diagnostics

MDPI

Systematic Review

Inter-Rater Reliability between Structured and Non-Structured Interviews Is Fair in Schizophrenia and Bipolar Disorders—A Systematic Review and Meta-Analysis

Hélio Rocha Neto [1,2,*], Ana Lúcia R. Moreira [3], Lucas Hosken [4], Joshua A. Langfus [5], Maria Tavares Cavalcanti [2,6], Eric Arden Youngstrom [5,7] and Diogo Telles-Correia [1,8]

1 Medical Faculty, Lisbon University, 1649-028 Lisbon, Portugal
2 Programa de Pós Graduação em Psiquiatria e Saúde Mental—PROPSAM, Instituto de Psiquiatria, Universidade Federal do Rio de Janeiro—UFRJ, Rio de Janeiro 22290140, RJ, Brazil
3 Cork Kerry Community Healthcare, Cork T12YE02, Ireland
4 Medical Psychology Sector, University Hospital Clementino Fraga Filho, HUCFF/UFRJ, Rio de Janeiro 21941913, RJ, Brazil
5 Clinical Psychology Program, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3270, USA
6 Medicine Faculty from Centro de Ciências da Saúde—Universidade Federal do Rio de Janeiro—UFRJ, Rio de Janeiro 21941902, RJ, Brazil
7 Helping Give Away Psychological Science, Chapel Hill, NC 27514, USA
8 Clinica Universitária de Psiquiatria e Psicologia Médica, Faculdade de Medicina, Universidade de Lisboa, 1649035 Lisbon, Portugal
* Correspondence: helio.neto@edu.ulisboa.pt or helio.neto@ufrj.br

Abstract: We aimed to find agreement between diagnoses obtained through standardized (SDI) and non-standardized diagnostic interviews (NSDI) for schizophrenia and Bipolar Affective Disorder (BD). Methods: A systematic review with meta-analysis was conducted. Publications from 2007 to 2020 comparing SDI and NSDI diagnoses in adults without neurological disorders were screened in MEDLINE, ISI Web of Science, and SCOPUS, following PROSPERO registration CRD42020187157, PRISMA guidelines, and quality assessment using QUADAS–2. Results: From 54231 entries, 22 studies were analyzed, and 13 were included in the final meta-analysis of kappa agreement using a mixed-effects meta-regression model. A mean kappa of 0.41 (Fair agreement, 95% CI: 0.34 to 0.47) but high heterogeneity ($\hat{I}^2$ = 92%) were calculated. Gender, mean age, NSDI setting (Inpatient vs. Outpatient; University vs. Non-university), and SDI informant (Self vs. Professional) were tested as predictors in meta-regression. Only SDI informant was relevant for the explanatory model, leaving 79% unexplained heterogeneity. Egger's test did not indicate significant bias, and QUADAS–2 resulted in "average" data quality. Conclusions: Most studies using SDIs do not report the original sample size, only the SDI-diagnosed patients. Kappa comparison resulted in high heterogeneity, which may reflect the influence of non-systematic bias in diagnostic processes. Although results were highly heterogeneous, we measured a fair agreement kappa between SDI and NSDI, implying clinicians might operate in scenarios not equivalent to psychiatry trials, where samples are filtered, and there may be more emphasis on maintaining reliability. The present study received no funding.

Keywords: standard diagnostic interview; meta-analysis; schizophrenia; bipolar affective disorder; reliability and validity

Citation: Neto, H.R.; Moreira, A.L.R.; Hosken, L.; Langfus, J.A.; Cavalcanti, M.T.; Youngstrom, E.A.; Telles-Correia, D. Inter-Rater Reliability between Structured and Non-Structured Interviews Is Fair in Schizophrenia and Bipolar Disorders—A Systematic Review and Meta-Analysis. *Diagnostics* **2023**, *13*, 526. https://doi.org/10.3390/diagnostics13030526

Academic Editor: Ahsan Khandoker

Received: 28 November 2022
Revised: 25 January 2023
Accepted: 28 January 2023
Published: 31 January 2023

1. Introduction

Low diagnostic reliability threatens the validity of both research and practice in psychiatry [1,2]. Accurate diagnosis forms the bedrock of treatment selection and

management of comorbidities, and the lack of a reliable diagnostic process can contribute to variability in outcomes, despite the availability of efficacious treatments. Nevertheless, diagnosing mental disorders poses a serious challenge, in part because of a lack of identifiable and specific biomarkers, leaving clinicians to rely on the evaluation of subjective characteristics susceptible to interpretation and potential bias [3–5].

The "operational revolution" popularized the definition of mental disorders using "operational criteria" comprising checklists of signs and symptoms [6]. Such definitions were considered "atheoretical" and thought to reduce the role of clinical judgment or interpretation, which may be tied to a particular conceptual model [6]. A standard diagnostic interview (SDI) is one way to evaluate whether a patient meets the operational definition of a disorder. The companion SDI for the Diagnostic and Statistical Manual (DSM) [7] is the Structured Clinical Interview for DSM (SCID) [8] which has become the prevailing standard for psychiatry research around the world [9].

The move toward operational diagnostic criteria and the use of SDIs aimed to solve the problem of unreliability in psychiatric diagnosis. Despite the increased reliability of SDIs, in practice clinicians often use non-standard diagnostic interviews (NSDI) [10,11]. These may be unstructured, impressionistic, guided by experience and intuition, and prototype-based diagnostic processes and their use can contribute to a gap between research evidence, which typically informs the construction of SDIs, and clinical practice. Nevertheless, NSDIs have some benefits. Evidence-Based Medicine (EBM) is described as the interaction of three areas of knowledge: clinical experience and expertise, patient values and expectations, and the best external evidence [12]. NSDIs can address the complexities that arise in specific cases but sacrifice the first two in favor of the third. Thus, the use of SDIs can err in the opposite direction, causing a tension between "rigor" and "relevance" [13]. Importantly, clinicians can still operationalize standardized diagnostic criteria without using an SDI.

