Soc Psychiatry Psychiatr Epidemiol (2014) 49:1677–1689
DOI 10.1007/s00127-014-0861-9

REVIEW

Psychiatric diagnosis by telephone: is it an opportunity?

Esther M. H. Muskens · Peter Lucassen · Wilke Groenleer · Chris van Weel · Richard Oude Voshaar · Anne Speckens

Received: 12 December 2013 / Accepted: 27 February 2014 / Published online: 15 March 2014

Abstract
Background For reasons of feasibility, diagnostic telephone interviews are frequently used in research of psychiatric morbidity. However, it is unknown whether diagnostic telephone interviews are as valid as diagnostic face-to-face interviews.
Research question Are diagnostic telephone interviews for psychiatric disorders as valid as diagnostic face-to-face interviews?
Method A systematic review of original studies in PubMed, PsychINFO and Embase was carried out. We included studies considering (1) the sensitivity and specificity of diagnostic telephone interviews using face-to-face interviews as a golden standard and (2) the agreement between diagnostic telephone and diagnostic face-to-face interviews. Eligible were studies in the general population, in patients at risk for psychiatric disorders and in psychiatric outpatients. We assessed risk of bias with the quality assessment of diagnostic accuracy studies (QUADAS) instrument.
Results We included sixteen studies. The included studies were generally small with thirteen studies reporting about <100 participants. Specificity was generally high in populations with low or intermediate prevalence of psychiatric morbidity. Sensitivity was low in these populations, but slightly higher in samples with more psychiatric disorders. Studies with a higher risk of psychiatric disorders generally reported higher percentages of agreement and higher kappa values. Considering the QUADAS-2 criteria, most studies had a medium or high risk of bias, especially concerning patient selection and unbiased judgement of the test. Of the six studies with a medium or low risk of bias, the three studies assessing current anxiety and depressive disorders yielded kappa values between 0.69 and 0.84, indicating good agreement.
Discussion 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.

Keywords Depression · Anxiety · Diagnosis · Telephone interview · Face-to-face interview

E. M. H. Muskens (✉) · P. Lucassen · W. Groenleer · C. van Weel
117 Department of Primary and Community Care, Radboud University Nijmegen Medical Centre, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands
e-mail: esther.muskens@gmail.com

C. van Weel
Australian Primary Health Care Research Institute, Australian National University, Canberra, Australia

R. Oude Voshaar
University Center for Psychiatry and Interdisciplinary Center for Psychopathology of Emotion Regulation (ICPE), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

A. Speckens
Department of Psychiatry, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Introduction

In psychiatric research projects a diagnosis is important for the selection of participants and as an outcome measure. To obtain a sample of participating patients who fulfill the criteria for the condition under study, or to assess the outcome, is demanding because of the length of the necessary psychiatric interview. Up until the 1970s, these

interviews were mainly face-to-face. Telephone interviews as alternative method were hardly ever mentioned in textbooks on survey methods [1], and they were seen as inferior to face-to-face interviews [2]. Researchers assumed that telephone interviews should be short and that they were only suitable for gathering factual data and not for more sensitive issues [1, 3]. The main advantage of telephone research is obvious: the low cost rate compared to face-to-face interviews [2, 4–6], which are about twice as expensive [1, 3]. Another advantage could be more control over the interview process [3, 4, 7, 8] thus decreasing interviewer influence [2]. The obvious drawback of the telephone interview is the lack of visual signs, which may be a cause of missing important diagnostic cues [2].

Telephone interviews in general show more compliance or acquiescence (yes-saying), evasiveness ("I don't know" answers, or no response at all) and more extreme responses compared with the face-to-face interviews [2, 3, 9–11]. Also, respondents tend to give more information in face-to-face interviews, especially following open-ended questions [2, 4, 7, 10]. Telephone interviews may be less suitable for people who are hearing impaired [2, 3, 10], mistrustful [8, 12], older, [3, 7, 10, 13] or very ill [3, 7]. The same goes for people from minorities or lower socioeconomic class [3, 12] and for people with lower education [3, 4, 7, 10, 11].

A systematic review comparing telephone and face-to-face interview for a specific psychiatric disorder—depression—showed a good comparability for the two methods, but the authors stated that the study quality was generally low [14]. There are, as far as we know, no reviews for psychiatric disorders in general. An important question is therefore, how valid telephone interviews are for psychiatric diagnosis in comparison with face-to-face interviews. This study reviews the value of telephone-administered standardized psychiatric diagnostic interviews from the following perspectives: (1) sensitivity and specificity of telephone interviews using face-to-face interviews as the golden standard and (2) agreement between telephone interviews and face-to-face interviews.

Methods

We performed a systematic review of the available literature in PubMed, PsychINFO and Embase, examining the value of telephone interviews in providing a psychiatric diagnosis as compared to face-to-face interviews.

Search strategy

In June 2012, we systematically searched for publications with a comparison between telephone and face-to-face diagnostic interviewing. We did not restrict our search by language or by age of participants. An academic reference librarian was consulted to ensure that search strategies and relevant articles were not overlooked.

We searched in three databases: PubMed, PsychINFO and EMBASE. For PubMed our search consisted of the All Fields and MeSH terms for "mental disorder(s)," "Diagnostic and Statistical Manual of Mental Disorders," "psychiatry," "psychiatric," "bipolar disorder(s)," "anxiety disorder(s)," "depressive disorder(s)" or "depression(s)," AND "interview(s)," "psychological," "interviewing," "Interviews as Topic," "telephone-administered," "face to face," "questionnaires," "diagnosis," "diagnoses," "diagnostic," "assessment," "measuring," "telephone" or phone (the complete search string for PubMed is shown in "Appendix"). We adapted the search for the other databases as required.

