assessments.

Seidel and Kilgus[12] conducted a prospective study of psychiatric patients presenting to the emergency department in Virginia. Patients were randomized to either a face-to-face assessment or a videoconference

Table 1.

Studies Comparing Telehealth With Face-to-Face Consultation for Initial Psychiatric Diagnosis in Various Clinical Settings (7 Studies)

Study year (location)	Setting of psychiatric diagnosis	Participants (N), mean age (SD)	Index interview, interview order (time interval)	Tests, instruments	Kappa (k) % agreement (agrm)
Bistre et al 2002 (US)[11]	Emergency room of mental health center Various conditions	Psychiatric patients (n = 38), 38 y (13)	Video, not randomized but balanced (20 vs 18) (same day)	Various psychiatric tests from *DSM-5* used	Recommended disposition, $k = 0.84$ Indication for involuntary admission, $k = 0.81$ Probable diagnosis, $k = 0.52$
Seidel and Kilgus 2014 (US)[12]	ED at hospital Various conditions	Patients >18 y Voluntarily presenting to ED	Video, simultaneously simultaneous	Disposition (discharge/hospital) Diagnosis (*DSM-IV* Axis I), HCR-20 Dangerousness Scale	Disposition, $k = 0.37$ Disposition rating scale, $k = 0.46$ HCR-20, $k = 0.45$ Diagnosis (*DSM-IV* Axis I), $k = 0.27$
Marcantonio et al 1998 (US)[13]	Urban teaching hospital Postsurgical delirium	Patients who had hip fractures and surgery (n = 41), 78 y (8)	Telephone, telephone first (1–4 d)	Diagnosis of delirium	$k = 0.82$
Singh et al 2007 (UK/NZ)[14]	New referral assessments Various conditions	New adult psychiatric referrals (n = 37), 35 y (12.5)	Video, randomly assigned (same day)	*DSM-IV* axis; Axis I–V, risk to self, admit-discharge-follow-up, etc, with a total of 27 attributes	Range across significant Axis I–V (excluding 2) ($k = 0.65$–$k = 0.90$) Combined *DSM-IV*, $k = 0.86$ Overall $k = 0.60$ (27 attributes)
Burke et al 1995 (US)[15]	University Medical Center Diagnosis of geriatric depression	Elderly outpatients of cognitive deficits or functional failure (n = 101), 77 y (7.1)	Telephone, telephone first, (1 wk)	GDS	GDS κ ranged from 0.29–0.75 with a mean $k = 0.52$
Paing et al 2010 (US)[16]	Diagnostic clinic of children's hospital Various initial conditions	Parents of children with suspected psychiatric illness (n = 12), 12 y (3.9)	Telephone, F2F first except 1 (unknown)	P-ChIPS which assesses for the presence of 21 disorders	% agrm (not kappa) ranged from 75% to 100%[a] Mean agrm (all diagnoses) 93.8%
Shore et al 2007 (US)[17]	Rural Native American community	American male native Vietnam veterans (n = 53), 54 y (range; 46–71)	Video, randomly assigned (2 wk)	SCID-III-R (past month, past year, and lifetime)	Major depression $k = 0.25$ (85% agrm) GAD $k = 1.0$ (100% agrm) Panic disorder, $k = 0.34$ (89.0% agrm) Alcohol dependence $k = 0.76$ (89% agrm) Overall,[b] majority $k > 0.6$ Externalizing disorders, $k = 0.70$ vs internalizing disorders, $k = 0.53$

[a] % agreement was also calculated as the prevalence was low (kappa underestimates when the true prevalence in the population is low).

[b] Overall across past month, past year, and lifetime.

Abbreviations: *DSM-IV = Diagnostic and Statistical Manual of Mental Disorders,* Fourth Edition, *DSM-5 = Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition, ED = emergency department, F2F = face-to-face, GDS = Geriatric Depression Scale, HCR-20 = Historical Clinical Risk Management-20, P-ChIPS = Parent's Version of the Children's Interview for Psychiatric Syndromes, SCID = Structured Clinical Interview for *DSM* Disorders.

assessment (lasting approximately 30 minutes), with a second psychiatrist as an observer who also provided a second opinion on the diagnosis (*DSM-IV* Axis I). For the 73 adult patients (48% of patients with depressive disorder, 18% with substance use disorder, 14% with bipolar disorder, 11% with psychosis, and 10% with other diagnoses), agreement between the assessments and diagnosis of the observer psychiatrist and the face-to-face or videoconference consultations was similar. The authors concluded that the results provided preliminary support for the safe use of telepsychiatry in the emergency department to determine the need for admission to inpatient care.

Postsurgical delirium. Marcantonio and colleagues[13] assessed patients for delirium 30 days after discharge for surgery for a hip fracture. The telephone interview was conducted first, with a face-to-face interview as soon as possible after the telephone interview (range: 1–4 days). Of the 41 subjects, by face-to-face assessment, 6 were diagnosed as delirious, and 35 patients were diagnosed as not delirious. All 6 patients with delirium were assessed as delirious by telephone, and of the 35 nondelirious, 33 patients were not delirious by telephone (ie, there were 2 false-positive assessments). However, 4 patients were unable to complete the telephone interview because they were "too confused" (but had been able to at baseline) and were classified as delirious.

New referral assessments. Singh and colleagues[14] evaluated the accuracy of psychiatric assessment for 37 consecutive new adult psychiatric referrals to the Napier Community Mental Health Team. Assessment was done on the same day—in random order—via both face-to-face and videoconference; assessors were blinded to the findings of the alternative mode. The intermode reliability was good, with the *DSM* diagnosis, risk assessment, and interventions κ values all above 0.76 and a combined overall accuracy ratio of 0.8. The authors concluded that telepsychiatry is a dependable mode of service delivery for diagnostic assessment and psychiatric intervention in routine new referrals.

Assessment of depression in the elderly. Burke and colleagues[15] evaluated consecutive patients scheduled for a US outpatient geriatric assessment clinic. Most patients were referred for cognitive dissonance deficits by their primary physician, social services, agency, or family. They underwent a geriatric depression scale assessment face-to-face and by telephone (in random order); the results of the assessment were compared with the final clinical diagnosis by a psychiatrist. The individual items showed good agreement, and the number of positive responses was not significantly different between the 2 methods. The authors concluded that administering the geriatric depression scale via telephone had good validity and reliability for both epidemiologic and clinical purposes.