One question is whether clinicians' diagnoses using NSDIs are less accurate compared to those made with SDIs. Since SDIs are currently considered the gold standard for diagnosis (particularly with a consensus review process), a more tractable question involves examining the agreement between the two approaches. If SDIs and NSDIs disagree, then typical psychiatric practice is at best less accurate and may be subject to systematic biases. Furthermore, if NSDIs do not reproduce the results of SDIs, evidence-based interventions, such as medications, psychotherapies, or alternative treatments, tested in trials with SDIs, are less likely to work as expected for patients diagnosed via NSDIs.

The disjuncture between research and practice may contribute to the shrinkage in treatment effect sizes moving from efficacy to effectiveness designs. Furthermore, NSDIs are subject to local and regional variations in practice. The Dartmouth Medical Atlas Project has found this at every level of analysis within the USA—national, regional, state, and local municipalities—and across every medical specialty examined [14,15]. Current "big data" projects, using statistical and machine-learning models, hinge on the accuracy of NSDIs as they mine medical records and claims data. If NSDIs are fundamentally prone to systematic biases, then these sophisticated models will be trained using unreliable targets and unable to generalize across regions or settings [16]. If they are not, we may consider that the use of SDIs in the research setting is dispensable, and studies using clinical NSDIs only are feasible. This further highlights the importance of understanding the level of agreement between SDIs and NSDIs for diagnostic decision-making.

Little previous work has examined agreement between SDI and NSDI diagnoses. A previous meta-analysis [9] found low agreement between SDI and NSDI diagnoses in children and adolescents. Later, Jensen-Doss [17] found an equivalent result comparing K-SADS and NSDIs, but again in a child and adolescent population. Rettew's work [9] is the latest review to address this question and is now 15 years old, without adult population evaluation. Thus, in order to update these findings and fill in the gap in adult psychiatry, the current work presents a systematic review of the reliability between SDI

and NSDI diagnoses in schizophrenia and Bipolar Affective Disorder (BD) patients, followed by a meta-analysis using kappa agreement as the effect size.

We focused on schizophrenia and BD diagnosis as index disorders, as their diagnostic constructs seem valid and persistent across the world, beyond cultural barriers [18–21]. This reduces the likelihood that our kappa estimates will be influenced by disagreement about the construct rather than differences between SDIs and NSDIs. As a result, our estimate here may be interpreted as near the upper limit of agreement, with other disorders showing lower overall agreement due to differences in conceptualization.

2. Materials and Methods

This review examined studies comparing diagnostic accuracy of SDIs and NSDIs, searching for each SDI by name and acronym. SDIs targeting both schizophrenia and BD (as is the case of SCID [8]) or just one of these diagnoses (as in the Mood Disorder Questionnaire; MDQ [22]) were then selected to build the search string. We initially sought to include the "missing gold standard" or Longitudinal, Expert, All Data (LEAD) approach [23,24]. However, use of "LEAD" in searches yielded few results. Therefore, the following SDIs were included: Composite International Diagnostic Interview—CIDI [25], Diagnostic Interview Schedule—DIS [26], Mini International Neuropsychiatric Interview—MINI [27], Schedules for Clinical Assessment in Neuropsychiatry—SCAN [28], Structured Clinical Interview for DSM—SCID [8], Standard for Clinicians Interview in Psychiatry—SCIP [29], Schedule for Affective Disorders—SADS [30], Diagnostic Interview for Genetic Disorders—DIGD [31], Bipolar Spectrum Diagnostic Scale—BSDS [32], General Behavior Inventory—GBI [33], Mood Disorder Questionnaire—MDQ [22], The Comprehensive Assessment of Symptoms and History—CASH [34]. As a generic reference for SDIs, we also included the term "standard diagnostic interview—SDI".

We conducted the search in MEDLINE, SCOPUS and ISI Web of Science databases. We restricted the year of publication to 2007 and beyond, since the Rettew et al. meta-analysis had collected data until that year. We augmented the search to include papers published in Portuguese and Spanish, in addition to English, though all articles recovered had an English version. The search string was built using both SDI acronyms and full length in title, abstract, subject and keywords, adapting Boolean operators for each database.

Beyond time span and language, inclusion criteria focused on original articles and reviews as publication type, and clinical trials, meta-analyses, randomized controlled trials, reviews and systematic reviews in research type. There are some reasons for including papers other than original diagnostic studies: Firstly, the number of studies that make a direct comparison between SDI and NSDI were surprisingly low; secondly, it is expected for clinical trials to recruit their patients with existing NSDI diagnoses, then to administer an SDI, and then extract their validated sample, which could give us more data than original diagnostic studies only; thirdly, we hoped to harvest references not included in MEDLINE, SCOPUS and ISI Web of Science through other reviews and meta-analyses. Table 1 details the inclusion and exclusion criteria. For quality assessment, we used the "Standards for Reporting of Diagnostic Accuracy Studies" (STARD) [35] criteria and applied the Quality Assessment of Diagnostic Accuracy Study (QUADAS–2) [36] tool. An "extraction tool" was built to get the information desired from each paper (described later).

Table 1. Inclusion and Exclusion criteria.

Inclusion Criteria	Exclusion Criteria
1. Is it a study that compares agreement between a SDI and a NSDI?	1. Were participants older than 18 and younger than 65 years old? If not, is it possible to separate them from the sample?
2. Is it a study using one of the 11 selected tools? (CIDI, DIS, MINI, SCAN, SCID, SCIP, SADS, DIGD, BSDS, GBI, MDQ, CASH). Refer to which one.	2. Were subjects with intelligence limitation excluded? If not, is it possible to separate them from the sample?
3. Was it published in a peer-reviewed journal?	3. Are SDI and NSDI diagnoses independently obtained?
4. Is it possible to extract diagnostic agreement for schizophrenia or BD? Sign which.	4. Was the NSDI diagnosis obtained by qualified health professional (Physician, Psychiatrist, psychologist or mental health at college grade professional)?
5. Does the reference show kappa agreement between SDI and NSDI? If not, is it possible to calculate it?	5. Was the NSDI diagnosis obtained exclusively for the present study, or was it obtained from medical archive?
	6. Was the diagnosis based on DSM III, DSMIIIr, DSMIV, DSMIVtr, DSMV, ICD 9, ICD 10 or ICD 11?