Selection of publications

For inclusion, we screened titles and abstracts. When title and abstract did not reveal sufficient information for inclusion or exclusion, the investigators read the full-text publication. Two investigators (EM, WG) independently selected publications from the list of retrieved publications. Disagreements about inclusion or exclusion were resolved by consulting a third investigator (PL). Inter-rater reliability on inclusion and exclusion was calculated as kappa; we considered kappa 0.6–0.8 as good and kappa 0.8–1.0 as excellent agreement [15]. After inclusion, we checked the references for additional publications.

To be included in the selection, studies had to be original studies comparing telephone and face-to-face interviews using the same standardized diagnostic criteria for a mental health problem. Each patient had to be subjected to both modes of interviewing. Studies were included that considered [1] the comparison between telephone and face-to-face interviewing as a criterion validity issue with face-to-face interviewing as the gold standard and [2] the agreement between the two methods. Agreement is based on all items of the questionnaire.

We excluded (1) studies with interviews about topics outside the field of mental health, [2] studies with non-standardized psychiatric interviews, (3) non-diagnostic interviews, (4) studies using different diagnostic interviews by telephone than face-to-face, (5) studies using different respondents for the two interview methods, (6) interviews using interactive voice response and (7) studies comparing scores of the two instruments with statistical testing or ICC values, as these studies did not determine whether a diagnosis was present or not.

Outcome assessment

We ranked the outcomes of the selected studies according to the risk of psychiatric morbidity. So, we considered studies in the general population as having a low risk of psychiatric morbidity, studies in general practice and studies with patients with risk factors as having intermediate risk, and studies in outpatients of psychiatric hospitals as having a high risk of psychiatric morbidity (Table 1). For the outcome assessment of the selected studies, we examined the sensitivity, specificity, percentage agreement and kappa values. Sensitivity is the proportion of true positives that are correctly identified by the test. Specificity is the proportion of true negatives that are correctly identified by the test. In general, the higher the sensitivity, the lower the specificity and vice versa [16]. Percentage agreement is defined as the extent to which the outcomes of the telephone and face-to-face interview agree with each other [17]. Kappa is a measure of reliability in which the agreement between two observers or two assessment methods is calculated, corrected for chance. A kappa of 0 means that the agreement rests fully on chance, a kappa of one means perfect agreement [18].

Quality assessment

We used the QUADAS-2 (quality assessment of diagnostic accuracy studies) tool to estimate the risk of bias in individual studies [19]. The use of this tool is recommended in systematic reviews of diagnostic accuracy by the Agency for Healthcare Research and Quality, Cochrane Collaboration and the U.K. National Institute for Health and Clinical Excellence. To estimate the risk of bias, the QUADAS-2 tool distinguishes four key domains that have to be rated: "patient selection" [question 1–3 (1) Was a consecutive or random sample of patients enrolled?, (2) Was a case–control design avoided?, (3) Did the study avoid inappropriate exclusions?], "index test" [question 4–5 (4) Were the index test results interpreted without knowledge of the results of the reference standard?, (5) If a threshold was used, was it pre-specified?], "reference standard" [question 6–7 (6) Is the reference standard likely to correctly classify the target condition?, (7) Were the reference standard results interpreted without knowledge of the results of the index test?] and "flow and timing" [question 8–11: (8) Was there an appropriate interval between index test(s) and reference standard?, (9) Did all patients receive a reference standard?, (10) Did patients receive the same reference standard?, (11) Were all patients included in the analysis?). We chose not to rank the included studies with numerical scores because quality scores have been shown to produce different results depending on how the individual items are weighted [20]. Two researchers (EM, WG) independently scored the risk of bias. Disagreements were resolved by consulting a third researcher (PL) [19].

Data extraction

Data extraction was performed independently by two researchers (EM, WG). For the construction of the data extraction form, we used the items of the STARD statement (Standards for Reporting of Diagnostic Accuracy) [21]. The items relevant for the quality assessment according to the QUADAS-2 tool [19] could be derived from this data extraction procedure.

Results

Selection of publications

Our database search retrieved 3,042 publications. We found six additional articles, four by checking the references of the retrieved articles and two on internet. After removing the duplicates, 1,879 publications remained to be screened (Fig. 1, flowchart). Applying the exclusion criteria on the title and abstract of these 1,879 publications resulted in the selection of 41 citations. The inter-investigator agreement was "good" with a kappa of 0.77 (95 % CI 0.71–0.83). Definite assessment of the full text of the 41 citations resulted in the exclusion of 25 studies, leaving 16 studies to be included.

Description of selected studies

The included studies were generally small with 13 studies reporting about less than 100 participants (Table 1). Many different instruments had been used. Studies using standardized psychiatric interviews (SCID [22], DIS [23] and CIDI [24]) frequently used only one diagnostic section. There was also a large heterogeneity concerning the age and psychiatric morbidity of the included participants. Most studies reported on outpatients visiting specialized clinics. The number of psychiatric disorders addressed in individual studies ranged from one to 21. Several small studies addressed a large range of disorders [25–28]. Two studies examined general population samples [26, 27], four studies examined samples with an intermediate risk of psychiatric disorder [28–31] and the remaining 10 studies examined high-risk samples with psychiatric outpatients [25, 32–40] (Table 2). Four studies used semi structured interviews; the outcomes did not differ from the outcomes of studies with structured psychiatric interviews (Table 3). The time between telephone and face-to-face interview did not influence the outcomes (Table 4). Finally, there were

