Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Psychiatry Research 316 (2022) 114728

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

The reliability of symptom assessment by telepsychiatry compared with face to face psychiatric interviews☆

Hiu Yan Yung *, Wai Tat Yeung, Chi Wing Law

Department of Psychiatry, Queen Mary Hospital, 102 Pok Fu Lam Road, Hong Kong

A B S T R A C T

Introduction: With the start of the COVID-19 pandemic, the various social distancing policies imposed have mandated psychiatrists to consider the option of using telepsychiatry as an alternative to face-to-face interview in Hong Kong. Limitations over sample size, methodology and information technology were found in previous studies and the reliability of symptoms assessment remained a concern.

Aim: To evaluate the reliability of assessment of psychiatric symptoms by telepsychiatry comparing with face-to-face psychiatric interview.

Method: This study recruited a sample of adult psychiatric patients in psychiatric wards in Queen Mary Hospital. Semi-structural interviews with the use of standardized psychiatric assessment scales were carried out in telepsychiatry and face-to-face interview respectively by two clinicians and the reliability of psychiatric symptoms elicited were assessed.

Results: 90 patients completed the assessments The inter-method reliability in Hamilton Depression Rating Scale, Hamilton Anxiety Rating Scale, Columbia Suicide Severity Rating Scale and Brief Psychiatric Rating Scale showed good agreement when compared with face-to-face interview.

Conclusion: Symptoms assessment by telepsychiatry is comparable to assessment conducted by face-to-face interview.

Abbreviations

BPRS	brief psychiatric rating scale
C-SSRS	Columbia suicide severity rating scale
HAM-A	Hamilton anxiety rating scale
HDRS	Hamilton depression rating scale
ICC	intraclass correlation coefficient
ICD	International statistical classification of diseases and related health problems
IT	information technology
QMH	Queen Mary Hospital
YMRS	Young Mania rating scale

1. Introduction

1.1. Literature review over reliability of symptoms assessment

Prior to the outbreak of the COVID-19 pandemic, there have been around 20 studies carried out in various countries which investigated the reliability of clinical assessments through telepsychiatry compared to face-to-face interviews. Upon reviewing the studies, while the results of those studies are likely reliable, there are certain limitations on studies setting, sample size, assessment scales and the use of technology (Frueh et al., 2000; Monnier et al., 2003; Hubley et al., 2016).

Since the pandemic, the interest in clinical outcome by telepsychiatry has gained its momentum and new studies have been conducted. Topics include the effectiveness of psychotherapy or counseling via videoconferencing and detecting changes in mental state or severity of symptoms via videoconferencing consultation. In summary, the latest studies suggest that clinical outcome by telepsychiatry is as effective as face-to-face care though they were limited by small sample sizes (Cheli et al., 2020; Dores et al., 2020; Erekson et al., 2020; Lynch et al., 2020; Sequeira et al., 2020; Wyler et al., 2021).

1.2. Development of telepsychiatry in Hong Kong

With the ongoing COVID-19 pandemic, there is a pressing need for the use of teleconsultation services including teleconsultations for patients with psychiatry follow up in both public and private sectors. Some allied health services need to convert their treatment interventions from traditional mode to videoconferencing. The Government of Hong Kong published a new policy in encouraging the adaptation of telehealth services in Hong Kong in 2021. Since then, there have been certain developments, such as launching mobile apps specifically designed for

☆ All authors have materially participated in the research and/or article preparation and all authors have approved the final article.

* Corresponding author.
E-mail address: yhy754@ha.org.hk (H.Y. Yung).

https://doi.org/10.1016/j.psychres.2022.114728
Received 22 March 2022; Received in revised form 10 July 2022; Accepted 14 July 2022
Available online 21 July 2022

patients to book and check appointments, development of smart hospital and setting up real time collection of data by mobile (CHENG, 2021).

Currently, telepsychiatry services in Hong Kong are underdeveloped. Previous study raised concerns on the reliability and quality of assessments. This study aims to preliminary evaluate the reliability of telepsychiatry in carrying out assessments.

1.3. Research limitation on assessing reliability in telepsychiatry

Some studies utilised a observer-interviewer configuration setting, meaning that during the assessment one rater conducted the interview while the other acted as an observer. Such setting could not mimic real-life consultations when patients perform telepsychiatry with clinicians (Baigent et al., 1997; Baer et al., 1995; Zarate et al., 1997; Jones et al., 2001; Matsuura et al., 2000).

The sample size of these studies varied but were generally small. Sample size of studies comparing reliability on psychiatric symptoms assessment (Baigent et al., 1997; Baer et al., 1995; Zarate et al., 1997; Yoshino et al., 2001; Jones et al., 2001; Menon et al., 2001; Matsuura et al., 2000) and agreement on diagnosis (Ruskin et al., 1998; Elford et al., 2000; Singh et al., 2007) ranged from 7 to 63. The studies which involved larger sample size up to a few hundred focused mainly on cognitive assessment via telepsychiatry (Turner et al., 2012; Timpano et al., 2013).

Technology and equipment could limit the reliability of assessment by telepsychiatry. The earliest study used microwave and cable to transmit audiovisual signal (Dwyer, 1973). Later studies in 2007 were conducted using improved technology in signal transmission by phone lines. One study found that the network bandwidth would significantly affect the reliability of telepsychiatry interview (Yoshino et al., 2001). Such issue was also reported by a subsequent meta-analysis, indicating that network bandwidth could be a significant moderator in assessing reliability (Hyler et al., 2005). According to Office of the Communications Authority in Hong Kong, Fiber-to-the-home/building household penetration rate was up to 79.3% by May 2021. This suggests that the technology available for telepsychiatry in Hong Kong currently should be much more advanced compared to settings in previous studies (Office of the Communications Authority in Hong Kong, 2021).

Table 1
Socio-demographics and diagnoses of participants and non-participants.