Footnote: SDI—standard diagnostic interview; NSDI—non-standard diagnostic interview; See Table 2 for SDI acronyms in full length.

Table 2. Number of entries by SDI acronym and full-length name.

Standard Diagnostic Interview	Acronym	Full Length
Composite International Diagnostic Interview—CIDI	1162	2662
Diagnostic Interview Schedule—DIS	769	619
Mini International Neuropsychiatric Interview—MINI	10967	2420
Schedules for Clinical Assessment in Neuropsychiatry—SCAN	9259	132
Structured Clinical Interview for DSM—SCID	3103	3872
Standard for Clinicians Interview in Psychiatry—SCIP	88	6
Schedule for Affective Disorders—SADS	584	971
Diagnostic Interview for Genetic Disorders—DIGD	0	0
Bipolar Spectrum Diagnostic Scale—BSDS	35	36
General Behavior Inventory—GBI	42	71
Mood Disorder Questionnaire—MDQ	287	320
The Comprehensive Assessment of Symptoms and History—CASH	1	1
Standard Diagnostic Interview—SDI	31	33

Rater Training and Reliability

Two authors (HGRN and LH) trained to use STARD, QUADAS–2, and the extraction tool in a dummy sample and then independently screened and selected references based on the instrument. Training was done in blocks of 10 papers, with the *a priori* protocol entailing a minimum of 3 training blocks and additional training until a kappa of 0.8 was achieved. After the third trial, inter-coder kappa was 0.81 ("Almost Perfect"; CI 0.69–0.93; $p < 0.001$) and article coding proceeded.

For the meta-analysis explanatory model, 10 variables were extracted: Number of subjects in each sample (*N*), female participants ratio, mean sample age in years, SDI, SDI informant (self vs. professional), informant profession, sample diagnosis, research setting (university vs. non-university), clinical setting (Inpatient vs. Outpatient), and country (later converted in Life Expectancy Index—LEI, using WHO database data, matching country data by publication year [37], as it seemed a better way to measure health system strength than countries name alone). The 2 coders also applied STARD and QUADAS–2 independently. Differences were reviewed directly in the reference or, whenever possible, contacting their authors to resolve any conflicts.

This review protocol was registered in PROSPERO under the registration number CRD42020187157 on the 19 May 2020, before reference extraction. The 3 databases were

accessed on the 10 June 2020. This study and report have been designed and written following PRISMA [38] orientation (PRISMA checklist appended).

Agreement (kappa) of SDI vs. NSDI diagnoses was directly extracted from papers where they were already reported or calculated when paper offered enough information or their authors provided it after direct request by email. For the meta-analysis, we followed Jansen's approach [39]. A power analysis using the *metapower* package v0.2.2[40] found that an effect size of 0.4 (fair agreement, and roughly the median in the DSM—5 field trials) [40] was detectable at a level of 99.8%, with a median sample size (N ~ 114), and 13 studies using a random effects model and high heterogeneity (e.g., I^2 ~ 0.9). Power would have been >86% to detect differences of k = 0.4 vs. 0.2 under moderate heterogeneity (I^2 ~ 0.5), though it dropped to 28% under conditions of high heterogeneity for random effects model testing moderators.

Once coded, kappas were pooled, and 95% CI was calculated using a random effects model. After pooled kappa calculation, mixed model meta-regression probed the heterogeneity ($\hat{I}^2$). Statistics were conducted using the *metafor* [41] and *metapower* [42] package for R statistical software (v4.1.2; R Core Team, Vienna, Austria), which also provided the funnel and forest plots.

3. Results

Our search protocol captured 54,231 initial entries. Further applications of inclusion/exclusion criteria, deletion of duplicates and unrelated references resulted in 49 references retained for eligibility assessment. A final list of 13 papers were coded for analysis, providing 15 kappas. Figure 1 presents the flow diagram from search to final inclusion.

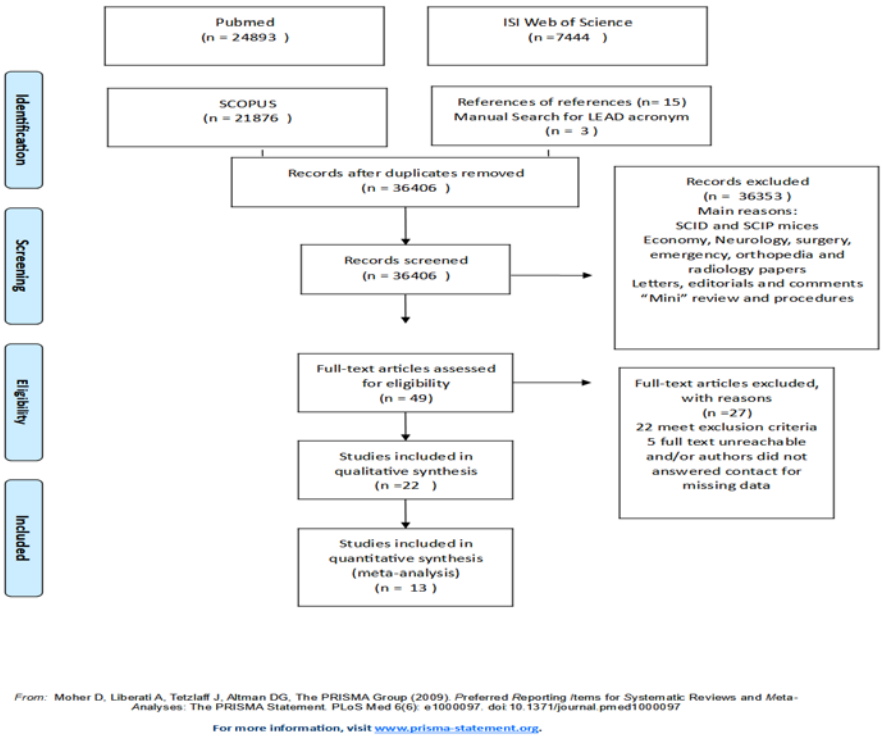

Figure 1. Screening, selection, inclusion and exclusion workflow.