Table 1 Included studies

Study	Population	*N*	Mean age; (M/F)	Test	Outcomes	Sensitivity	Specificity	Kappa	% Agreement	Risk of bias
Watson [25]	Community volunteers recruited from acquaintances	49	40, 4 years (22/27)	DIS	5 substance use disorders	Inconclusive	98 % or higher	0.92		High
					8 anxiety disorders			0.62		
					Affective disorders			Below 0.20, except dysthimic disorder 0.76		
Cacciola [26]	Convenience sample from a larger longitudinal study of college-aged men with or without a history of paternal alcoholism	41	21.9 years (41/0)	SCID	Lifetime diagnoses: major depression	57.1 %	100 %	0.64	57.1	High
					Panic disorder	0 %	100 %	–	0	
					Social phobia	25 %	97.3 %	0.29	20.0	
					Simple phobia	50 %	97.4 %	0.47	33.3	
					any disorder	62.5 %	96.0 %	0.62	58.8	
					Current diagnoses: major depression	50 %	100 %	0.66	50.0	
					Panic disorder	0 %	100 %	–	0	
					Social phobia	0 %	100 %	–	0	
					Simple phobia	0 %	97.5 %	0.03	0	
					Generalized anxiety disorder	0 %	95.1 %	–	0	
					Any disorder	28.6 %	94.1 %	0.27	22.2	
Wells [32]	Random sample of general population stratified by the presence or absence of two indicators of	230	39 years (125/105)	DIS depression section	Lifetime major depression	56 %	89 %	0.45		Medium
					Lifetime dysthimia	55 %	95 %	0.48		
					Lifetime MDD and/or dysthimia	71 %	89 %	0.57		
Crippa [31]	Volunteering undergraduate students who were screened with MS (a screening instrument for	100	20.8 years (37/63)	SCID social phobia module	Social anxiety disorder			0.84		Medium

Table 1 continued

Study	Population	*N*	Mean age; (M/F)	Test	Outcomes	Sensitivity	Specificity	Kappa	% Agreement	Risk of bias
Evans [33]	General practice sample of consecutive attenders	98	51 years (31/67)	GHQ-12 CI5-R	Psychiatric caseness			0.75		High
					Common mental disorders			0.72		
Paulsen [27]	Members of families of probands with panic disorder or agoraphobia with panic attacks	39	nm	SADS L	Panic disorder			0.81	92	Medium
					Agoraphobia with panic			0.69	90	
					Agoraphobia without panic			nm	100	
					Major depression			0.69	90	
					Alcoholism			0.84	97	
					No mental disorder			0.69	85	
Burke [37]	Consecutive elderly outpatients referred for the evaluation of cognitive deficits or unexplained	83	76,9 years (54/29)	CS-GDS	Depression (cutoff point 14)	94 %	42 %			Low
					Depression (cutoff point 19)	61 %	58 %			
Lyneham [40]	24 children from anxiety clinic, 24 children from advertisement who never sought help, 25 children from advertisement who sought help for externalizing	73	9.2 years (49/24)	ADIS-C-IV	Anxiety, mood and externalizing disorders			0.86		High
Hajebi [29]	Psychiatric outpatients, sampling directed so that approx. half of the patients had a history of lifetime	72	35.2 years (36/36)	SCID-I psychotic disorder module	Primary psychotic disorder in the past 12 months	73.7 %	67.9 %			Medium
					Primary psychotic disorder in lifetime	80.6 %	80.6 %			
					Any psychotic disorder in lifetime	86.5 %	82.9 %			

Table 1 continued

Study	Population	*N*	Mean age; (M/F)	Test	Outcomes	Sensitivity	Specificity	Kappa	% Agreement	Risk of bias
Rohde [35]	Participants from follow up study; 20 selected based on prior diagnosis of depression, 20 selected based on prior psychiatric disorder, 20 selected with no psychiatric disorder	60	24.4 years	KIDDIE-SADS	Major depressive disorder			067		High
					Anxiety			0.84		
					Alcohol use			0.70		
					Substance use			0.73		
					adjustment disorder with depressed mood			0.31		
Aziz [30]	Outpatients, male veterans seeking help for PTSD	34	54 years (34/0)	CAPS, HAM-D	PTSD (CAPS cut off point 60)	86 %	69 %	0.72	82	High
					PTSD (CAPS cut off point 65)	84 %	80 %	0.75 %		
					Depression	79 %	100 %	0.70	85	
Simon [36]	Consecutive outpatients starting with antidepressant	31	40 years (7/24)	SCID	Current major depression			0.73	90	Low
Revicki [34]	Outpatients with bipolar disorder	30	36,3 years (13/17)	CIDI for mania and major depression	Mania			0.78		Medium
					Major depression			1.00		
					Alcohol			0.80		
Tunstall [38]	Outpatients from geriatric and psycho geriatric day hospitals	29	15 subjects 80.3 years (7/8) 14 subjects 77.4 years	DDS	Depression			0.79		High
Ward-King [39]	Outpatients, children previously diagnosed as	20 children and their parents	8.92 years (14/6)	ADI-R	Autism				No data presented about agreement	High
Paing [28]	Convenience sample of referred parents of children with suspected psychiatric	12 parents	12.2 years (7/5)	P-ChlPS	21 psychiatric disorders				Range from 75 to 100 %. Mean agreement 93.8 %	High