Parent interview for childhood psychiatric syndromes. Paing and colleagues[16] administered the Parent's version of the Children's Interview for Psychiatric Syndromes to a sample of 12 clinically referred parents of children and adolescents with suspected psychiatric illness. The interview aims to make a psychiatric diagnosis. The most common diagnoses were oppositional defiant disorder, major depressive disorder, bipolar disorder, and anxiety disorder. The percentage of agreement was generally high for each specific diagnosis, ranging from 75% to 100% agreement between telephone and face-to-face diagnosis. However, the authors characterized this as a preliminary study due to a very small sample.

Native American Vietnam veterans. A total of 53 male Native American veterans were randomly assigned[17] to undergo the Structured Clinical Interview for *DSM-III-R* (SCID) of psychiatric assessments, over 2 separate occasions, by different interviewers, via face-to-face and real-time interactive videoconferencing within 2 weeks. Percent agreement between modalities was greater than 80%, except for lifetime drug abuse (76%), lifetime substance abuse (72%), and lifetime major depressive disorder (66%). The authors concluded that SCID assessment by live interactive videoconferencing did not differ significantly from face-to-face assessment in this population.

While all these studies are relatively small, they consistently found a relatively high level of agreement between face-to-face and telepsychiatry assessments. Four of the studies used videoconference, while 3 used telephone. There were no studies of the comparative performance of video vs telephone for initial psychiatric diagnosis in clinical settings.

Assessments in Nonclinical Settings

Most of the remaining studies (n = 28) were not conducted in clinical settings but instead compared the reliability of standardized diagnostic instruments conducted face-to-face and by telehealth. All suggested similar reliability for telehealth and face-to-face assessment, with the most common diagnostic areas including depression (7 studies) and miscellaneous/multiple conditions (10 studies). There were 3 studies each for psychosis, PTSD, and bipolar disease and 1 study each for autism spectrum disorder (ASD) and social anxiety disorder (SAD).

Assessing depression. Most of the 7 studies (Table 2), which assessed the correlation or agreement between telehealth and in-person assessments of depression, found substantial levels of agreement or excellent interrater reliability between telehealth and face-to-face assessments of depression using a variety of subjects and different tests.

Kobak[18] assessed whether administering the Hamilton Depression Rating Scale (HDRS) via video affected the psychometric properties or equivalence of the test. The interrater reliability, as measured by the ICC, was considered acceptable (ICC = 0.80; 95% CI, 0.74–0.95). In 2008, Kobak et al[19] assessed both video and telephone interviews vs in-person interviews using the Montgomery-Asberg Depression Rating Scale (MADRS) and found that telephone (ICC = 0.94 [$P < .0001$]) had an ICC comparable with video (ICC = 0.93 ($P < .0001$)) and concluded that the assessment of patients using the MADRS by telehealth (both video and telehealth methods) was comparable to face-to-face administration. Furthermore, Hermens et al[20] also assessed the interrater reliability of the MADRS by telephone vs in-person and measured a lower ICC of 0.65 but still considered a good level of agreement.

Tunstall et al[21] and Burke et al[22] assessed the agreement between telephone and face-to-face assessments involving elderly patients. Burke assessed the GDS (mean $k = 0.62$), and Tunstall assessed the Depression Diagnostic Scale ($k = 0.79$); both indicated substantial levels of agreement.

Simon et al[23] evaluated the agreement between telephone and in-person assessments in people with an average age of 40 years, using the Structured Clinical Interview, finding a κ value of 0.73 (good level of agreement) for current major depression.

Table 2.

Studies Comparing Telehealth With Face-to-Face Consultation for Assessing Specific Conditions (n = 18 Studies)