SCID was the most reported SDI (n = 3872) (based on full length, to avoid cross

references with other acronyms), followed by CIDI (n = 2662) and MINI (n = 2420). DIGD was not found in any reference, and CASH was used in a single report (see Table 2 for details). Almost all years had at least 1 reference in the final list, but only 5 SDIs were represented (SCID, MINI, CIDI, MDQ and BSDS). Table 3 presents the final list of included sources with author, publication year, and diagnosis' details.

Table 3. Selected papers by author, year, diagnosis, sample size, kappa, applied SDI, sample country, clinical scenario, NSDI scenario and SDI applicant.

Author	Year	Diagnosis	Sample Size	Kappa	SDI/Applicant	Country	NSDI Clinical Scenario
Unenge et al. [43]	2012	SCZ	46	--	SCID/Health professional	Sweden	General outpatient
Adelufosi et al. [44]	2012	SCZ	324	--	SCID	Nigeria	General outpatient
Rafrafi et al. [45]	2013	SCZ	114	0.410	CIDI/Health professional	Tunisia	General inpatient
Yazici et al. [46]	2018	SCZ	131	--	SCID	Turkey	Universitary outpatient
Nordgaard et al. [47]	2012	SCZ	100	0.330	SCID/Health professional	Denmark	Universitary inpatient
Stewart et al. [48]	2007	BD	21	--	SCID/Health professional	USA	General inpatient
Zimmerman et al. [49]	2008	BD	700	0.450	SCID/Health professional	USA	Universitary outpatient
Jon et al. [50]	2009	BD	238	0.370	MDQ	South Korea	Universitary outpatient
Vázquez et al. [51]	2010	BD	101	0.550	BSDS	Argentine	General outpatient
Jiménez et al. [52]	2012	BD	138	--	SCID	Spain	Universitary outpatient
Suresh et al. [53]	2013	BD	42	0.250	MDQ	USA	Universitary inpatient
Asaad et al. [54]	2014	BD	390	--	SCID/Health professional	Egypt	General outpatient
Verhoeven et al. [55]	2017	BD	7016	0.480	MINI/Health professional	Netherlands	Universitary outpatient
Ince et al. [56]	2019	BD	183	0.520	BSDS	Turkey	Universitary outpatient
Imamura et al. A [57]	2015	BD	55	0.340	MDQ	Japan	General outpatient
Imamura et al. B [57]	2015	BD	55	0.300	BSDS	Japan	General outpatient
Rajkumar et al. [58]	2016	BD	139	--	MINI	India	General outpatient
Wesley et al. [59]	2018	BD	168	--	MINI/Health professional	India	Universitary outpatient
Hebbrecht et al. [60]	2020	BD	276	0.660	MINI/Health professional	Belgium	Universitary inpatient
Hong et al. [61]	2014	BD	345	0.360	MDQ	South Korea	Universitary outpatient
Lee et al. A [62]	2013	BD	113	0.120	MDQ	South Korea	Universitary inpatient
Lee et al. B [62]	2013	BD	113	0.400	BSDS	South Korea	Universitary inpatient
Kung et al. [63]	2015	BD	860	0.410	MDQ	USA	General inpatient

Footnote: SCZ—schizophrenia; BD—Bipolar Affective Disorder; SCID—Structured Clinical Interview for DSM; CIDI—Composite International Diagnostic Interview; MDQ—Mood Disorder Questionnaire; BSDS—Bipolar Spectrum Diagnostic Scale; MINI—Mini International Neuropsychiatric Interview.

References were of "average" quality based on QUADAS–2 scores. The most common issue was that subjects were usually recruited from settings dedicated to a specific disease or to similar diagnostic spectra (e.g., schizophrenia spectrum) when performing reliability calculations. In two studies, it was not possible to check patient selection bias [51,57], and a third may have excluded patients with previous mood-related psychotic symptoms [62]. In Suresh et al. [53], it was not clear if clinicians knew SDI results (i.e., failure of masking), but that was not an issue for all other references. Whenever a

gross disruption in case flow and timing of diagnoses was identified, the reference was excluded ($k = 1$), but in the final sample, only eight studies explicitly reported the interval between SDI and NSDI diagnosis, resulting in most studies receiving an "unknown" classification. Most studies used methodologies considered equivalent to usual clinical settings, except for Nordgaard et al. [47], where the reference standard was a diagnostic consensus among two highly trained researchers in diagnostic interviews. Figures 2 and 3 report the full QUADAS–2 coding.