DIS Diagnostic interview schedules, *KIDDIE SADS* Schedule for Affective Disorders and Schizophrenia for school-aged children, *GHQ-12* General Health Questionnaire, *CIS-R* Clinical interview schedule revised, *SCID* Structured clinical interview for DSM disorders, *CS-GDS* Collateral source version of the Geriatric Depression Scale, *ADIS-CIV* Anxiety disorders interview schedule for children for DSM-IV, *SADS-L* Schedule of affective disorders and schizophrenia-lifetime version, *CIDI* Composite international diagnostic interview, *PRIME-MD* Primary care evaluation of mental disorders, *CAPS* Clinician-Administered PTSD Scale, *HAM-D* Hamilton Rating Scale for depression, *DDS* Depression Diagnostic Scale, *ADI-R* Autism diagnostic interview-revised, *GDS* Geriatric Depression Scale, *P-ChIPS* Parent's version of the children's interview for psychiatric syndromes, *nm* not mentioned

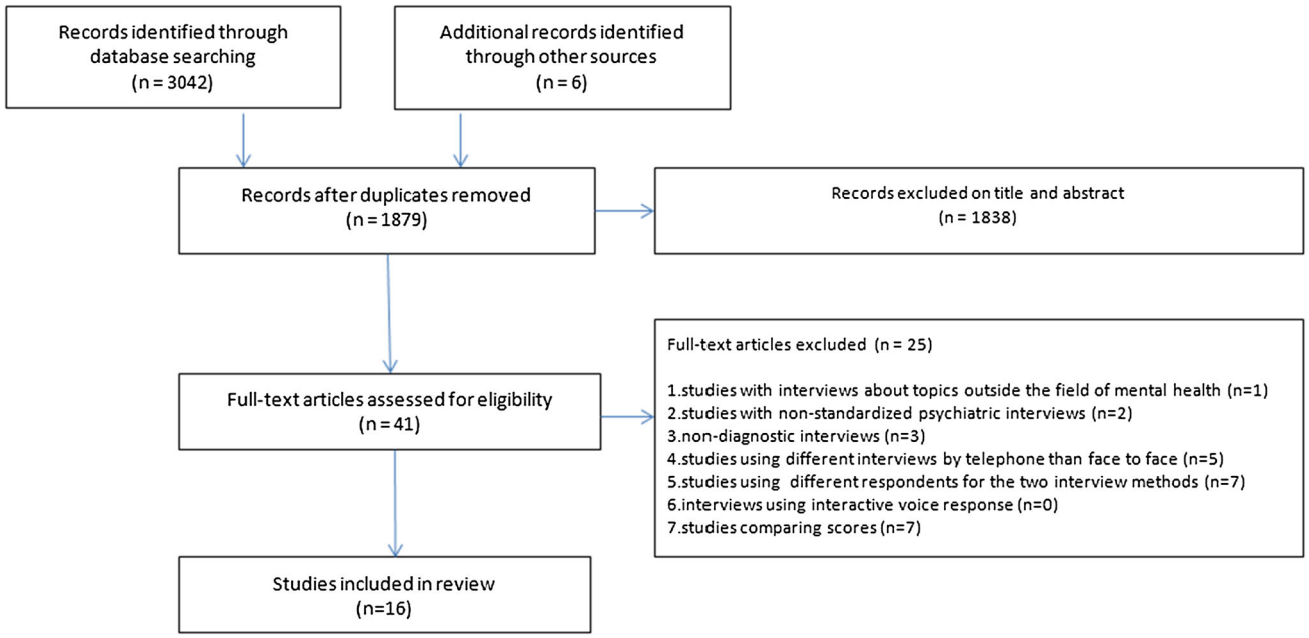

Fig. 1 Flowchart

no differences between outcomes from interviews by trained lay interviewers and interviews by professionals (Table 5).

Sensitivity and specificity

The two studies in samples with a *low risk* of psychiatric morbidity [25, 26] mainly aimed at diagnosing depressive and anxiety disorders. The study of Cacciola [26] with 41 respondents found a specificity of 94.1 % for any disorder. The study of Watson [25] with 49 respondents found a specificity of 98 % or higher for substance use disorders. Sensitivity was low in both studies (Table 1). From the four studies with *intermediate risk* of psychiatric morbidity [27, 31–33] only the study by Wells [32] with 230 patients provided data about criterion validity. They found high specificities for lifetime major depression (89 %), lifetime dysthymia (95 %) and lifetime MDD and/or dysthymia (89 %). Sensitivity was 55 % for lifetime dysthymia, 56 % for lifetime major depression and 71 % for the combination of both disorders. From the remaining 10 studies with a *high risk* of psychiatric morbidity, three studies provided data about criterion validity [25, 37]. Hajebi [29] assessed 72 outpatients with the SCID psychotic disorder module. Sensitivity and specificity were 86.5 and 82.9 % for any psychotic disorder in lifetime, 80.6 and 80.6 % for primary psychotic disorder in lifetime, and 73.3 and 67.9 % for primary psychotic disorder in the past 12 months, respectively. Aziz [30] tested the CAPS for detection of PTSD and the HAM-D for depression in 34 outpatients. The sensitivity and specificity for CAPS 65 84 and 80 %, and for HAM-D 79 and 100 %. Burke [37] assessed the criterion validity of a version of the Geriatric Depression Scale in 83 elderly outpatients. They used a cutoff point of 14. Specificity was 42 % and sensitivity 94 %.