	Participants (N = 90) Median (IQR) / N (%)	Non-participants (N = 46) Median (IQR) / N (%)	p-value
Age	39 (32.75–52.25)	41.5 (31.75–53.25)	0.95
Gender			0.703
Female	52 (57.8%)	25 (54.3%)	
Male	38 (42.2%)	21 (45.7%)	
Educational level			0.131
Primary school or below	9 (10%)	5 (10.9%)	
Secondary school	44 (48.9%)	30 (65.2%)	
University or above	37 (41.1%)	11 (23.9%)	
Duration known to mental health services (up to time of interview, in years)	3.5 (0 −14)	4 (0–15)	0.636
Diagnosis according to ICD-10			0.704
Organic, including symptomatic, mental disorder (F00-F09)	1 (1.1%)	0 (0%)	
Mental and behavioral disorders due to psychoactive substance use (F10-F19)	8 (8.9%)	7 (15.2%)	
Schizophrenia, schizotypal and delusional disorders (F20-F29)	33 (36.7%)	15 (32.6%)	
Mood disorders (F30-F39)	33 (36.7%)	16 (34.8%)	
Neurotic, stress related and somatoform disorder (F40-F48)	12 (13.3%)	8 (17.4%)	
Disorders of adult personality and behaviors (F60-F69)	3 (3.3%)	0 (0%)	

Table 2
Inter-rater reliability of total score of HDRS, HAM-A, YMRS and C-SSRS.

	Wilcoxon signed-rank test	Spearman's rho correlation coefficient
HDRS	$Z = -0.501$, $p = 0.617$	$r = 0.922$, $p<0.001$
HAM-A	$Z = -0.655$, $p = 0.502$	$r = 0.93$, $p<0.001$
YMRS	$Z=-2.382$, $p = 0.017$	Nil
C-SSRS	$Z=-0.816$, $p = 0.414$	$r = 0.992$, $p<0.001$

Some of the studies utilised standardized assessment scales to assess the reliability in telepsychiatry with face-to-face interview. Limited conclusion could be reached as most of them had limitation over sample size, methodology and technology used. The standardized assessment scales most frequently used in previous studies were Hamilton Depression Rating Scale and Brief Psychiatric Rating Scale. Studies using HDRS identified a remarkable result that the inter-method reliability of assessment of retardation was reasonably high while that of agitation was low (Kobak, 2004, 2008). In studies using BPRS, most items were found to be highly correlated with no significant difference except suboptimal reliability over "blunt affect" (Baigent et al., 1997). It was also found that the reliability in observational items were generally poorer when compared with subjective reported items in telepsychiatry (Jones et al., 2001).

Only one study which assessed patients with obsessive compulsive disorder used HAM-A. The result showed that inter-method reliability of total score of HAM-A was high between telepsychiatry and face-to-face interview but no information was provided regarding individual item of the scales (Baer et al., 1995). For suicidal risk assessment, one study specifically assessed the agreement of suicidal ideations by the Beck Scale for Suicidal Ideation in emergency department for psychiatric patients and showed that the reliability of telepsychiatry was high (Seidel and Kilgus, 2014). More studies focused on use of standardized assessment scales in assessment of cognitive function rather than mood or psychotic symptoms in psychiatric patients (Kirkwood et al., 2000; Montani et al., 1996, 1997; Craig et al., 1999; Amarendran et al., 2011; Turner et al., 2012; Timpano et al., 2013).

1.4. Importance of the study

With the emergence of COVID-19 and various social distancing policies, telehealth care services have become a popular topic. Given the new advocate on social distancing, unpredictable risk of infection of COVID-19 and border control, telehealth care may well be an alternative route to provide medical services.

With the advancing technology over the past decades and the increasing use of telepsychiatry since COVID-19, the study aims to investigate the reliability of psychiatric symptoms assessement using telepsychiatry consultation. We also aim to standardize the sample size, using of multiple standardized assessment tools and establish a proper setting for telepsychiatry consultation in order to have an effective comparison of reliability of psychiatric symptoms assessment.

2. Materials and methods

2.1. Study setting

The study is conducted in the adult psychiatric wards in Queen Mary Hospital. Ethics approval for this study has been obtained from the Hong Kong West cluster Institutional Review Board on human subjects (IRB reference number: UW 20–476)

2.2. Subjects' recruitment

Both male and female patients admitted to psychiatric wards of Queen Mary Hospital who are 18 or above and could communicate by

Cantonese, Putonghua or English were recruited. Subjects who (a) could not provide written consent; (b) were diagnosed with dementia or mental retardation; (c) had communication or language barrier; (d) could not complete questionnaires or assessments; and (e) were admitted only for clinical procedures were excluded. Written informed consents have been obtained from all subjects who agreed to participate in the study.

2.3. Outcome measures

The following standard psychiatric assessment scales were chosen for the study as they are commonly used in clinical interviews during practice.

2.3.1. Hamilton depression rating scale

HDRS, containing 17 items is a widely used clinician-administered depression assessment scale and it was used to assess depressive symptoms in this study. The Chinese version has been well validated. The Interrater reliability, item total-score correlations and the internal reliability were satisfactory (Hamilton, 1960; Zheng et al., 1988).

2.3.2. Hamilton anxiety rating scale

HAM-A, containing 14 items, is one of the first rating scales developed to measure the severity of anxiety symptoms and is widely used in both clinical and research settings. It has been validated in Chinese population. It was used to assess anxiety symptoms in this study (Hamilton, 1959).

2.3.3. Young mania rating scale

YMRS is used to assess manic symptoms in this study. The scale has 11 items and is based on the patient's subjective report of his or her clinical condition and clinical observation made during clinical interviews (Young et al., 1978).