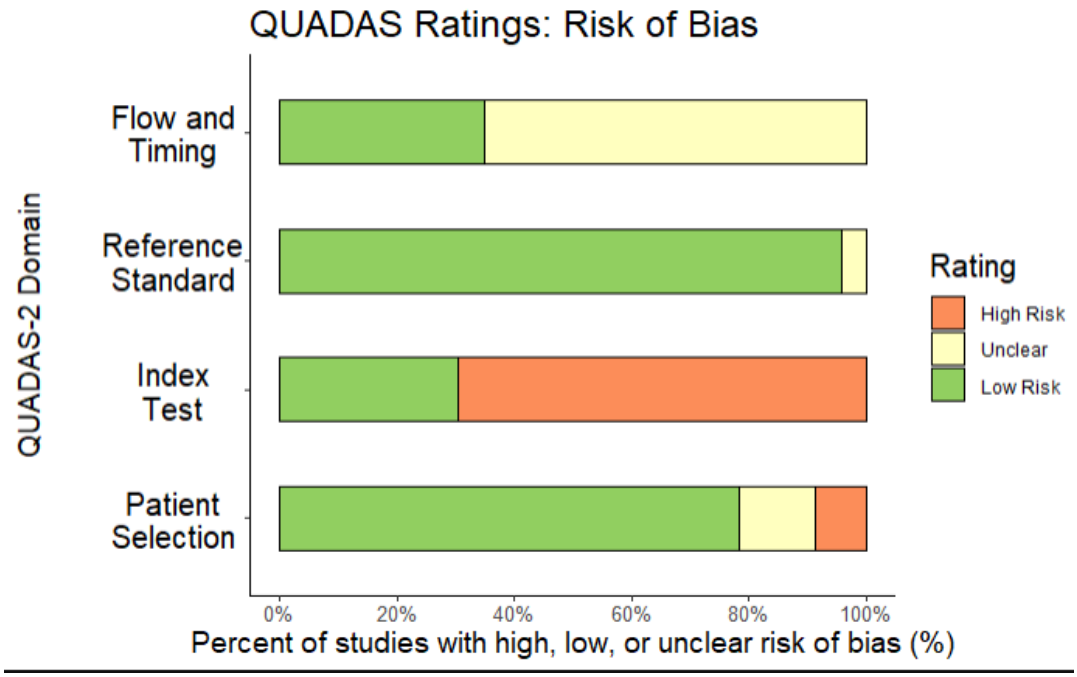

Figure 2. QUADAS Risk of Bias report.

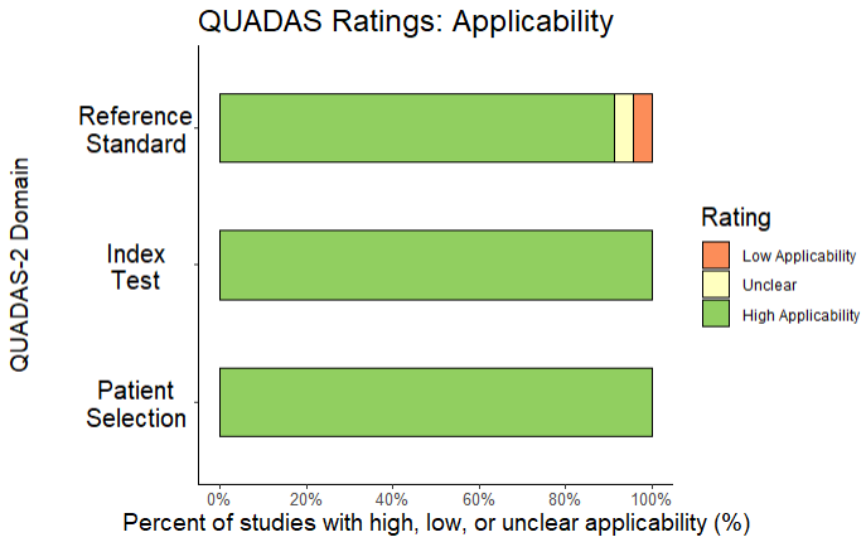

Figure 3. QUADAS Applicability report.

Of the final analyzed entries, 15 results were included for meta-analysis. These studies reported kappas ranging from 0.12 to 0.66. The trim-and-fill funnel plot (Figure 4) indicated that if there was bias, it would have been due to unpublished studies having a small sample size and high kappas (e.g., three implied studies in that region of the plot). Egger's test indicated no significant bias. The weighted mean kappa was 0.41 (Fair agreement, 95% CI: 0.34 to 0.47), however, with a high heterogeneity ($\hat{I}^2 = 92\%$) (Figure 5).

An augmented meta-regression model tested female percentage, publication year, NSDI setting (inpatient vs. outpatient; university vs. non-university) and SDI interview (CIDI, MDQ, MINI, BSDS, SCID) as potential moderators. The model accounted for R^2 = 10.1% of variance in kappas, Q_{model} (8 *df*) = 9.20, *n.s.*, leaving 79% unexplained heterogeneity, Q_{error} (6 *df*) = 38.31, $p < 0.00005$. An alternate model collapsing SDI interview into a format (self-administered vs. interview) performed similarly: R^2 = 18.8%, Q_{model} (5 *df*) = 7.50, *n.s.*, leaving 79% unexplained heterogeneity, Q_{error} (9 *df*) = 54.46, $p < 0.00005$.