Study (location)	Setting	Participants (N), mean age (SD)	Index interview, interview order, time interval	Tests, instruments	ICC, kappa (*k*), or sensitivity and specificity
Depression (n = 7)					
Kobak 2004 (US)[18]	Recruited through newspapers	Subjects met the criteria for an affective disorder (n = 21), 37 y (12)	Video, not randomized but balanced, same day	HDRS	ICC = 0.80 (0.74–0.95)
Kobak et al 2008 (US)[19]	Recruited through newspapers	Subjects diagnosed with mood disorder (n = 35 F2F vs tel), 43 y (range 20–72) (n = 35 F2F vs video), 44 y (range 21–66)	Telephone, counterbalanced, same day Video, balanced, same day	MADRS MADRS	ICC = 0.94 (*P* < .0001) ICC = 0.93 (*P* < .0001)
Hermens et al 2006 (The Netherlands)[20]	Primary GP clinic	Primary care patients minor or mild-major depression (n = 66), 44 y (17)	Telephone, F2F first, a few days	MADRS	ICC = 0.65
Tunstall et al 1997 (UK)[21]	Geriatric or psychogeriatric day hospital	Elderly patients with depression/dementia (n = 29)	Telephone, randomized, 7 d	DPDS DDS	DPDS *k* = 0.79 (0.57–1)
Burke et al 1997 (US)[22]	Geriatric Assessment Center (University Medical Center)	Psychiatric outpatients (n = 83), 77 y (7.0)	Telephone, telephone first, 1 wk	CS-GDS	*K* range: 0.33–0.85, mean *k* = 0.62
Simon et al 1993 (US)[23]	Mental Health Clinic of a large staff-model HMO-major depression	Subjects >18 y starting antidepressants (n = 30), 40 y (range 24–68 y)	Telephone, not randomized, 1–7 d	SCID current major depression HDRS	SCID current major depression *k* = 0.73
Wells et al 1988[24]	Recruited from 2nd wave of study	50% of subjects had symptoms of depression (n = 230)	Telephone, F2F first, 3 mo on average	Depression section of DIS: Lifetime depression	Lifetime major depression, *k* = 0.45 Lifetime dysthymia, *k* = 0.48 Lifetime major depression and/or dysthymia, *k* = 0.57
Bipolar disorder (n = 3)					
Brar et al 2002 (US)[25]	Outpatient clinic of Psychiatric Institute and community	Psychiatric outpatients and others suspected of having BPD-I (n = 20)	Telephone, randomly assigned, median 20 d (3–120 d)	DIGS	IRR ranged from −0.05 for number of clean episodes of mania as part of mania to duration of most severe mania = 1.0, age at most severe depression = 0.90
Feldman-Naim et al 1997 (US)[26]	Outpatient research clinic	Outpatients with rapid cycling bipolar disorder (n = 14), 42 y (6.8)	Telephone, telephone first, same day	HIGH-SAD SIGH-SAD	HIGH-SAD ICC = 0.85 SIGH-SAD ICC = 0.94
Revicki et al 1997 (US)[27]	2 psychiatric treatment centers	Outpatients with bipolar (n = 28), 36 y (7.9)	Telephone, randomly assigned, average 4 d (range 1–19 d)	*DSM-III-R*	Mania, *k* = 0.78 Major depression, *k* = 1.00 Alcohol abuse, *k* = 0.61 Suicide history, *k* = 0.80
PTSD (n = 3)					
Aziz and Kenford 2004 (US)[28]	Veterans Affairs Medical Center	Male veterans (n = 34), 54 y (4.8)	Telephone, randomized, 30 d	CAPS HDRS	CAPS-60 *k* = 0.72 (0.50–0.94) CAPS-65 *k* = 0.75 (0.53–0.97) HDRS-14 *k* = 0.70 (0.46–0.93)
Porcari et al 2009 (US)[29]	PTSD clinic	Male veterans referred to PTSD clinic (n = 20), 50–59 y	Video, randomized, assumed same day	CAPS	All subscales, *k* = 1.0 Overall total score, *k* = 0.32
Litwack et al 2014 (US)[30]	VHA hospital	Trauma-exposed veterans (n=30) F2F and video, 53.2 y (11.5)	Video, pseudorandom, 2 wk	CAPS (*DSM-IV*)	Diagnosis, *k* = 0.83
Psychosis (n = 3)					
Michel et al 2014 (Switzerland)[31]	Psychiatric hospital	Both inpatients and outpatients, n = 94	Telephone, varied but counterbalanced, 1 wk	Past/present psychosis COPER, COGDIS, APS, BLIPS	*K* only for symptom presence Past or present psychosis *k* = 0.97, COPER *k* = 0.57, COGDIS *k* = 0.65, APS *k* = 0.75, BLIPS *k* = 1.0
Hajebi et al 2012 (Iran)[32]	Psychiatric outpatient services	Psychiatric outpatients, n = 72	Telephone, F2F first, 5–10 d	SCID for *DSM-IV*	Primary psychotic disorder (12 mo) Sensitivity = 73.7; specificity = 67.9

(continued)

Table 2 (continued).

Study (location)	Setting	Participants (N), mean age (SD)	Index interview, interview order, time interval	Tests, instruments	ICC, kappa (*k*), or sensitivity and specificity
Yoshino et al 2001 (Japan)[33]	Psychiatric hospital	Chronically ill inpatients met schizophrenia criteria. (n = 42): 14 F2F vs F2F, 14 broadband, 14 narrowband	Video, unknown, no > 4 d	BPRS	F2F vs F2F, ICC = 0.87 Broadband, ICC = 0.88 Narrowband, ICC = 0.44
SAD (n = 1)					
Crippa et al 2008 (US)[34]	Recruited from 2 universities	Undergraduate students with and without SAD (n = 100), 21 y (2.06)	Telephone, Telephone first, Min 1 mo – 2 mo max	SCID (*DSM-IV*)	*k* = 0.84
ASD (n = 1)					
Reese et al 2013 (US)[35]	University Medical Center	Children with ASD or developmental delay (n = 21), 3–5 y	Video, simultaneous, same time	ADOS ADI-R	ADOS *k* = 0.47 ADI-R *k* = 0.74

Abbreviations: ADI-R = Revised Autism Diagnostic Interview, ADOS = Autism Diagnostic Observation Schedule, APS = attenuated psychotic symptoms, ASD = autism spectrum disorder, BLIPS = brief limited intermittent psychotic, BPRS = Brief Psychiatric Rating Scale, CAPS = Clinician-Administered PTSD Scale, COGDIS = cognitive disturbances, COPER = cognitive-receptive basic symptoms, CS-GDS = Collateral Source Version of the Geriatric Depression Scale, DDS = Dementia Diagnostic Scale, DIGS = Diagnostic Interview for Genetic Studies, DIS = Diagnostic Interview Schedule, DPDS = Depression Diagnostic Scale, *DSM-III* = *Diagnostic and Statistical Manual of Mental Disorders*, Third Edition, *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, F2F = face to face, GP = general practitioner, HDRS = Hamilton Depression Rating Scale, HIGH-SAD = Hypomania Interview Guide Including Hyperthymia-Seasonal Affective Disorder, ICC = intraclass correlation coefficient, HMO = health maintenance organization, MADRS = Montgomery-Asberg Depression Rating Scale, PTSD = posttraumatic stress disorder, SCID = Structured Clinical Interview for *DSM* Disorders, SIGH-SAD = Structured Interview Guide for Hamilton Depression Rating Scale-Seasonal Affective Disorder, VHA = Veterans Health Administration.

Finally, Wells et al[24] assessed lifetime depression using the depression section of the Diagnostic Interview Schedule (DIS), with κ values ranging from 0.45 to 0.57, considered a moderate level of agreement.

Assessing bipolar disorder. Three studies (Table 2) compared the agreement or interrater reliability between telephone and face-to-face assessments of bipolar disorder using a variety of tests. All patients either had or were suspected of having bipolar disorder.

Brar et al[25] assessed the Diagnostic Interview for Genetic Studies, which consists of up to 25 items. Seven items had unsatisfactory reliability; however, overall, the telephone interview was considered reliable for most items tested, and the authors determined that it seemed more reliable to assess bipolar I disorder in the absence of psychotic features or substance abuse.

Feldman-Naim et al[26] found a high level of correlation between the telephone and face-to-face administration of both the Hypomania Interview Guide Including Hyperthymia-Seasonal Affective Disorder (ICC = 0.85) and the Structured Interview Guide for Hamilton Depression Rating Scale-Seasonal Affective Disorder (ICC = 0.94).