2.3.4. Columbia suicide severity rating scale

C-SSRS is an assessment tool that evaluates suicidal ideation and behavior and to assess suicidal risk, which is an essential aspect in psychiatric assessment in this study. It divides the result into 11 categories. From a prospective cohort study in assessing predictive validity, receiver operating characteristics curve showed a good ability to predict suicidal-related behaviors 6 months after discharge with a sensitivity of 70% and specificity of 65–67% (Madan et al., 2016).

In a retrospective designed analysis, the scale demonstrated good convergent validity and divergent validity. It reported 100% sensitivity and 96% specificity in detecting lifetime suicidal attempts retrospectively (Posner et al., 2011).

2.3.5. Brief psychiatric rating scale

BPRS is mainly used to assess psychotic symptoms in this study. The scale involves 18 items and each item is rated from 0 to 7. The scale is one of the most widely used scales to measure psychotic symptoms (Overall and Gorham, 1962, 1988).

2.4. Study procedures

Socio-demographic characteristics and diagnoses according to ICD, 10th version of all subjects who agreed to participate were collected from records in the Clinical Management System of Hospital Authority of Hong Kong. The raters in this study were two psychiatric trainees who have similar working experience in psychiatry in Hong Kong public services setting. Training over the assessment scales were completed online with calibration between two raters. Discussions were made between the two raters to review the scoring conventions and interview guides so as to provide a chance for raters to calibrate on scoring of the scales. Pilot study was then launched between two raters before the commencement of the formal study.

For each of the participants, a telepsychiatry interview and a face-to-face interview were carried out by the two raters separately The duration of interview was set to be within 1 h for both interviews. Randomization was carried out for the sequence of telepsychiatry interviews and for the rater who carried out telepsychiatry interviews. The two interviews are completed within two consecutive days so as to minimize the variation of symptoms over time. The raters are blinded from the assessment results of each other.

2.5. Equipment and settings for clinical interviews

2.51. Telepsychiatry

Equipment was provided by the Hospital Authority in Queen Mary Hospital. Two iPad (10.2-inch,8th generation,32GB WIFI) were used for telepsychiatry interviews. The app ZOOM was used to carry out telepsychiatry interviews. The study fully complied with security protocol of the Information Technology Department of Hospital Authority.

During each telepsychiatry interview, a ward staff would be present to assist participants for any technical issue. Each participant would stay in a private consultation room in psychiatric wards during the interview while the rater conducted the telepsychiatry interview in doctor's office. The distance between patient's seat and the iPad was fixed at 45 cm to maximize the rater's view of participants' facial expression and gestures over upper part of the body when seated.

Participants were asked to start the interview by pressing the "Start button" in the ZOOM app. Before commencement of the interview, the raters would first confirm with the patients the quality of the reception of image and sound.

2.52. Face to face interviews

Both the participant and the rater stayed in the same private consultation room in psychiatric wards.

2.6. Sample size calculation

Calculation of sample size was based on the intraclass correlation coefficient which planned for comparison of the total score of the standardized assessment scales (Bujang and Baharum, 2017).

Minimum ICC value to be achieved was set at 0.6. Expected value of ICC to be achieve was set at 0.75 (Cicchetti, 1994; Koo and Li, 2016). Power was set at 80%. Significance level was set at 0.05. Number of raters is 2. Minimal sample size calculated was 82.

2.7. Statistical analysis

Statistical analysis was performed with the aid of IBM SPSS Statistics for Windows, Version 26.0 (IBM Corp., Armonk. NY, USA computer software) and R 4.0.2 with library "psych".

Descriptive statistics were used to summarize demographics and clinical data. Continuous data including age and duration known to mental health service was expressed as Median and categorical data including gender, educational level and diagnosis was expressed as absolute proportions. continuous data was analyzed by Mann-Whitney U test and Categorical data was analyzed by Pearson Chi-square test.

Kappa statistics (2 levels of categories) or weighted Kappa statistics (>2 levels of categories) were first used to assess the reliability of items in HDRS, HAM-A and YMRS. When the data was affected by the high prevalence in one homogenous category i.e., high prevalence index (Sim and Wright, 2005), absolute agreement was used for further analysis on the actual agreement between two methods (Viera and Garrett, 2005).

Wilcoxon signed-rank test and Spearman's rho correlation coefficient was used to replace ICC for assessing the agreement between two methods for total score of HDRS, HAM-A, YMRS, C-SSRS and items of BPRS as the data was not normally distributed.

3. Result

3.1. Subjects' recruitment

There were 274 admissions to psychiatric wards J6& J7 in Queen Mary Hospital recorded from 1sth July 2020 to 28th February 2021. Among them, 21 were found to be repeated admissions. 9 were minor (under 18) and hence no valid consent could be obtained for the study. 15 admissions were excluded as the patients were admitted for clinical procedures only.93 were excluded according to the exclusion criteria. In the end, 90 out of 136 patients consented for the study.

3.2. Socio-demographics for participants vs non-participants

Socio-demographics characteristics of participants who completed the study and those who refused to participate were compared. There was no significant difference in age, gender, education level, duration known to mental health services and diagnosis between the groups Table 1).

3.3. Result of pilot study

Overall result showed nearly perfect agreement between two raters for all the scales.

3.4. Result of core study

The combined results of Wilcoxon signed-rank test and Spearman's rho correlation coefficient show that the total score of HDRS, HAM-A and C-SSRS had no significant difference and were positively correlated by both methods of interview. Result of Wilcoxon signed-rank test shows significant difference between the total score of YMRS (Table 2). Reviewing the raw data, due to the limitation in sample collection, 78 out of 90 participants have zero marks by both methods. For the remaining sample, the total score assessed by face-to-face interview in general showed a higher score when compared to telepsychiatry. Details for statistics in sub-items for the assessment scales are attached in Appendix.

The overall inter-method reliability of the individual items in HDRS, HAM-A and YMRS were ranged from 0.5 to 1 except the item "agitation" in HDRS was noted to be 0.33. Actual agreement was reviewed for "hypochondriasis" due to the negative value obtained. Actual agreement obtained was 96.6%.