Agreement

From the studies with *low risk* of psychiatric morbidity [26, 27], Cacciola reported low agreement and low kappa values; for any disorder, these were 22.2 % and 0.27, respectively. Watson only reported kappa values, which were generally low, with the exception of a kappa of 0.92 for substance use disorders. In the *intermediate risk* studies [28–31], one study reported about percentage agreement. Paulsen [27] found high values for percentage agreement; percentage agreement for no mental disorder (85 %) was the lowest. Kappa values in the four studies ranged between 0.45 and 0.84. Paulsen found kappa values between 0.69 (agoraphobia with panic, major depression, no mental disorder) and 0.84 (alcoholism). Evans [33] reported kappa values of 0.72 and 0.75 for common mental disorders and psychiatric caseness, respectively, in a study of general practice attendees. Crippa [31] assessed 100 volunteering undergraduate students with the SCID social phobia module and found a kappa value of 0.84; this study enriched the sample by a screening for social phobia before the study. Wells [32] found kappa values of 0.45, 0.48 and 0.57 for lifetime major depression, lifetime dysthymia, and lifetime MDD and/or dysthymia, respectively; they

Table 2 Possible sources of bias in studies

	1. Was a consecutive or random sample of patients enrolled?	2. Was a case – control design avoided?	3. Did the study avoid inappropriate exclusions?	Risk of bias patient selection	4. Were the index test results interpreted without knowledge of the results of the reference standard?	5. If a threshold was used, was it pre-specified?*	Risk of bias index test	6. Is the reference standard likely to correctly classify the target condition?	7. Were the reference standard results interpreted without knowledge of the results of the index test?	Risk of bias reference standard	8. Was there an appropriate interval between index test(s) and reference standard?	9. Did all patients receive a reference standard?	10. Did patients receive the same reference standard?	11. Were all patients included in the analysis?	Risk of bias flow and timing	Total risk bias
Watson [25]	–	+	+	Medium	–	+	High	+	–	High	+	+	+	+	Low	High
Cacciola [26]	–	–	+	High	+	+	Low	+	+	Low	+	+	+	+	Low	High
Wells [32]	+	–	+	Medium	?	+	Unclear	+	+	Low	+	+	+	–	Low	Medium
Crippa [31]	+	–	+	Medium	+	+	Low	+	+	Low	?	+	+	+	Low	Medium
Evans [33]	+	+	+	Low	–	+	High	+	–	High	+	+	+	?	Low	High
Paulsen [27]	–	+	+	Medium	+	+	Low	+	+	Low	–	+	+	?	Unclear	Medium
Burke [37]	+	+	+	Low	+	+	Low	+	+	Low	+	+	+	–	Low	Low
Lyneham [40]	?	–	?	High	+	+	Low	+	+	Low	+	+	+	–	Low	High
Hajebi [29]	+	_	+	Medium	+	+	Low	+	+	Low	+	+	+	+	Low	Medium
Rohde [35]	–	–	?	High	?	+	Unclear	+	?	Unclear	+	+	+	+	Low	High
Aziz [30]	–	?	+	High	+	+	Low	+	+	Low	+	+	+	+	Low	High
Simon [36]	+	+	+	Low	+	+	Low	+	+	Low	+	+	+	–	Low	Low
Revicki [34]	?	+	+	Medium	?	+	Unclear	+	?	Unclear	+	+	+	?	Low	High
Tunstall [38]	–	–	+	High	?	+	Unclear	+	?	Unclear	+	+	+	?	Low	High
Ward-King [39]	–	+	+	Medium	–	+	High	+	–	High	+	+	+	+	Low	High
Paing [28]	–	+	+	Medium	–	+	High	+	+	Low	7	+	+	7	Unclear	High
	Patient selection				Index test			Reference standard			Flow and timing					

* All studies scored positive because they all used a standardized questionnaire

Table 3 Subdivision of studies in structured and semi structured questionnaires

	Sensitivity	Specificity	Kappa	% Agreement
Semistructured				
Lyneham [40] ADIS-C-IV			0.86	
Tunstall [38] DDS			0.76	
Ward-King [39] ADI-R			No data	
Paing [28] P-Chips				Mean agreement 93.8 %
Structured				
Aziz [30] Caps*	84 %	80 %	0.75	
Aziz [30] HAM-D	79 %	100 %	0.70	
Evans [33] GHQ			0.75	
Evans [33] CIS-R			0.72	
Rohde [35] KIDDIE-SADS*			0.31–0.84	
Revicki [34] CIDI			0.78–1.00	
Revicki [34] PRIME-MD			0.80	
Wells [32] DIS*	55–71 %	89–95 %	0.03–0.66	0–58.8
Burke [37] CS-GDS	94 %	42 %		
Watson [25] DIS	Inconclusive	98 % or higher	Below 0.20–0.92	
Paulsen [27] SADS-L			0.69–0.84	85–100
Crippa [31] SCID			0.84	
Simon [36] SCID			0.73	
Cacciola [26] SCID*	0–62.5 %	94.1–100 %	0.03–0.66	0–58.8
Hajebi [29] SCID*	73.3–86.5 %	67.9–82.9 %		

* Questionnaire has several outcomes, see Table 1

stratified the sample for the presence of indicators for depression prior to the study. The studies in *high-risk* samples generally reported high percentages of agreement and high kappa values. Six of these studies, however, reported on <40 participants [27, 30, 34, 36, 38, 39]. Paing assessed 12 parents of children for assessing 21 psychiatric disorders. From the larger studies [29, 35, 37, 40], two reported on agreement only providing data about the kappa values [35, 40]. Lyneham [40] assessed 73 outpatient children with the ADIS-C-IV for anxiety, mood and externalizing disorders. They found a Kappa of 0.86. Rohde [35] used the KIDDIE-SADS in 60 psychiatric outpatients and found kappa values for major depressive disorder of 0.96, for anxiety disorder of 0.87, for alcohol and substance use of 1.00, and for adjustment disorder with depressed mood of 0.74.