Revicki et al[27] demonstrated good to excellent levels of agreement using the *DSM-III-R*, with most κ values ranging from 0.61 to 0.78.

Overall, telephone assessments are deemed an acceptable alternative for assessing patients with bipolar disorder compared to in-person assessments.

Assessing PTSD. Three studies compared telehealth with face-to-face assessment for PTSD (Table 2). One study (Aziz and Kenford[28]) compared the agreement between telephone and face-to-face assessments, and 2 studies (Porcari et al[29] and Litwack et al[30]) compared the agreement between video and face-to-face assessments. Both video and telephone methods of interviewing demonstrated acceptable levels of agreement, suggesting that the Clinician-Administered PTSD Scale (CAPS) can be delivered via telehealth.

Aziz and Kenford in 2004 measured the agreement of the CAPS at 2 different cut-points: CAPS-60 ($k = 0.72$) and CAPS-65 ($k = 0.75$), finding a substantial level of agreement.

Porcari et al[29] found perfect agreement ($k = 1.0$) on all the subscales; however, the agreement for the overall total score for PTSD diagnosis was lower than expected at $k = 0.32$, which is considered a fair agreement level. This may have been due to a small number of patients (N = 20) who were referred to the PTSD clinic but did not have an existing PTSD diagnosis.

Litwack et al demonstrated high interrater reliability between video and face-to-face assessments of the diagnosis of PTSD using the CAPS ($k = 0.83$).

Assessing psychosis. Three studies assessed psychosis (Table 2). Two studies (Michel et al[31] and Hajebi et al[32]) evaluated the agreement between telephone and face-to-face assessments. One study (Yoshino et al[33]) compared the correlation between video and face-to-face assessments. Both video and telephone are acceptable alternative interview methods compared to face-to-face interviews.

Yoshino et al[33] compared video with face-to-face assessment (assessed both narrowband and broadband video) using the Brief Psychiatric Rating Scale (BPRS). Narrowband (ICC = 0.44), which was based on older technology from the early 2000s, had a significantly lower ICC than broadband (ICC = 0.88). However, broadband video, which is common today, was just as reliable as face-to-face (ICC = 0.87).

Michel et al[31] compared telephone with face-to-face assessment, finding moderate to perfect agreement (ranging from 0.57 to 1.0) for symptom presence.

Finally, in 2013, Hajebi et al[32] compared the SCID for *DSM-IV* between telephone and in-person and found the sensitivity to be 73.7% and the specificity to be 67.9%.

Assessing social anxiety disorder

One study (Crippa et al[34]) assessed social anxiety disorder (SAD). The study found that the test-retest κ agreement between the telephone interview and the face-to-face interview for assessing SAD in students with and without SAD using the SCID for *DSM-IV* had an excellent agreement with a κ value of 0.84. This study concluded that the use of the SCID via telephone for SAD assessments is supported (Table 2).

Assessing ASD. One study (Reese et al[35]) assessed children (11 children had ASD and 10 children with developmental delays) from 3 to 5 years old using the Autism Diagnostic Observation Schedule (ADOS) and the Revised Autism Diagnostic Interview (ADI-R) by simultaneous videoconferencing (Table 2). One test instrument, ADOS ($k = 0.47$), had a weaker agreement (at a moderate level) compared to the other, ADI-R ($k = 0.74$; substantial agreement). Overall, there was no significant difference in the reliability of the ADOS and ADI-R between video and in-person assessments. However, the authors concluded that future research should be completed using a larger sample size and with children without an already existing diagnosis of ASD.

Assessing miscellaneous and/or multiple conditions. Ten studies assessed multiple conditions or various conditions (Table 3). Seven studies evaluated the agreement or reliability between telephone and face-to-face interviews. Three studies (Baer et al,[36] Grob et al,[37] Jones et al[38]) compared video with in-person interviews. Overall, most studies found acceptable to substantial levels of agreement; the only exception was the assessment of adjustment disorder with depressed mood, which indicated an unacceptable level of agreement ($k = 0.31$).

Baer et al[36] assessed current patients of an obsessive-compulsive disorder clinic using various scales such as the Yale-Brown Obsessive Compulsive Scale (ICC = 0.99), HDRS (ICC = 0.98), and the Hamilton Anxiety Rating Scale. The ICC of 0.99 demonstrated a very strong agreement between video and in-person interviews. Grob et al[37] also compared video to face-to-face assessment with the Mini-Mental State Examination (ICC = 0.95), GDS (ICC = 0.82), and BPRS (ICC = 0.81) within nursing home residents. They found an excellent level of agreement. Furthermore, Jones et al[38] compared video to in-person assessment using the BPRS (ICC = 0.83) within geriatric patients of a psychiatric unit, finding comparable results.

Cacciola et al[39] assessed the conditions in the SCID-III-R within college men, finding that for current diagnoses, the κ value ranged widely from 0.03 for simple phobia to 0.66 for major depression. Ruskin et al[40] also assessed conditions using the SCID-III-R within psychiatric inpatients, with κ values ranging from 0.70 for major depression to 1.0 for panic disorder.

Hajebi et al[41] also assessed the SCID-I (version *DSM-IV*) and found the telephone to be an acceptable method of interviewing for diagnosing lifetime psychotic disorders (sensitivity and specificity = 80.6), but telephone was not as sensitive (sensitivity = 73.7) or specific (specificity = 67.9) for diagnosing current psychotic disorders.

Evans et al[42] sampled patients from 2 different general practitioner clinics, finding excellent levels of agreement with the 12-Item General Health Questionnaire (GHQ-12) ($k = 0.75$) and the Revised Clinical Interview Schedule (CIS-R) with a κ value of 0.72.

Watson et al[43] assessed community volunteers with the DIS, which assesses a variety of disorders, finding an overall κ >60, which indicates quite good levels of agreement overall.

Rohde et al[44] assessed younger people with (mean age of 24 years) with Axis I and Axis II. The values of κ ranged from 0.67 to 0.84, which indicated excellent agreement between telephone and in-person interviews. The exception was for the adjustment disorder with depressed mood, finding a κ value of 0.31, indicating only slight agreement.