Table 3 shows the absolute scores of C-SSRS also demonstrate the suicidal assessment for different degree of active suicidal ideations and different types of suicidal attempts.

The result of Wilcoxon signed-rank test for all items in BPRS except "blunt affect" showed no significant difference. The items including emotional withdrawal, tension, hostility, motor retardation and suspiciousness had a relatively low spearman's rho correlation coefficient of 0.57–0.7. Other items showed value of spearman's rho correlation coefficient around 0.8 or above. The item "blunt affect" showed significant difference result when compared across two methods.

Table 3
Socio-demographics and diagnoses of participants and non-participants.

	Telepsychiatry Scores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Face-to-face	**0**	46	0	0	0	0	0	0	0
	1	0	1	0	0	0	0	0	0
	2	0	0	1	1	0	0	0	0
	3	0	0	0	6	0	0	0	0
	4	0	0	0	0	4	0	0	0
	5	0	0	0	0	0	5	0	0
	6	0	0	0	0	0	0	12	0
	7	0	0	0	0	0	0	0	13
	8	0	0	0	0	0	0	0	1

4. Discussion

4.1. Overall reliability for depressive symptoms

Two groups of depressive symptoms can be classified from HDRS, including retarded depression and agitated depression. Overall result shows that the assessment for both groups by telepsychiatry is comparable to face-to-face interviews for all items except agitation obtained substantial agreement (Hamilton, 1960).

4.2. Reliability in psychomotor behaviors

The assessment on agitation is found to have significantly poorer agreement when compared to other items. The finding is consistent with the study by Kobak. The Kobak study found that observation on agitation was limited in telepsychiatry as the assessment on agitation was based on motor movement while the assessment on retardation was relatively better as it was based on both facial expressions and rate of speech (Kobak, 2004).

We also reviewed the score of "Retardation" as its the inter-method reliability is the second lowest among all items. Although a large number of participants in our study scored zero in both modalities of assessment for agitation and retardation, subsequent review shows that telepsychiatry rated lower absolute score in general for both items when compared with face-to-face interviews.

Most participants presented with subtle changes in retardation and agitation with at most 2 out of 4 marks due to the limitation in study design in recrutiment of participants for telepsychiatry. The results suggested that telepsychiatry is less reliable to pick up subtle difference in retardation and agitation, compared to face to face clinical observation.

We expected that the more extreme features of agitation or retardation would be easily pick up by both methods though we could not make a conclusion in the study as the recruited participants did not show these obvious symptoms.

Our raters also provide suggestion based on observation in view of the suboptimal agreement in agitation and retardation. Some participants were observed to show less degree in agitated movement when they were facing the monitor. It appeared that they have limited their actions intentionally to avoid their actions being out of capture range from the camera. We therefore propose that the limitation of image capturing could cause limitation in assessment on agitation. While most of the participants were observed to have tendency to speak at a slower pace and louder volume when compared with face-to-face interviews, the postulation is that participants may perceive such to allow the raters to assess their conditions better via telepsychiatry with the distance between the monitors and participants. Therefore, the raters might score inaccurately during telepsychiatry in these items. Further studies which involved more raters would be needed to verify the hypotheses.

4.3. Reliability in anxiety symptoms

Regarding assessment on anxiety symptoms, HAM-A further divides the anxiety symptoms into two groups (Hamilton, 1959). The first group is called somatic symptoms which include gastro-intestinal, genitourinary, respiratory, cardiovascular, somatic general and autonomic symptoms. Assessment on somatic symptoms by telepsychiatry is comparable with face-to-face interviews.

The second group is psychic symptoms, which include tension, fears, insomnia, anxiety, intellectual changes, depression, and behavior at interview. Reviews on absolute scores for the item "Fear" show mild difference over the scoring only while a large group of participants actually reviewed zero score. It is likely related to the high prevalence in homogeneous category. The only outlier which telepsychiatry rated

score of 3 while face-to-face interview rated score of 0 was reviewed. The patient suffered from delusional disorder. The significant change of marks could be related to the fluctuated mental state due to environmental change after admission.

Our study is the first study that investigate the reliability of telepsychiatry in assessing anxiety symptoms among a general inpatient psychiatric population as previous study only compared to total scores in a group of patients with Obsessive compulsive disorder (Baer et al., 1995).

4.4. Reliability in assessing suicidal risk

The C-SSRS is found to have strong convergent validity with established ideation and behavior scales including the clinician administrated Beck Scale for Suicidal Ideation (Posner et al., 2011). A recent study investigated the agreement of suicidal ideations by the Beck Scale for Suicidal Ideation in emergency department for psychiatric patients showed that the reliability of telepsychiatry is high (Seidel and Kilgus, 2014). As suicidal risk assessment is a major component in psychiatric consultations, the reliability of assessing suicidal risk is a major focus of this study. Half of the participants did not present with suicidal ideas. Among the other half of the participants, 18 participants presented with active suicidal ideas with or without plans and 26 participants presented with actual suicidal attempt, or interrupted suicidal attempt. It was remarkable that there were found to have nearly perfect agreement between two modalities. Our result is comparable with the study by Seidel & Kilgus and confirms that assessment of suicidal risk by telepsychiatry is comparable to that as face-to-face interview.

4.5. Reliability in psychotic symptoms

The items related to symptoms of paranoid belief and schizophrenia in BPRS can be divided into assessment based on observation and assessment based on verbal report. The overall reliability of all items in telepsychiatry is comparable to face-to-face interviews except the item “blunt affect”.

Items based on observation include tension, emotional withdrawal, mannerism and posturing, motor retardation and uncooperativeness. Others are based on verbal report (Overall and Gorham, 1962). However, the two studies which included BPRS for assessing reliability in telepsychiatry showed the result that blunt affect was actually also rated on observation and the assessment was limited by telepsychiatry (Baigent et al., 1997; Jones et al., 2001). The items involving observation showed lower inter-method reliability in general (Jones et al., 2001).