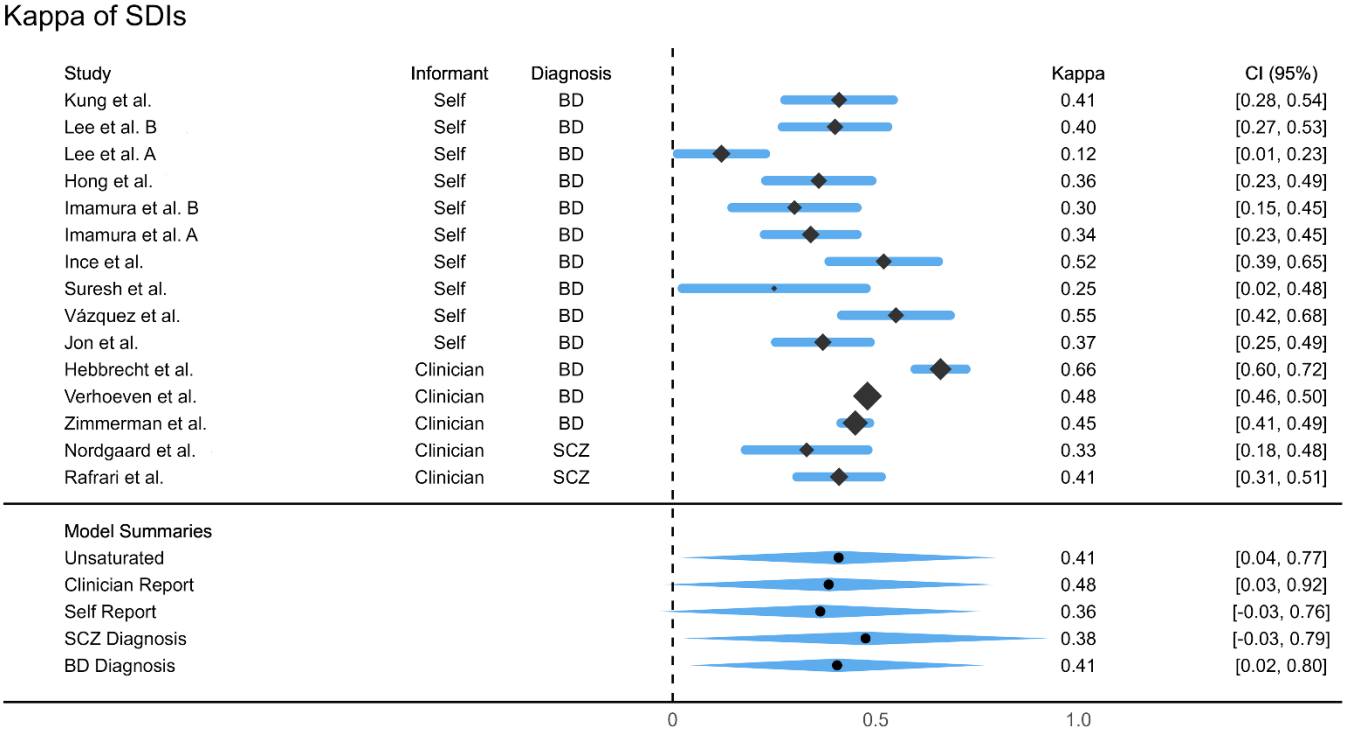

Figure 4. Trim and fill funnel plot. White dots indicate implied missing studies, [45,47,49–51,53,55–57,60–63].

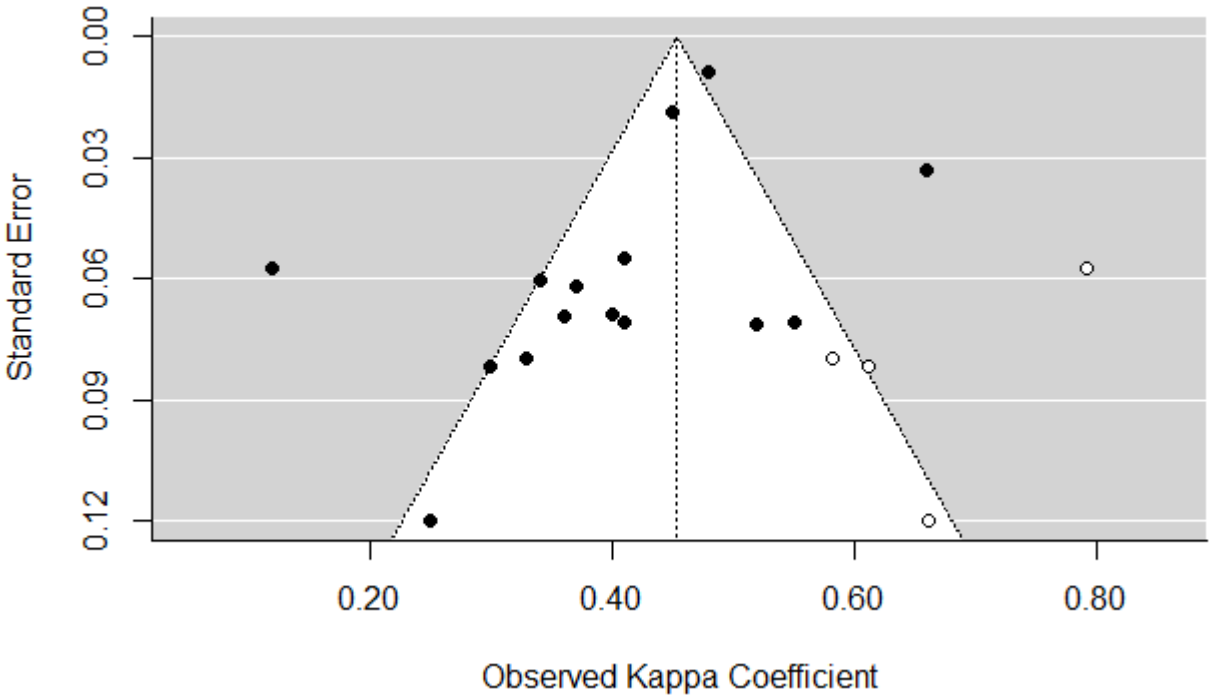

Figure 5. Forest Plot.

4. Discussion

The goal of the present study was to meta-analyze agreement between diagnoses based on SDIs versus NSDIs in patients with BD and schizophrenia. The average agreement between the two methods was "fair" based on a literature of "average" reporting quality. High heterogeneity persisted, even after exploring a variety of potential predictors using mixed meta-regressions.

The type of information obtained with SDIs versus NSDIs, as well as clinicians' use of diagnostic prototypes instead of standardized criteria, may be reasons for the low agreement. However, clinicians' prototype-based approach usually match ICD or DSM criteria, even with NSDIs as information-gathering procedure [11]. NSDIs allow clinicians' use of clinical judgment to uncover relevant information not probed in a SDI [64,65]; however, this can also incur biases to jeopardize the evidence-gathering process. Thus, the lack of agreement may be due to different information being uncovered with the use of SDIs versus NSDIs, even with clinicians applying standardized criteria. If SDIs and NSDIs result in different diagnoses, despite the use of operational criteria for the disorders themselves, then research in psychiatry works with diagnostic models that do not represent clinical practice and vice versa.