Lyneham and Rapee[45] assessed children with or without anxiety using the Anxiety Disorders Interview Schedule for Children for *DSM-IV*. They showed an excellent agreement between face-to-face and telephone interviews ($k = 0.86$) for the overall principal diagnosis.

Risk of Bias

Most studies were rated at high risk of bias or some concerns (in aggregate: n = 27, 77% of studies) for Domain 1, patient selection. For Domain 2, index test, most of the studies (n = 24, 69%) were rated at low risk of bias, with the remainder of the 35 studies rated at some concerns or high risk of bias. Similarly, for Domain 3, reference standard, most of the studies (n = 22, 63%) were rated at low risk of bias, with the remainder rated at some concerns or high. Domain 4, flow and timing, showed a similar pattern, with most of the studies being rated at low risk of bias (n = 26, 74%) and the remainder at some concerns or high.

Overall, very few studies were rated at an overall low risk of bias (11%, n = 4); most of the studies were rated overall as having some concerns (37%, n = 13) or at a high risk of bias (51%, n = 18).

The risk of bias of the included studies is presented in Figure 2.

Table 3.

Studies Comparing Telehealth With Face-to-Face Consultation for Assessing the Following Conditions in Miscellaneous and/or Multiple Conditions (n = 10 Studies)

Study (location)	Setting	Participants (N), mean age (SD)	Index interview, interview order, time interval	Tests, instrument	ICC, κ (k), or sensitivity and specificity
Baer et al 1995 (US)[36]	OCD clinic within a general hospital	Current patients of OCD clinic (n = 10), 38 y (13)	Video, balanced (20 vs 18), simultaneous	YBOCS HDRS HARS	YBOCS-ICC = 0.99 HDRS-ICC = 0.98 HARS-ICC = 0.99
Grob et al 2001 (US)[37]	VA nursing home	Nursing home residents (n = 27), 72.5 y (2.8)	Video, sequentially assigned, 1 wk	MMSE GDS BPRS	MMSE, ICC = 0.95 GDS, ICC = 0.82 BPRS, ICC = 0.81
Jones et al 2001 (US)[38]	Adult psychiatry unit from University Medical Center	Geriatric patients collected from Adult Psychiatry Unit (n = 27), 66.5 y (9.1)	Video vs F2F observer, simultaneously, simultaneous	BPRS	Total BPRS, ICC = 0.83 (0.67–0.92)
Cacciola et al 1999 (US)[39]	Recruited from local community colleges	College men with or without paternal alcoholism (n = 41), 22 y (0.8)	Telephone, balanced first 21 F2F, then 20 telephone, mean = 5.9 d (3.2)	SCID-III-R	Current diagnoses Simple phobia *k* = 0.03 Major depression *k* = 0.66
Ruskin et al 1998 (US)[40]	VA mental hygiene clinic Various conditions assessed	Psychiatric inpatients (n = 30, 15 F2F vs F2F, 15 F2F vs telephone)	Telephone (also F2F and F2F) balanced, 1–2 d	SCID-III-R	Major depression, *k* = 0.70 (vs 0.73 F2F vs F2F) Bipolar disorder, *k* = 0.81 (*k* = 0.76 F2F vs F2F) Panic disorder, *k* = 1.0 (*k* = 1.0, F2F vs F2F) Alcohol dependence, *k* = 0.86 (*k* = 0.86, F2F vs F2F)
Hajebi et al 2012 (Iran)[41]	2 Psychiatric outpatient services	Psychiatric outpatients (*n* = 72), 35 y (9.9)	Telephone, F2F first, Max 2 wk	SCID-I (*DSM-IV*)	Primary psychotic disorder (Past 12 mo) sensitivity = 73.7 and specificity = 67.9 (Lifetime) sensitivity = 80.6 and specificity = 80.6
Evans et al 2004 (UK)[42]	2 General practices	Sample of patients in GP clinics (n = 98), 51 y (17.4)	Telephone, alternating, within 48 h	GHQ-12 CIS-R	GHQ (*k* = 0.75 [SE = 0.10]) CIS-R (*k* = 0.72 [SE = 0.10])
Watson et al 1992 (US)[43]	Veterans Administration Medical Center	Community volunteers (n = 49), 40 y (9.2)	Telephone, mostly balanced (first 26 patients F2F), 1 wk	DIS	5 substance use disorders, mean *k* = 0.92 8 anxiety disorders, mean *k* = 0.62 Overall *k* > 60
Rohde et al 1997 (US)[44]	Follow-up from the Oregon Adolescent Depression Project	Patients with and without a disorder (n = 120; 60 for Axis I and 60 for Axis II), 24 y (0.3)	Telephone, counterbalanced, median=14 d	Axis I Axis II	Test-retest reliability Axis I=major depressive disorder, *k* = 0.67 Anxiety disorders, *k* = 0.84 Alcohol and substance abuse, *k* = 0.70, 0.73 Adjustment disorder (depressed mood) *k* = 0.31 Axis II, any personality disorder ICC = 0.8
Lyneham and Rapee 2005 (Australia)[45]	Community anxiety clinic/ newspaper advertisement	Children with or without anxiety (n = 73), 9 y (1.9)	Telephone, counterbalanced, 6.8 d (range 1–19)	ADIS-C-IV	Overall principal diagnosis between F2F and telephone, *k* = 0.86

Abbreviations: ADIS-C-IV = Anxiety Disorders Interview Schedule for Children for *DSM-IV*, BPRS = Brief Psychiatric Rating Scale, CIS-R = Revised Clinical Interview Schedule, DIS = Diagnostic Interview Schedule, *DSM-IV* = *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, F2F = face to face, GAD = generalized anxiety disorder, GDS = Geriatric Depression Scale, GHQ-12 = 12-Item General Health Questionnaire, HARS = Hamilton Anxiety Rating Scale, HDRS = Hamilton Depression Rating Scale, ICC = intraclass correlation coefficient, MMSE = Mini-Mental State Examination, OCD = obsessive-compulsive disorder, SCID-I = Structured Clinical interview, VA = Veterans Affairs, YBOCS = Yale-Brown Obsessive Compulsive Scale.