Comparing with our study, even though the technology has improved from telephone to fiber network with a setting more compatible as the daily routine consultation, the items based on observation showed relatively lower correlation when compared with those based on verbal report in general, especially for the items tension, emotional withdrawal and motor retardation.

The result of “Mannerism and posturing” and “uncooperativeness” could not be considered as valid after reviewing the absolute scores as only 1 to 2 patients have scores on these two items.

A similar result that shows significant difference between the two methods on the item “blunt affect” is found. The result suggests that there is limitation in using telepsychiatry to assess items that required observation for rating. The possible reason would be similar to the observation of retardation as blunting of affect is based on observation on facial expressions.

4.6. Further discussion on the strength and limitation of the study

4.6.1. Precise sample size calculation

Many of the previous studies on reliability between telepsychiatry interviews and face-to-face interviews did not provide detailed information on calculation of sample size. Also, the number of participants involved varied from twenty to few hundreds. In order to demonstrate both clinical and statistical significances, the sample size of this study was precisely calculated before it proceeded.

4.6.2. More comprehensive and appropriate statistical methods in calculating reliability

Many of the previous studies applying standard psychiatric assessment scales used intraclass coefficient to compare the reliability for telepsychiatry and face-to-face interview (Kobak, 2004, 2008; Baigent et al., 1997; Jones et al., 2001). However, the data collected in reliability study in health care researches is usually not normally distributed, which means that ICC is not the proper statistical measures to use. (Bobak et al., 2018; Liu et al., 2016) Instead, Wilcoxon signed-rank test and Spearman’s rho correlation coefficient should be used for data that are not normally distributed.

Many studies used either Kappa statistics or absolute agreement only in calculating the reliability for categorical data. In this study, both methods were used. Kappa statistics alone can be affected by the extreme prevalence of positive rating and the pattern of disagreement especially in data which is not normally distributed. Absolute agreement alone could not rule out the effect caused by chance agreement.

4.6.3. Keeping up with advances in information technology in this study

Studies on reliability of telepsychiatry interviews were carried out in earlier years when knowledge of computer literacy and IT was not widely available within the public domain. With the advancement in technology over the past decades, the cost of using advanced IT has reduced while the quality has improved. The general public has ample chances to engage and use computers and related information technology. It is therefore considered necessary to carry out the reliability study again as technology, an important factor on the reliability of telepsychiatry, is constantly developing and becomes more available to all walks of life. It is expected that the attitudes and perception towards telepsychiatry interview to be different when compared with the past decades.

4.6.4. Multiple standardized psychiatric assessment scales used

In this study, psychiatric assessment scales which were not used in previous telepsychiatry studies have been adopted, including C-SSRS and YMRS. Besides, a detailed result of inter-method reliability for individual items in HAM-A has also been performed, which is not attempted in other studies. The inter-method reliability of HDRS and BPRS are confirmed to have similar results comparable to previous studies after the bandwidth has been changed to 1.2 Mbps in this study.

4.6.5. Limitation in numbers of raters involved

In the study, two raters were involved in carrying out assessments. While every attempts have been implemented to minimize possible bias, it is not possible to be a zero bias situation due to different assessment styles, assessment methods and experience of the two raters. To minimize the bias issue, standardization of the raters was implemented by training, guidelines and discussion about the scoring on assessment scales. To improve the study in future, increasing number of raters would be a way to further minimize the possibility of bias during assessment and the agreement by chance.

4.6.6. Sampling bias

Under the COVID-19 epidemic the feasibility in carrying out the study in the community was constrained. Thus this study was carried out in inpatient setting. However, the symptoms presentation of participants would be different when compared to outpatient settings, emergency department and consultation liaison settings due to the hospital environment and different strategies of management and treatment plan given during inpatient setting. The participants recruited would likely present with more acute mood and psychotic symptoms that are severe and require admission. As calculation on inter-method reliability can be

affected by prevalence, the sampling bias can lead to error in calculation. It is the reason why we also include the review on raw data and calculation by actual agreement. At the same time, patients who had strong resistance towards telepsychiatry would likely refuse to participate. Therefore, the recruited participants were limited to those who at least had initial acceptance in use of telepsychiatry and this can lead to bias in data for satisfaction score. Indeed, we consider that this sampling bias resembles the actual targeted population as telepsychiatry is usually performed with those patients who consented and preferred to use in clinical practice.

4.6.7. Time lag in between assessments by two methods

Although strategies have been implemented to minimize the time gap between assessments by the two raters, it is unavoidable that there may be a change in mental state in between the two assessments which resulted in different scoring in the assessments, particular in cases where significant fluctuation of mental state is expected within a short period of time such as mania.

4.6.8. Suboptimal in translating study findings to other settings

Due to COVID-19 pandemic and related social distancing regulations, it was not feasible to carry out the study in community setting. Therefore, we decided to carry out the study in inpatient setting. We tried to mimic community setting by providing iPad which the local patients are familiar with. At the same time, we allowed participants to control the iPad themselves and reduce any unnecessary interference from clinical staff in a private room. Although we try to mimic the setting as much as possible, it would be difficult to completely replicate other settings such as outpatient clinics, consultation liaison setting and community outreach program. Therefore, further research on different settings would be necessary to confirm the result before applying in other clinical settings.

4.6.9. Setting to mimic psychiatric interview

Common standards assessment scales were chosen and Semi-structured interview was used with an attempt to mimic clinical psychiatric interviews. However, we understand that the standardized assessment scales would be used only in very limited occasions in usual clinical practice in Hong Kong. This could limit how the current results reflect on a clinical interview in usual practice. Further researches by assessing the change on mental state or symptoms could be carried out to assess the clinical effect of telepsychiatry.