4.1. Assessing Model Heterogeneity

None of the variables examined as potential moderators significantly reduced heterogeneity in kappa estimates. Previous studies suggested that patients give more information and are more reliable in their statements on self-reporting instruments compared to clinician-guided interviews, particularly about sensitive or stigmatized topics [66]. However, self-administered interviews may lead to failure to accurately report symptoms due to difficulty in comprehending technical language [67]. Additionally, both mania and psychosis can involve a lack of insight into one's mental state or behavior. It is possible that patients misunderstood questions, reported more information than requested by doctors or did not classify certain signs and symptoms in the same way a clinician would [68].

Considering the other explanatory variables, we anticipated that a semi-structured

format would be more sensitive and specific than a fully structured SDI. However, the retrieved studies largely did not report which format was used, with the exception of Nordgaard et al. [63], who raised this hypothesis. Thus, the effect of format (structured vs. semi-structured) could not be tested as a heterogeneity explanation.

We also expected that strong public health systems would be associated with better practices, more professional training, and the adoption of quality protocols. Using Life Expectancy Index (LEI) as a proxy for health system quality, we tested whether it would explain further reliability between SDI and NSDI; however, it had no impact on our explanatory model.

The setting where NSDI was performed was also expected to be a predictor of heterogeneity. University settings might be more adherent to diagnostic protocols and have clinicians that are up to date regarding diagnostic protocols compared to non-university services. Furthermore, we expected to see a difference between inpatient and outpatient clinics due to the number of assessments and intensity of behavior observation. However, none of these factors were significant in the explanatory model.

4.2. Limits for Systematic Review of Agreement Studies

This review was limited by the number of studies that reported adequate information for coding, which represented <1% of the citations captured in the pre-registered search strategy. Furthermore, although most studies showed a QUADAS–2 rating of "adequate" quality (Figure 2), we encountered challenges due to inadequate reporting, including difficulty extracting information about potential moderators, as well as an extremely low yield of usable studies compared to initial search results.

There were several common weaknesses in the reporting of results that resulted in the exclusion of potentially interesting predictors of agreement. Since SDIs are the dominant standard for research in psychiatry, we expected studies to report agreement statistics of NSDIs vs. SDIs as part of the study (e.g., patients initially diagnosed with schizophrenia using NSDI then recruited for research and tested with a SDI to confirm diagnosis). Unfortunately, very few papers reported the initial number of tested subjects and most reported only SDI-positive recruited participants. This makes it impossible to estimate the base rate, the kappa, and other statistics needed to assess agreement between SDIs and NSDIs [69].

Another challenge in reviewing the literature was that studies often used a specific module of SDI instead of the whole instrument. Both DSM and ICD have exclusion criteria for disorders that should render impossible at least some types of comorbidity (such as schizophrenia and BD). Triage tools developed for a single diagnosis, like MDQ and BSDS, will be particularly prone to such problems [70]. These instruments can only consider whether BD symptoms are present or absent, never checking or excluding other hypotheses. This increases the probability of random agreement between SDI and NSDI, lowering the estimated reliability (kappa) and also the validity of the diagnosis. Thus, restricting the SDI to a single module likely affects both a tool's sensitivity and specificity and raises concerns about validity.

Despite having excellent power to evaluate the kappa, we were unable to explain a significant proportion of the heterogeneity in kappa estimates. Heterogeneity was extremely high, and the power to test moderators using a random effects model (as specified a priori) was not optimal. Results are consistent with the possibility that clinicians in "NSDI mode" access different clinical information from SDIs, consequently establishing different diagnoses. Another explanation is that clinicians might be using specific naturalistic and regional prototypes [71] or that diagnostic criteria were interpreted differently across the many cultural contexts. Thus, even if NSDIs and SDIs were targeting the same clinical criteria, there may be differences in how they are framed due to different norms or expectations. Both ICD and DSM manuals draw attention to the possibility that the disorder construct might have relevant differences among people from different countries. Our study included studies from nine different countries on five

continents, introducing the possibility of cultural heterogeneity; however, LEI (which differed by country) had no effect in the explanatory model. Finally, linguistic differences may also affect reliability; although SDIs are usually validated after translation, the same could not be said of clinicians using NSDIs.

One initial goal of this study was to examine agreement between SDIs and LEAD standard diagnoses. Despite recent ICD and DSM field trials [40,72], we have not found any paper considering a LEAD gold standard against SDI. Furthermore, the number of codable papers comparing SDI and NSDI diagnoses were bigger than in the Rettew et al. article. Our results show that very few SDIs are actually used. DIGS was not used at all, and most other SDIs have fewer reports when compared with the three most used (SCID, MINI and CIDI). Overall, there is a lack of reporting on the agreement between methods of diagnosis (i.e., LEAD, SDI, NSDI). A major strength of the current work is that it is the only study in the last decade to compare SDIs and NSDIs, a very relevant issue for translational psychiatry. Other relevant strengths were the use of an extraction tool, parallel reviewing strategy, and a very inclusive screening methodology, searching for papers from all continents. It is unlikely that any relevant report was not accessed.

4.3. Study Limitations

Due to changes in institutional access, we were unable to screen PsycINFO; although it is unlikely that a relevant journal was indexed in that library, but not in MEDLINE, ISI Web of Science or SCOPUS, that was a departure from our predefined protocol. Additionally, we did not systematically check gray literature and non-indexed journals, which may have resulted in missing smaller studies. However, that would likely have resulted in studies with low kappa, as usually very positive findings are published. This concern is mitigated, however, by the funnel plot we obtained (fig 4), which points toward a lack of literature with high kappa findings, not low ones.

Working with schizophrenia and BD was a choice as we wanted to measure agreement in two highly valid and prevalent disorders around the world, with supposedly little cultural influence in their definitions across cultures. However, our results cannot be translated to other mental disorders. Indeed, we hypothesize that other disorders might have a poorer reliability performance due to cultural and values interference in NSDI evaluation, which would require further testing outside the scope of this study.