DISCUSSION

Our systematic review included 35 studies across different clinical settings and psychiatric conditions. Their findings suggest that clinical psychiatric diagnoses by telehealth (by either telephone or video) vs face-to-face generally achieve an overall acceptable to excellent level of agreement or interrater reliability. The assessment of diagnosis for various psychiatric conditions by telehealth is likely to be acceptable, especially in circumstances where it is not practical or expedient to see the patient face-to-face. It is important

Figure 2.
Risk of Bias for the Included Studies

	D1	D2	D3	D4	Overall
Autism spectrum disorder					
Reese 2013	×	+	+	+	×
Bipolar					
Brar 2002	−	×	×	+	×
Feldman-Naim 1997	×	+	+	+	×
Revicki 1997	+	+	−	+	−
Depression					
Kobak 2008	−	+	+	+	−
Hermens 2006	+	+	+	+	+
Kobak 2004	−	+	+	+	−
Wells 1998	+	−	×	+	×
Burke 1997	+	+	+	×	×
Tunstall 1997	×	−	−	+	×
Simon 1993	+	+	+	+	+
Diagnosis					
Bistre 2022	−	+	+	×	×
Seidel 2014	−	+	+	+	−
Paing 2010	×	−	−	−	×
Shore 2007	−	+	+	+	−
Singh 2007	+	+	+	+	+
Marcantonio 1998	−	×	−	×	×
Burke 1995	+	+	+	+	+
Psychosis					
Michel 2014	−	+	+	−	−
Buckley 2013	−	−	−	−	−
Yoshino 2001	×	+	−	+	×

	D1	D2	D3	D4	Overall
PTSD					
Litwack 2014	−	−	−	+	−
Porcari 2009	−	+	+	−	−
Aziz 2004	−	+	+	+	−
Social anxiety disorder					
Crippa 2008	×	+	+	×	×
Miscellaneous					
Hajebi 2012	−	+	+	+	−
Lyneham 2005	×	+	+	+	×
Evans 2004	+	×	×	+	×
Grob 2001	−	−	−	−	−
Jones 2001	−	+	+	+	−
Cacciola 1999	×	+	+	+	×
Ruskin 1998	×	+	+	+	×
Rohde 1997	×	−	−	+	×
Baer 1995	×	+	+	+	×
Watson	×	×	×	+	×

Domains:
D1: Patient selection
D2: Index test
D3: Reference standard
D4: Flow and timing

Judgement
× High
+ Low
− Unclear

to note that very few studies were overall rated as having a low risk of bias (11%, n = 4); most studies were rated as having some concerns (37%, n = 13) or a high risk of bias (51%, n = 18) overall.

Seven studies were conducted in real clinical settings—eg, in emergency departments, new psychiatric referrals, or checking for postsurgery delirium. Four studies used videoconference, and 3 used telephone. While all these studies are relatively small, they were consistent in finding a relatively high level of agreement between face-to-face and telepsychiatry assessments.

Most of the remaining 28 studies were not in clinical settings but rather examined the reliability of standardized diagnostic instruments. All suggested similar interrater reliability or agreement between telehealth and face-to-face assessments, with the most common diagnostic areas including depression (7 studies) and miscellaneous/ multiple conditions (10 studies) as well as 3 studies each for psychosis, PTSD, and bipolar disease and 1 study each for ASD and SAD.

The 2014 review by Muskens et al[6] included 16 studies (14 we included) compared to the total 28 studies we identified, concluding that "There is insufficient evidence that diagnostic telephone interviews for the diagnosis of psychiatric disorders are valid, although results for depression and anxiety disorders seem promising." The evidence since 2014 has strengthened, and importantly, additional studies have been conducted in real clinical settings. However, as might be expected with the broad range of psychiatric problems, most conditions only have a few relevant studies. It is worth noting, however, that our findings are consistent with several other reviews of telehealth for diagnostic purposes in the psychiatric and mental health space. For example, a scoping review of 10 studies comparing telehealth (both synchronous and asynchronous) to face-to-face diagnosis of ASD found that the accuracy of telehealth diagnosis was 80%–91%.[46] A systematic review of telehealth diagnosis of dementia and mild cognitive impairment found the sensitivity of telehealth of 0.8–1.0 for the dementia diagnosis and 0.71 (95% CI, 0.54–0.84) for the mild cognitive impairment diagnosis.[47] Another systematic review of telehealth diagnosis in children with developmental concerns likewise found a high diagnostic agreement between telehealth and face-to-face diagnoses and additionally reported a high level of stakeholder satisfaction.[48] Systematic reviews for other conditions have also generally found a high rate of diagnostic accuracy for telehealth, compared to face-to-face, for example, for the diagnosis of otorhinolaryngological diseases (accurate diagnosis for 86% of patients)[49] and surgical site infections of adult patients (diagnostic accuracy ranging from 70% to 100%).[50] This is not uniform, however. For example, live teleophthalmology compared to face-to-face diagnosis of common eye health conditions was found to be superior or comparable[51] while asynchronous telehealth (store-and-forward) diagnosis of dental caries and enamel defects found equivalent or superior diagnostic for store-and-forward for dental caries, but mixed evidence for diagnosis of enamel defects.[52]

The strength of the present review includes its rigorous methodology and comprehensive searches, which identified evidence across a broad range of mental health conditions and patient populations. We also did not restrict the eligibility of includable studies by language, although only studies in English met the inclusion criteria. However, it is worth noting that of the 35 studies that met the inclusion criteria, the majority (28 studies) compared telehealth to face-to-face interviews for the administration of standardized diagnostic instruments rather than for the initial clinical diagnosis. Only 7 studies compared telehealth with face-to-face consultations for the initial diagnosis, suggesting an urgent need for additional evidence of the value of telehealth for this purpose. Six of those 7 studies were conducted in the United States, which may limit the generalizability of their findings. Studies were also small (the median sample size was 37, and only 4 studies had sample sizes equal to or greater than 100), and their heterogeneity in terms of the studied populations, conditions, and outcome reporting precluded the ability to conduct prespecified meta-analyses. Finally, a wide range of both synchronous (live) and asynchronous interventions fall under the umbrella of telehealth, for example, mobile apps, store-and-forward platforms, and interactive voice response systems. The present findings apply specifically to live telehealth and cannot be generalized beyond that. Overall, a variety of small studies suggest that psychiatric diagnoses or assessments of various psychiatric conditions by telehealth seem to be a viable option and should be considered for certain patients during situations, settings, or environments. An area of concern that should be focused on in future research is impact of nonverbal cues and physical appearance. Although these findings are generally reassuring, additional research is necessary to verify the applicability of these findings. Furthermore, more investigation is needed in areas that have not been adequately addressed, such as determining the initial training required to reduce the limitations of telehealth. In addition, many of the studies are old and use different technologies to those available today, which also warrants additional investigation.