Quality of included studies

Both studies in low-risk samples had a high risk of bias [36, 37]; three of four studies in intermediate risk samples had a medium risk of bias [27, 29, 31, 32]; from the remaining 10 studies in high-risk samples, two had low risk of bias (Table 2). In 13 studies, there were problems concerning patient selection [25–32, 34, 35, 38–40]: for instance, oversampling of patients with depressive symptoms [32] or with any lifetime psychotic disorder [29] or other sampling strategies leading to one group with cases and one group with non-cases. This strategy likely causes an exaggerated diagnostic accuracy. Three studies used a convenience sample resulting in uncertainty about the direction in which the results are biased [25, 26, 30]. Apart from patient selection, the other main cause of bias is interpretation of the index test with knowledge of the results of the reference test or vice versa. This also causes favorable results in validity or agreement. In one study, the same interviewer performed all tests [25], thus introducing bias in the direction of favorable validity measures.

Discussion

Is it valid to perform telephonic interviews instead of a face-to-face format? The use of telephone interviews relies on the premise that the diagnosis obtained with this method should be as valid as the diagnosis obtained in face-to-face interviews [29]. Generally, our conclusion is that there are too few studies properly performed to draw a definite conclusion about the comparability of telephone and face-to-face interviews for psychiatric morbidity.

The included studies are very heterogeneous (considering patient groups, setting, type of instruments and quality of the

Table 4 Subdivision of studies in time duration between telephonic and face-to-face interview

Mean week	Sensitivity	Specificity	Kappa	% Agreement
0–2 week				
Hajebi [29] SCID*	73.3–86.5 %	67.9–82.9 %		
Lyneham [40] ADIS-C-IV			0.86	
Evans [33] GHQ			0.75	
Evans [33] CIS-R			0.72	
Cacciola [26] SCID*	0–62.5 %	94.1–100 %	0.03–0.66	0–58.5
Tunstall [38] DDS			0.76	
Simon [36] SCID			0.73	
Watson [25] DIS	Inconclusive	98 % or higher	Below 0.20–0.92	
Burke [37]CS-GDS	94 %	42 %		
Revicki [34] CIDI			0.78–1.00	
Revicki [34] PRIME-MD			0.80	
2–4 week				
Rohde [35] KIDDIE-SADS*			0.31–0.84	
>4 week				
Aziz [30] Caps*	84 %	80 %	0.75	
Aziz [30] HAM-D	79 %	100 %	0.70	
Ward-King [39] ADI-R			No data	
Paulsen [27] SADS-L			0.69–0.84	85–100
Wells [32] DIS*	55–71 %	89–95 %	0.03–0.66	0–53.8
Crippa [31] SCID			0.84	
*NM**				
Paing [28] P-Chips				Mean agreement 93.8

NM not mentioned

Table 5 Subdivision of studies in interviewer type

	Sensitivity	Specificity	Kappa	% Agreement
Trained lay interviewer				
Revicki [34] CIDI			0.78–1.00	
Revicki [34] PRIME-MD			0.80	
Simon [36] SCID			0.73	
Watson [25] DIS	Inconclusive	98 % or higher	below 0.20–0.92	
Wells [32] DIS*	55–71 %	89–95 %	0.03–0.66	0–58.8
Tunstall [38] DDS			0.76	
Professional				
Lyneham [40] ADIS-C-IV			0.86	
Aziz [30] Caps*	84 %	80 %	0.75	
Aziz [30] HAM-D	79 %	100 %	0.70	
Rohde [35] KIDDIE-SADS*			0.31–0.84	
Burke [37] CS-GDS	94 %	42 %		
Crippa [31] SCID			0.34	
Hajebi [29] SCID*	73.3–86.5 %	67.9–82.9 %		
Ward-King [39] ADI-R			No data	
Not mentioned				
Paing [28] P-Chips				Mean agreement 93.8
Paulsen [27] SADS-L			0.69–0.84	85–100
Evans [33] GHQ			0.75	
Evans [33] CIS-R			0.72	
Cacciola [26] SCID*	0–62.5 %	94.1–100 %	0.03–0.66	0–58.8

* Questionnaire has several outcomes see Table 1

data). The two studies in the general population (*low risk* of psychiatric disorder) had high specificities for the DIS (diagnostic interview schedules) and SCID (structured clinical interview for DSM Disorders). This implies that the cases identified by telephone would probably be identified by the face-to-face interview as well. This conclusion is subject to doubt because these studies had a high risk of bias. Moreover, the sensitivity was low, implying that many of the cases might be missed by the telephone interview in comparison with the face-to-face interview. Agreement measures also showed that both interview modes are not leading to comparable results. The studies with *intermediate risk* of psychiatric disorder still had reasonably high specificity and low sensitivity, but medium to high risk of bias. The study in general practice had good kappa's for the broad category of psychiatric caseness. The studies with *high risk* of psychiatric disorder had higher sensitivity (less false-negative diagnoses) but lower specificity (more false-positive diagnoses) and were of low quality.

The reliability of the assessment of the lifetime prevalence of a psychiatric diagnosis is questionable regardless of which method has been used [41]. Therefore, if we restrict our conclusion to the studies assessing current psychiatric morbidity with agreement measures and medium or low risk of bias, there remain three studies [27, 31, 36] comparing the two modes of interviewing in patients with anxiety and depressive disorders. Kappa values in these studies range between 0.69 and 0.84, indicating good agreement. Possibly, in this field, the results of telephone interviewing are comparable with face-to-face interviews.