5. Conclusion

To conclude, the study result demonstrated that assessments on depressive, anxiety and psychotic symptoms and suicidal risk by telepsychiatry have good reliability in general when compared with face-to-face interview. The inter-method reliability of items of mild agitation and retardation, emotional withdrawal, tension and blunting of affect which are observation-based are lower in telepsychiatry.

Declaration of Competing Interest

I declare that the study and manuscript represent my own work. It is part of my dissertation for completing the HKCPsych Part III Examination. Dr. Chi Wing Law is my supervisor. Dr. Yeung Wai Tat and I conducted all the clinical interviews and collected the clinical data from the assessment scales, questionnaires and the Hospital Authority Clinical Management System for all subjects. Diagnosis of the diseases were confirmed by team meetings with consultants in individual teams and documented in clinical records. Statistical analysis was conducted by me with the assistance by Mr. Edward Choi and I held several meetings with him in between to confirm on the relevant statistical methods used. Tables and figures related to statistical analysis were done by me. As the investigator of this study, I take direct responsibility for the study design, subject recruitment, data collection, results interpretation and write up of the dissertation.

Acknowledgments

I have benefited greatly from the mentoring of Dr. Law Chi Wing and the comments received from Dr. Wong Ming Cheuk Michael, Dr Chang Wing Chung, Dr. Wong Sze Nga, Dr. Lee Chi Kei, Dr. Cheng Pak Wing and Dr. Chan Wai Chi when they reviewed the dissertation. The clinical staff in Queen Mary Hospital provide numerous assistances during the subject recruitment process. Finally, this study would not have been possible without the participation of patients.

Appendix

Result of core analysis of Hamilton Depression Rating Scale.

Hamilton Depression Rating Scale	
Items	Weighted Kappa / Kappa statistics
Depressed mood	0.82
Feeling of guilt	0.71
Suicide	0.95
Insomnia: early in the night	0.69
Insomnia: middle of the night	0.63
Insomnia: early hours of the morning	0.75
Work and activities	0.76
Retardation	0.67
Agitation	0.42
Anxiety psychic	0.68
Anxiety somatic	0.68
Somatic symptoms gastro-intestinal	0.7
General somatic symptoms	0.61
genital symptoms	0.62
Hypochondriasis	−0.015
Lost of weight	0.76
insight	0.79

Result of core analysis of Hamilton Anxiety Rating Scale.

Items	Weighted Kappa / Kappa statistics
Anxious mood	0.73
Tension	0.8
Fears	0.52
Insomnia	0.77
Intellectual	0.78
Depressed mood	0.83
Somatic (muscular)	0.65
Somatic (sensory)	0.78
cardiovascular symptoms	0.85
Respiratory symptoms	0.81
gastrointerstinal symptoms	0.85
genitourinary symptoms	0.81
Autonomic symptoms	0.61
Behavior at interview	0.62

Result of core analysis of Young Mania Rating Scale.

Young Mania Rating Scale	
Items	Weighted Kappa / Kappa statistics
Elevated mood	0.96
Increased motor activity energy	0.72
Sexual interest	0.61
Sleep	0.89
Irritability	0.89
Speech	0.94
Language-thought disorder	0.82
Content	0.96
Disruptive- aggressive behaviors	1
Appearance	0.74
Insight	0.94

Result of core analysis of Brief Psychiatric Rating Scale

Brief Psychiatric Rating Scale		
Items	Wilcoxon Signed-rank test	Spearman's rho correlation coefficient
Emotional withdrawal	$Z=-0.182$, $p = 0.856$	$r = 0.569$, $p<0.001$
Conceptual disorganization	$Z=-0.997$, $p = 0.319$	$r = 0.81$, $p<0.001$
Tension	$Z=-0.131$, $p = 0.896$	$r = 0.69$, $p<0.001$
Mannerism and posturing	$Z<0.0001$, $p = 1$	$r = 1$
Grandiosity	$Z=-0.122$, $p = 0.903$	$r = 0.795$, $p<0.001$
Hostility	$Z=-0.447$, $p = 0.655$	$r = 0.577$, $p<0.001$
Suspiciousness	$Z=-0.437$, $p = 0.662$	$r = 0.615$, $p<0.001$
Hallucinatory behaviors	$Z=-0.272$, $p = 0.785$	$r = 0.912$, $p<0.001$
Motor retardation	$Z=-0.489$, $p = 0.625$	$r = 0.695$, $p<0.001$
Uncooperativeness	$Z=-1$, $p = 0.317$	$r = 1$
Unusual thought content	$Z=-1.63$, $p = 0.103$	$r = 0.937$, $p<0.001$
Blunt affect	$Z=-2.078$, $p = 0.038$	$r = 0.908$, $p<0.001$
Excitement	$Z=-2$, $p = 0.317$	$r = 0.924$, $p<0.001$
Disorientation	$Z=-0.378$,$p = 0.705$	$r = 0.999$, $p<0.001$

References

Amarendran, V., George, A., Gersappe, V., Krishnaswamy, S., Warren, C., 2011. The reliability of telepsychiatry for a neuropsychiatric assessment. Telemed. e-Health 17 (3), 223–225. https://doi.org/10.1089/tmj.2010.0144.

Baer, L., Cukor, P., Jenike, M.A., Leahy, L., O'Laughlen, J., Coyle, J.T., 1995. Pilot studies of telemedicine for patients with obsessive-compulsive disorder. Am. J. Psychiatry. https://doi.org/10.1176/ajp.152.9.1383.

Baigent, M.F., Lloyd, C.J., Kavanagh, S.J., Ben-Tovim, D.I., Yellowlees, P.M., Kalucy, R. S., Bond, M.J., 1997. Telepsychiatry: 'tele' yes, but what about the 'psychiatry'? J. Telemed. Telecare 3 (1_suppl), 3–5. https://doi.org/10.1258/1357633971930346.