Article Information

Published Online: September 16, 2024. https://doi.org/10.4088/JCP.24r15296

Submitted: February 18, 2024; accepted July 1, 2024.

To Cite: van der Merwe M, Atkins T, Scott AM, et al. Diagnostic assessment via live telehealth (phone or video) versus face-to-face for the diagnoses of psychiatric conditions: a systematic review. *J Clin Psychiatry.* 2024;85(4):24r15296.

Author Affiliations: Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia (van der Merwe, Atkins, Scott, Glasziou); Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom (Scott).

Corresponding Author: Madeleen van der Merwe, MBiostatistics, Institute for Evidence-Based Healthcare, Bond University, 14 University Dr, Robina, QLD 4226 (mavander@bond.edu.au).

Relevant Financial Relationships: The authors report no actual or potential conflicts of interest.

Funding/Support: This systematic review was commissioned by the Department of Health and Aged Care, Canberra, Australia, as part of a series of systematic reviews on the effectiveness of telehealth within primary care in 2020–21 and their update in 2023. The funder was involved in establishing the parameters of the study question (PICO).

Role of the Funder/Sponsor: The funder was not involved in the conduct, analysis, or interpretation of the systematic review or in the decision to submit the manuscript for publication.

Additional Information: Supplementary material, including the search strings, summary of excluded studies, and PRISMA 2020 checklist, is available from the authors upon request.

ORCID: Tiffany Atkins: https://orcid.org/0009-0007-4564-6859; Madeleen van der Merwe: https://orcid.org/0000-0001-7871-4300; Anna M. Scott: https://orcid.org/0000-0002-0109-9001; Paul P. Glasziou: https://orcid.org/0000-0001-7564-073X

References

1. Scott AM, Bakhit M, Greenwood H, et al. Real-time telehealth versus face-to-face management for patients with PTSD in primary care: a systematic review and meta-analysis. *J Clin Psychiatry*. 2022;83(4):21r14143.
2. Scott AM, Clark J, Greenwood H, et al. Telehealth *v.* face-to-face provision of care to patients with depression: a systematic review and meta-analysis. *Psychol Med*. 2022;52(14):2852–2860.
3. Krzyzaniak N, Greenwood H, Scott AM, et al. The effectiveness of telehealth versus face-to face interventions for anxiety disorders: a systematic review and meta-analysis. *J Telemed Telecare*. 2024;30(2): 250–261.
4. Scott AM, Peiris R, Atkins T, et al. Telehealth versus face-to-face delivery of cognitive behavioural therapy for insomnia: a systematic review and meta-analysis of randomised controlled trials. *J Telemed Telecare*. 2023:1357633X231204071. 10.1177/1357633X231204071
5. Greenwood H, Krzyzaniak N, Peiris R, et al. Telehealth versus face-to-face psychotherapy for less common mental health conditions: systematic review and meta-analysis of randomized controlled trials. *JMIR Ment Health*. 2022;9(3): e31780.
6. Muskens EMH, Lucassen P, Groenleer W, et al. Psychiatric diagnosis by telephone: is it an opportunity? *Soc Psychiatry Psychiatr Epidemiol*. 2014;49(10): 1677–1689.
7. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. 2021; 372:n160.
8. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2_ a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8): 529–536.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.

10. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4): 284–290.
11. Bistre M, Juven-Wetzler A, Argo D, et al. Comparable reliability and acceptability of telepsychiatry and face-to-face psychiatric assessments in the emergency room setting. *Int J Psychiatry Clin Pract*. 2022;26(3):228–233.
12. Seidel RW, Kilgus MD. Agreement between telepsychiatry assessment and face-to-face assessment for emergency department psychiatry patients. *J Telemed Telecare*. 2014;20(2):59–62.
13. Marcantonio ER, Michaels M, Resnick NM. Diagnosing delirium by telephone. *J Gen Intern Med*. 1998;13(9):621–623.
14. Singh SP, Arya D, Peters T. Accuracy of telepsychiatric assessment of new routine outpatient referrals. *BMC Psychiatry*. 2007;7:55.
15. Burke WJ, Roccaforte WH, Wengel SP, et al. The reliability and validity of the Geriatric Depression Rating Scale administered by telephone. *J Am Geriatr Soc*. 1995;43(6):674–679.
16. Paing WW, Weller RA, Dixon TA, et al. Face-to-face versus telephone administration of the parent's version of the children's interview for psychiatric syndromes (P-ChIPS). *Curr Psychiatry Rep*. 2010;12(2):122–126.
17. Shore JH, Savin D, Orton H, et al. Diagnostic reliability of telepsychiatry in American Indian veterans. *Am J Psychiatry*. 2007;164(1):115–118.
18. Kobak KA. A comparison of face-to-face and videoconference administration of the Hamilton Depression Rating Scale. *J Telemed Telecare*. 2004;10(4):231–235.
19. Kobak KA, Williams JBW, Jeglic E, et al. Face-to-face versus remote administration of the Montgomery-Asberg Depression Rating Scale using videoconference and telephone. *Depress Anxiety*. 2008;25(11): 913–919.
20. Hermens MLM, Ader HJ, van Hout HPJ, et al. Administering the MADRS by telephone or face-to-face: a validity study. *Ann Gen Psychiatry*. 2006;5:3.
21. Tunstall N, Prince M, Mann A. Concurrent validity of a telephone-administered version of the Gospel Oak instrument (including the SHORT-CARE). *Int J Geriatr Psychiatry*. 1997;12(10):1035–1038.
22. Burke WJ, Rangwani S, Roccaforte WH, et al. The reliability and validity of the collateral source version of the Geriatric Depression Rating Scale administered by telephone. *Int J Geriatr Psychiatry*. 1997;12(3):288–294.
23. Simon GE, Revicki D, VonKorff M. Telephone assessment of depression severity. *J Psychiatr Res*. 1993;27(3):247–252.
24. Wells KB, Burnam MA, Leake B, et al. Agreement between face-to-face and telephone-administered versions of the depression section of the NIMH diagnostic interview schedule. *J Psychiatr Res*. 1988;22(3):207–220.
25. Brar LK, Brar JS, Deily NG, et al. Can clinical features of bipolar-I disorder be assessed reliably on the telephone? *J Affect Disord*. 2002;71(1–3):221–227.
26. Feldman-Naim S, Myers FS, Clark CH, et al. Agreement between face-to-face and telephone-administered mood ratings in patients with rapid cycling bipolar disorder. *Psychiatry Res*. 1997;71(2):129–132.
27. Revicki DA, Tohen M, Gyulai L, et al. Telephone versus in-person clinical and health status assessment interviews in patients with bipolar disorder. *Harv Rev Psychiatry*. 1997;5(2):75–81.
28. Aziz MA, Kenford S. Comparability of telephone and face-to-face interviews in assessing patients with posttraumatic stress disorder. *J Psychiatr Pract*. 2004; 10(5):307–313.
29. Porcari CE, Amdur RL, Koch EI, et al. Assessment of post-traumatic stress disorder in veterans by videoconferencing and by face-to-face methods. *J Telemed Telecare*. 2009;15(2):89–94.
30. Litwack SD, Jackson CE, Chen M, et al. Validation of the use of video teleconferencing technology in the assessment of PTSD. *Psychol Serv*. 2014;11(3): 290–294.
31. Michel C, Schimmelmann BG, Kupferschmid S, et al. Reliability of telephone assessments of at-risk criteria of psychosis: a comparison to face-to-face interviews. *Schizophr Res*. 2014;153(1–3):251–253.
32. Hajebi A, Motevalian A, Amin-Esmaeili M, et al. Telephone versus face-to-face administration of the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, for diagnosis of psychotic disorders. *Comprehensive Psychiatry*. 2012;53(5):579–583.
33. Yoshino A, Shigemura J, Kobayashi Y, et al. Telepsychiatry: assessment of televideo psychiatric interview reliability with present- and next-generation internet infrastructures. *Acta Psychiatr Scand*. 2001;104(3):223–226.
34. Crippa JAS, de Lima Osório F, Del-Ben CM, et al. Comparability between telephone and face-to-face structured clinical interview for DSM-IV in assessing social anxiety disorder. *Perspect Psychiatr Care*. 2008;44(4):241–247.
35. Reese RM, Jamison R, Wendland M, et al. Evaluating interactive videoconferencing for assessing symptoms of autism. *Telemed J E Health*. 2013; 19(9):671–677.
36. Baer L, Cukor P, Jenike MA, et al. Pilot studies of telemedicine for patients with obsessive-compulsive disorder. *Am J Psychiatry*. 1995;152(9):1383–1385.
37. Grob P, Weintraub D, Sayles D, et al. Psychiatric assessment of a nursing home population using audiovisual telecommunication. *J Geriatr Psychiatry Neurol*. 2001;14(2):63–65.
38. Jones BN 3rd, Johnston D, Reboussin B, et al. Reliability of telepsychiatry assessments: subjective versus observational ratings. *J Geriatr Psychiatry Neurol*. 2001;14(2):66–71.
39. Cacciola JS, Alterman AI, Rutherford MJ, et al. Comparability of telephone and In-person structured clinical interview for DSM-III-R (SCID) diagnoses. *Assessment*. 1999;6(3):235–242.
40. Ruskin PE, Reed S, Kumar R, et al. Reliability and acceptability of psychiatric diagnosis via telecommunication and audiovisual technology. *Psychiatr Serv*. 1998;49(8):1086–1088.
41. Hajebi A, Motevalian A, Amin-Esmaeili M, et al. Telephone versus face-to-face administration of the structured clinical interview for diagnostic and statistical manual of mental disorders, fourth edition, for diagnosis of psychotic disorders. *Compr Psychiatry*. 2012;53(5):579–583.
42. Evans M, Kessler D, Lewis G, et al. Assessing mental health in primary care research using standardized scales: can it be carried out over the telephone? *Psychol Med*. 2004;34(1):157–162.
43. Watson CG, Anderson PE, Thomas D, et al. Comparability of telephone and face to face diagnostic interview schedules. *J Nerv Ment Dis*. 1992;180(8):534–535.
44. Rohde P, Lewinsohn PM, Seeley JR. Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. *Am J Psychiatry*. 1997;154(11): 1593–1598.
45. Lyneham HJ, Rapee RM. Agreement between telephone and in-person delivery of a structured interview for anxiety disorders in children. *J Am Acad Child Adolesc Psychiatry*. 2005;44(3):274–282.
46. Stavropoulos KKM, Bolourian Y, Blacher J. A scoping review of telehealth diagnosis of autism spectrum disorder. *PLoS One*. 2022;17(2):e0263062.
47. McCleery J, Laverty J, Quinn TJ. Diagnostic test accuracy of telehealth assessment for dementia and mild cognitive impairment. *Cochrane Database Syst Rev*. 2021;7(7):CD013786.
48. La Valle C, Johnston E, Tager-Flusberg H. A systematic review of the use of telehealth to facilitate a diagnosis for children with developmental concerns. *Res Dev Disabil*. 2022;127:104269.
49. Shah SR, Munhall CC, Nguyen SA, et al. Diagnostic accuracy and management concordance of otorhinolaryngological diseases through telehealth or remote visits: a systematic review & meta-analysis. *J Telemed Telecare*. 2023: 1357633x231156207. 10.1177/1357633X231156207
50. Ng HJH, Huang D, Rajaratnam V. Diagnosing surgical site infections using telemedicine: a systematic review. *Surgeon*. 2022;20(4):e78–e85.
51. Tan IJ, Dobson LP, Bartnik S, et al. Real-time teleophthalmology versus face-to-face consultation: a systematic review. *J Telemed Telecare*. 2017;23(7):629–638.
52. Inês Meurer M, Caffery LJ, Bradford NK, et al. Accuracy of dental images for the diagnosis of dental caries and enamel defects in children and adolescents: a systematic review. *J Telemed Telecare*. 2015;21(8):449–458.