Strengths and weaknesses

A strength of our study is that, to our knowledge, this is the first systematic review studying the diagnostic agreement between telephone and face-to-face interviewing, using methodological criteria as the QUADAS statement. We performed a broad search in three databases for publications with a comparison between telephone and face-to-face diagnostic interviewing. We performed inclusion and exclusion of the publications and the data extraction with two researchers. An important weakness of our study is that it was impossible to perform a meta-analysis for this review because the eligible studies were too heterogeneous with respect to sampling, number of participants and study quality. We limited our review to studies using the diagnostic instrument for making a diagnosis and excluded studies comparing the scores of both modes of questioning.

Comparison with the literature

There are few systematic reviews comparing telephone diagnostic interviewing with face-to-face interviewing for mental health. One Dutch study [14] about depression concluded that telephone interviewing for depression is feasible and yields comparable results as face-to-face interviews, but the selected studies are weak concerning methodological quality. According to the authors, the psychiatric face-to-face interview is still the gold standard. The reliability of psychiatric interviews, however, is not perfect when considering agreement for interviews considered to be golden standards. For example, Segal [41] reports test–retest reliabilities (kappa) of the SCID interview of 0.32–1.00. Witchen [42] found very high kappa's for test–retest reliability of the CIDI interview, probably due to the fully structured nature of the CIDI interview. Our results are broadly in line with these studies. Another review [42] which compares telephone and video conference for assessing cognitive function concludes that the telephone interview had much to offer for the clinician and researcher, but nevertheless, the choice of the cognitive interview should fit for the limitations of telephone interviewing (lack of visual cues). This finding could also apply for mental health interviews. Psychiatric interviews are frequently clinician-administered. Clinicians always use more information than the direct answers on the questions; non-verbal cues probably play an important role in the final judgment about the diagnosis [2]. Therefore, a telephone interview cannot be as specific as a face-to-face interview.

Implications for future research

We recommend that further studies in this field should adhere to the guidelines in the QUADAS statement. Specifically, researchers should pay attention to patient selection and unbiased judgment of the tests. Patients should be consecutively (or randomly) enrolled in a study to avoid a case–control design. Future studies should include larger samples of participants, for example, at least 200 respondents (pilot) or 400 for reliability studies and even more for validity studies [43]. Finally, it would be desirable to study a specific disorder with a specific instrument instead of a combination of disorders including psychotic disorders and affective disorders with a general instrument. The study should use a structured interview, not a semistructured one because of the variability inherent for these interviews. For example, a study of depression with a specific structured depression questionnaire in a group of psychiatric outpatients. We propose to start with the field of depressive and anxiety disorders.

Conclusion

Taking this altogether, we conclude that there is inconsistent evidence that telephone interviews for the diagnosis of psychiatric disorders are valid compared to face-to-face

interviews. Telephone interviewing in the general population may not be valid because comparability measures are lowest in these low-risk populations. Finally, telephone interviewing for research purposes in depression and anxiety disorders might be a proper and valid method. Future research on depression and anxiety disorders may benefit the field and should preferably be conducted with fully structured interviews leaving no room for clinical interpretation of the answers.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix: PubMed search string

(bipolar disorders[mesh] OR bipolar disorders[tiab] OR bipolar disorder[tiab] OR "Anxiety Disorders"[Mesh] OR Anxiety Disorder[tiab] OR Anxiety Disorders[tiab] OR "depressive disorder"[Mesh] OR depressive disorder[tiab] OR depressive disorders[tiab] OR depression[tiab] OR depressions[tiab] OR Mental Disorders[tiab] OR mental disorder[tiab] OR DSM[tiab] OR "Diagnostic and Statistical Manual of Mental Disorders"[Mesh] OR psychiatric) AND ("Interview, Psychological"[Mesh] OR interview[tiab] OR interviews[tiab] OR interviewing[tiab] OR "Interviews as Topic"[Mesh] OR telephone-administered[tiab] OR face to face[tiab] OR questionnaires[mesh]) AND ("Diagnosis"[Mesh] OR Diagnosis[tiab] OR diagnoses[tiab] OR diagnostic[tiab] OR assessment[tiab] OR measuring[tiab]) AND ("Telephone"[Mesh] OR telephone[tiab] OR phone[tiab]).