Bobak, C.A., Barr, P.J., O'Malley, A.J, 2018. Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. BMC Med. Res. Methodol. 18 (1), 1–11. https://doi.org/10.1186/s12874-018-0550-6.

Bujang, M.A., Baharum, N., 2017. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. Arch. Orofac. Sci. 12 (1) https://doi.org/10.1002/sim.1108.

Cheli, S., Cavalletti, V., Petrocchi, N., 2020. An online compassion-focused crisis intervention during COVID-19 lockdown: a cases series on patients at high risk for psychosis. Psychosis 12 (4), 359–362. https://doi.org/10.1080/17522439.2020.1786148.

Cheng, I. (2021, January 21). Development of telehealth services. Retrieved January 01, 2022, from https://www.legco.gov.hk/research-publications/english/essentials-2021ise14-development-of-telehealth-services.htm.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284. https://doi.org/10.1037/1040-3590.6.4.284.

Craig, J.J., McConville, J.P., Patterson, V.H., Wootton, R., 1999. Neurological examination is possible using telemedicine. J. Telemed. Telecare 5 (3), 177–181. https://doi.org/10.1258/1357633991933594.

Dores, A.R., Geraldo, A., Carvalho, I.P., Barbosa, F., 2020. The use of new digital information and communication technologies in psychological counseling during the