References

1. Quinn RP, Gutek BA, Walsh JT (1980) Telephone interviewing: a reappraisal and a field experiment. Basic Appl Soc Psychol 1(2):127–153
2. Leeuw ED (ed) (1992) Data quality in mail, telephone and face to face surveys. TT-Publikaties, Amsterdam
3. Marcus AC, Crane LA (1986) Telephone surveys in public health research. Med Care 24(2):97–112
4. Tyzoon TT (1979) Telephone survey methods: the state of the art. J Mark 43(3):68–78
5. Semiatycki J (1979) A comparison of mail, telephone, and home interview strategies for household health surveys. Am J Public Health 69(3):238–245
6. Hoek L, Hovens JE (2011) Psychiatric diagnosis by telephone? Tijdschr Psychiatr 53(7):419–424
7. Groves RM (1990) Theories and methods of telephone surveys. Ann Rev Sociol 16:221–240
8. Aquilino WS, Sciuto LAL (1990) Effects of interview mode on self-reported drug use. Public Opin Q 54(3):362–393
9. Bowling A (2005) Mode of questionnaire administration can have serious effects on data quality. J Public Health 27(3):281–291
10. Herzog AR, Rodgers WL, Kulka RA (1983) Interviewing older adults: a comparison of telephone and face-to-face modalities. Public Opin Q 47(3):405–418
11. Holbrook AL, Green MC, Krosnick JA (2003) Telephone versus face-to-face interviewing of national probability samples with long Questionnaires: comparisons of respondent satisficing and social desirability response bias. Public Opin Q 67(1):79–125
12. Aquilino WS (1994) Interview mode effects in surveys of drug and alcohol use: a field experiment.I. Public Opin Q 58(2): 210–240
13. Erdman C (2001) The medicolegal dangers of telephone triage in mental health care. J Leg Med 22(4):553–579
14. Hoek L, Hovens JE (2011) Psychiatric diagnosis by telephone? Tijdschr Psychiatr 53(7):419–424
15. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
16. Lalkhen AG, McCluskey A (2008) Clinical tests: sensitivity and specificity. Contin Educ Anaesth Crit Care Pain 8(6):221–223
17. Terwee CBBS, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60(1):34–42
18. Viera AJGJ (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363
19. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155(8):529–536
20. Whiting P, Harbord R, Kleijnen J (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5:19
21. Bossuyt PM (2008) STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. Radiology 248(3):713–714
22. Spitzer RL, Williams JB, Gibbon M, First MB (1992) The structured clinical interview for DSM-III-R (SCID). I: history, rationale, and description. Arch Gen Psychiatry 49(8):624–629
23. Robins LN, Helzer JE, Croughan J, Ratcliff KS (1981) National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. Arch Gen Psychiatry 38(4): 381–389
24. Kessler RC, Ustun TB (2004) The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Method Psychiatr Res 13(2):93–121
25. Watson CG, Anderson PE, Thomas D, Nyberg K (1992) Comparability of telephone and face to face diagnostic interview schedules. J Nerv Ment Dis 180(8):534–535
26. Cacciola JS, Alterman AI, Rutherford MJ, McKay JR, May DJ (1999) Comparability of telephone and in-person structured clinical interview for DSM-III-R (SCID) diagnoses. Assessment 6(3):235–242
27. Paulsen AS, Crowe RR, Noyes R, Pfohl B (1988) Reliability of the telephone interview in diagnosing anxiety disorders. Arch Gen Psychiatry 45(1):62–63
28. Paing WW, Weller RA, Dixon TA, Weller EB (2010) Face-to-face versus telephone administration of the parent's version of the children's interview for psychiatric syndromes (P-ChIPS). Curr Psychiatry Rep 12(2):122–126
29. Hajebi A, Motevalian A, Amin-Esmaeili M, Hefazi M, Radgoodarzi R, Rahimi-Movaghar A et al (2012) 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 53(5):579–583

30. Aziz MA, Kenford S (2004) Comparability of telephone and face-to-face interviews in assessing patients with posttraumatic stress disorder. J Psychiatr Pract 10(5):307–313
31. Crippa JA, de Lima Osorio F, Del-Ben CM, Filho AS, da Silva Freitas MC, Loureiro SR (2008) Comparability between telephone and face-to-face structured clinical interview for DSM-IV in assessing social anxiety disorder. Perspect Psychiatr Care 44(4):241–247
32. Wells KB, Burnam MA, Leake B, Robins LN (1988) Agreement between face-to-face and telephone-administered versions of the depression section of the NIMH diagnostic interview schedule. J Psychiatr Res 22(3):207–220
33. Evans M, Kessler D, Lewis G, Peters TJ, Sharp D (2004) Assessing mental health in primary care research using standardized scales: can it be carried out over the telephone? Psychol Med 34(1):157–162
34. Revicki DA, Tohen M, Gyulai L, Thompson C, Pike S, Davis-Vogel A et al (1997) Telephone versus in-person clinical and health status assessment interviews in patients with bipolar disorder. Harv Rev Psychiatry 5(2):75–81
35. Rohde P, Lewinsohn PM, Seeley JR (1997) Comparability of telephone and face-to-face interviews in assessing axis I and II disorders. Am J Psychiatry 154(11):1593–1598
36. Simon GE, Revicki D, VonKorff M (1993) Telephone assessment of depression severity. J Psychiatr Res 27(3):247–252
37. Burke WJ, Roccaforte WH, Wengel SP, Conley DM, Potter JF (1995) The reliability and validity of the Geriatric Depression Rating Scale administered by telephone. J Am Geriatr Soc 43(6): 674–679
38. Tunstall N, Prince M, Mann A (1997) Concurrent validity of a telephone-administered version of the Gospel Oak instrument (including the SHORT-CARE). Int J Geriatr Psychiatry 12(10):1035–1038
39. Ward-King J, Cohen IL, Penning H, Holden JJ (2010) Brief report: telephone administration of the autism diagnostic interview–revised: reliability and suitability for use in research. J Autism Dev Disord 40(10):1285–1290
40. Lyneham HJ, Rapee RM (2005) Agreement between telephone and in-person delivery of a structured interview for anxiety disorders in children. J Am Acad Child Adolesc Psychiatry 44(3):274–282
41. Patten S (2009) Accumulation of major depressive episodes over time in a prospective study indicates that retrospectively assessed lifetime prevalence estimates are too low. BMC Psychiatry 9:19
42. Ball C, McLaren P (1997) The tele-assessment of cognitive state: a review. J Telemed Telecare 3(3):126–131
43. Charter RA (1999) Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. J Clin Exp Neuropsychol 21(4):559–566