COVID-19 pandemic. Int. J. Environ. Res. Public Health 17 (20), 7663. https://doi.org/10.3390/ijerph17207663.
Dwyer, T.F., 1973. Telepsychiatry: psychiatric consultation by interactive television. Am. J. Psychiatry 130 (8), 865–869. https://doi.org/10.1176/ajp.130.8.865.
Elford, R., White, H., Bowering, R., Ghandi, A., Madiggan, B., John, K.S., 2000. A randomized, controlled trial of child psychiatric assessments conducted using videoconferencing. J. Telemed. Telecare 6 (2), 73–82. https://doi.org/10.1258/1357633001935086.
Erekson, D.M., Bailey, R.J., Cattani, K., Fox, S.T., Goates-Jones, M.K., 2020. Responding to the Covid-19 pandemic at a university counseling center: administrative actions, client retention, and psychotherapy outcome. Couns. Psychol. Q. 34 (3–4), 729–743. https://doi.org/10.1080/09515070.2020.1807914.
Frueh, B.C., Deitsch, S.E., Santos, A.B., Gold, P.B., Johnson, M.R., Meisler, N., Magruder, K.M., Ballenger, J.C., 2000. Procedural and methodological issues in telepsychiatry research and program development. Psychiatr. Serv. 51 (12), 1522–1527. https://doi.org/10.1176/appi.ps.51.12.1522.
Hamilton, M., 1960. A rating scale for depression. J. Neurol. Neurosurg. Psychiatr. 23 (1), 56. https://doi.org/10.1136/jnnp.23.1.56.
Hamilton, M.A.X., 1959. The assessment of anxiety states by rating. Br. J. Med. Psychol. 32 (1), 50–55. https://doi.org/10.1111/j.2044-8341.1959.tb00467.x.
Hubley, S., Lynch, S.B., Schneck, C., Thomas, M., Shore, J., 2016. Review of key telepsychiatry outcomes. World J. Psychiatry 6 (2), 269. https://doi.org/10.5498/wjp.v6.i2.269.
Hyler, S.E., Gangure, D.P., Batchelder, S.T., 2005. Can telepsychiatry replace in-person psychiatric assessments? A review and meta-analysis of comparison studies. CNS Spectr. 10 (5), 403–415. https://doi.org/10.1017/s109285290002277x.
Jones, B.N., Johnston, D., Reboussin, B., McCall, W.V., 2001. Reliability of telepsychiatry assessments: subjective versus observational ratings. J. Geriatr. Psychiatry Neurol. 14 (2), 66–71. https://doi.org/10.1177/089198870101400204.
Kirkwood, K.T., Peck, D.F., Bennie, L., 2000. The consistency of neuropsychological assessments performed via telecommunication and face to face. J. Telemed. Telecare 6 (3), 147–151. https://doi.org/10.1258/1357633001935239.
Kobak, K.A., 2004. A comparison of face-to-face and videoconference administration of the Hamilton depression rating scale. J. Telemed. Telecare 10 (4), 231–235. https://doi.org/10.1258/1357633041424368.
Kobak, K.A., Williams, J.B., Engelhardt, N., 2008. A comparison of face-to-face and remote assessment of inter-rater reliability on the Hamilton depression rating scale via videoconferencing. Psychiatry Res. 158 (1), 99–103. https://doi.org/10.1016/j.psychres.2007.06.025.
Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.
Liu, J., Tang, W., Chen, G., Lu, Y., Feng, C., Tu, X.M., 2016. Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Arch. Psychiatry 28 (2), 115–120. https://doi.org/10.11919/j.issn.1002-0829.216045.
Lynch, D.A., Medalia, A., Saperstein, A., 2020. The design, implementation, and acceptability of a telehealth comprehensive recovery service for people with complex psychosis living in NYC during the COVID-19 crisis. Front. Psychiatry 11. https://doi.org/10.3389/fpsyt.2020.581149.
Madan, A., Frueh, B.C., Allen, J.G., Ellis, T.E., Rufino, K.A., Oldham, J.M., Fowler, J.C., 2016. Psychometric reevaluation of the Columbia–suicide severity rating scale. J. Clin. Psychiatry 77 (07). https://doi.org/10.4088/jcp.15m10069.
Matsuura, S., Hosaka, T., Yukiyama, T., Ogushi, Y., Okada, Y., Haruki, Y., Nakamura, M., 2000. Application of telepsychiatry: a preliminary study. Psychiatry Clin. Neurosci. 54 (1), 55–58. https://doi.org/10.1046/j.1440-1819.2000.00637.x.
Menon, A.S., Kondapavalru, P., Krishna, P., Chrismer, J.B., Raskin, A., Hebel, J.R., Ruskin, P.E., 2001. Evaluation of a portable low cost videophone system in the assessment of depressive symptoms and cognitive function in elderly medically ill veterans. J. Nerv. Ment. Dis. 189 (6), 399–401. https://doi.org/10.1097/00005053-200106000-00009.
Monnier, J., Knapp, R.G., Frueh, B.C., 2003. Recent advances in telepsychiatry: an updated review. Psychiatr. Serv. 54 (12), 1604–1609. https://doi.org/10.1176/appi.ps.54.12.1604.
Montani, C., Billaud, N., Couturier, P., Fluchaire, I., Lemaire, R., Malterre, C., Lauvernay, N., Piquard, J., Frossard, M., Franco, A., 1996. Telepsychometry": a remote psychometry consultation in clinical gerontology: preliminary study. Telemed. J. 2 (2), 145–150. https://doi.org/10.1089/tmj.1.1996.2.145.
Montani, C., Billaud, N., Tyrrell, J., Fluchaire, I., Malterre, C., Lauvernay, N., Couturier, P., Franco, A., 1997. Psychological impact of a remote psychometric consultation with hospitalized elderly people. J. Telemed. Telecare 3 (3), 140–145. https://doi.org/10.1258/1357633971931048.
Office of the Communications Authority in Hong Kong (2021, July). Key communications statistics Retrieved August 28, 2021 from https://www.ofca.gov.hk/en/media_focus/data_statistics/key_stat/.
Overall, J.E., Gorham, D.R., 1962. The brief psychiatric rating scale. Psychol. Rep. 10 (3), 799–812. https://doi.org/10.2466/pr0.1962.10.3.799.
Overall, J.E., Gorham, D.R., 1988. The brief psychiatric rating scale (BPRS): recent developments in ascertainment and scaling. Psychopharmacol. Bull. https://doi.org/10.1007/978-3-319-56782-2_1976-2.
Posner, K., Brown, G.K., Stanley, B., Brent, D.A., Yershova, K.V., Oquendo, M.A., Currier, G.W., Melvin, G.A., Greenhill, L., Shen, S., Mann, J.J., 2011. The Columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am. J. Psychiatry 168 (12), 1266–1277. https://doi.org/10.1176/appi.ajp.2011.10111704.
Ruskin, P.E., Reed, S., Kumar, R., Kling, M.A., Siegel, E., Rosen, M., Hauser, P., 1998. Reliability and acceptability of psychiatric diagnosis via telecommunication and audiovisual technology. Psychiatr. Serv. 49 (8), 1086–1088. https://doi.org/10.1176/ps.49.8.1086.
Seidel, R.W., Kilgus, M.D., 2014. Agreement between telepsychiatry assessment and face-to-face assessment for emergency department psychiatry patients. J. Telemed. Telecare 20 (2), 59–62. https://doi.org/10.1177/1357633x13519902.
Sequeira, A., Alozie, A., Fasteau, M., Lopez, A.K., Sy, J., Turner, K.A., Werner, C., McIngvale, E., Björgvinsson, T., 2020. Transitioning to virtual programming amidst COVID-19 outbreak. Couns. Psychol. Q. 34 (3–4), 538–553. https://doi.org/10.1080/09515070.2020.1777940.
Sim, J., Wright, C.C., 2005. The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. 85 (3), 257–268. https://doi.org/10.1093/ptj/85.3.257.
Singh, S.P., Arya, D., Peters, T., 2007. Accuracy of telepsychiatric assessment of new routine outpatient referrals. BMC Psychiatry 7 (1). https://doi.org/10.1186/1471-244x-7-55.
Timpano, F., Pirrotta, F., Bonanno, L., Marino, S., Marra, A., Bramanti, P., Lanzafame, P., 2013. Videoconference-based mini mental state examination: a validation study. Telemed. J. E-Health 19 (12), 931–937. https://doi.org/10.1089/tmj.2013.0035.
Turner, T.H., Horner, M.D., VanKirk, K.K., Myrick, H., Tuerk, P.W., 2012. A pilot trial of neuropsychological evaluations conducted via telemedicine in the Veterans Health Administration. Telemed. e-Health 18 (9), 662–667.
Viera, A.J., Garrett, J.M., 2005. Understanding interobserver agreement: the kappa statistic. Fam. Med. 37 (5), 360–363.
Wyler, H., Liebrenz, M., Ajdacic-Gross, V., Seifritz, E., Young, S., Burger, P., Buadze, A., 2021. Treatment provision for adults with ADHD during the COVID-19 pandemic: an exploratory study on patient and therapist experience with on-site sessions using face masks vs. telepsychiatric sessions. BMC Psychiatry 21 (1). https://doi.org/10.1186/s12888-021-03236-9.
Yoshino, A., Shigemura, J., Kobayashi, Y., Nomura, S., Shishikura, K., Den, R., Wakisaka, H., Kamata, S., Ashida, H., 2001. Telepsychiatry: assessment of televideo psychiatric interview reliability with present- and next-generation internet infrastructures. Acta Psychiatr. Scand. 104 (3), 223–226. https://doi.org/10.1034/j.1600-0447.2001.00236.x.
Young, R.C., Biggs, J.T., Ziegler, V.E., Meyer, D.A., 1978. A rating scale for mania: reliability, validity and sensitivity. Br. J. Psychiatry 133 (5), 429–435. https://doi.org/10.1192/bjp.133.5.429.
Zarate, C.A., Weinstock, L., Cukor, P., Morabito, C., Leahy, L., Burns, C., Baer, L., 1997. Applicability of telemedicine for assessing patients with schizophrenia. J. Clin. Psychiatry 58 (1), 22–25. https://doi.org/10.4088/jcp.v58n0104.