The methodology was not inclusive of comorbidities that might be reasonably prevalent in both disorders. However, failing to diagnose schizophrenia or BD in subjects with other disorders would also be considered an agreement failure, and so we believe that it would have no impact on our findings. Also, our methodology included article types, such as reviews and clinical trials, that would not have been adequately evaluated by our quality tools. The inclusion of these types of articles was a choice in order to increase our sample size, but since none of them were included in the analysis, this methodology option had no impact on the present study.

Finally, the unexplained heterogeneity may jeopardize the interpretation of meta-analysis results. However, the overall estimated kappa aligns with two prior meta-analyses [9,17] as well as what is usually measured in single reports of very well-conducted studies, like Kottwicki [73] longitudinal study of reliability between SDI and NSDI. Moreover, our study used best practices for conducting systematic reviews, including PRISMA guidelines. Thus, since we reached a result that is equivalent to similar studies in the field and employed a rigorous methodology, the heterogeneity warrants consideration as observation in itself rather than as an artifact of our methods.

Reliability has been a major challenge in psychiatry over at least the last 70 years [74]. Most studies showing an increase in reliability with the use of DSM criteria are based on research in academic rather than clinical settings. This reinforces the idea that standardized criteria are not used in clinical practice [11], where a prototype approach may seem more feasible to clinicians [75]. Future work should investigate the extent to

which the heterogeneity in agreement between SDIs and NSDIs diagnoses may be attributable to clinicians using clinical prototypes that do not align with categorical diagnostic constructs such as the DSM or to the unreliability of data achieved by SDIs and NSDIs approach.

Our results corroborate previous findings showing only fair kappas between SDIs and NSDIs in clinical settings. Most studies that use SDIs in a previous NSDIs-diagnosed sample do not report the size and results of the tested sample. Also, it is necessary to be more explicit about the full or partial use of an SDI when selecting subjects for research. We would like to suggest that reviewers and journals request this information during the peer review process, but also that guidelines including such information are available for best practices in psychiatry research.

Author Contributions: H.R.N. conceived the review, developed the search strategy and the data extraction instrument, collected the data, and wrote the main draft. L.H. made the parallel screening, data extraction, instrument application and final data consolidation. A.L.R.M. validated and organized the data for the meta-analysis, suggested meta-regression variables, and made substantive contributions to writing the main draft. E.A.Y. conducted the meta-analysis, wrote the analytic scripts, provided statistical supervision for the results and explanatory models, and edited the final draft. J.A.L. produced figures and made substantive edits to the final draft. M.T.C. and D.T.-C. organized the data collection, provided support from university libraries, and provided suggestions and comments throughout the writing process. All authors reviewed and agreed on analyses, as well as the final version of the manuscript, tables, and figures. All authors have read and agreed to the published version of the manuscript.

Funding: The present manuscript is not part of a funded project.

Institutional Review Board Statement: This article does not contain any studies with human participants performed by any of the authors.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: Hélio G Rocha Neto receives a monthly wage from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazilian Education Ministry, due to his PhD dedication. Youngstrom has received royalties from the American Psychological Association and Guilford Press, consulted with Signant Health about psychological assessment, and received funding from NIMH. He is the founder and Executive Director of Helping Give Away Psychological Science (HGAPS.org). Ana Moreira, Lucas H. de Carvalho, Joshua A. Langfus, Maria T. Cavalcanti and Diogo Telles-Correia have no conflict of interest to declare.

References

1. Kreitman, N.; Sainsbury, P.; Morrissey, J.; Towers, J.; Scrivener, J. The Reliability of Psychiatric Assessment: An Analysis. *Br. J. Psychiatry* **1961**, *107*, 887–908.
2. Nussbaum, A.M. Questionable Agreement: The Experience of Depression and DSM-5 Major Depressive Disorder Criteria. *J. Med. Philos. A Forum Bioeth. Philos. Med.* **2020**, *45*, 623–643.
3. Helzer, J.E. Reliability of Psychiatric Diagnosis. *Arch. Gen. Psychiatry* **1977**, *34*, 129.
4. Wing, J.K.; Birley, J.L.; Cooper, J.E.; Graham, P.; Isaacs, A.D. Reliability of a procedure for measuring and classifying 'present psychiatric state'. *Br. J. Psychiatry* **1967**, *113*, 499–515.
5. Aboraya, A.; First, M.B. Point/counterpoint: The reliability of psychiatric diagnosis. *Psychiatry* **2007**, *4*, 22–25.
6. Andreasen, N.C. DSM and the Death of Phenomenology in America: An Example of Unintended Consequences. *Schizophr. Bull.* **2006**, *33*, 108–112.
7. *Diagnostic and Statistical Manual of Mental Disorders*, 3rd ed.; American Psychiatric Association: Washington, DC, USA, 1980.
8. Spitzer, R.L.; Williams, J.B.; Gibbon, M.; First, M.B. The Structured Clinical Interview for DSM-III-R (SCID): I: History, Rationale, and Description. *Arch. Gen. Psychiatry* **1992**, *49*, 624–629.
9. Rettew, D.C.; Lynch, A.D.; Achenbach, T.M.; Dumenci, L.; Ivanova, M.Y. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *Int. J. Methods Psychiatr. Res.* **2009**, *18*, 169–184.
10. Aboraya, A. Do psychiatrists use structured interviews in real clinical settings? *Psychiatry* **2008**, *5*, 26–27.
11. Rocha Neto, H.G.; Sinem, T.B.; Koiller, L.M.; Pereira, A.M.; de Souza Gomes, B.M.; Veloso Filho, C.L.; Cavalcanti, M.T.; Telles-Correia, D. Intra-rater Kappa Accuracy of Prototype and ICD-10 Operational Criteria-Based Diagnoses for Mental Disorders: A Brief Report of a Cross-Sectional Study in an Outpatient Setting. *Front. Psychiatry* **2022**, *13*, 1–8.

12. Sackett, D.L.; Rennie, D. The science of the art of the clinical examination. *JAMA* **1992**, *267*, 2650–2652.
13. Schon, D. *The Reflective Practitioner: How Professionals Think in Action*; Basic Books: New York, NY, USA, 1983.
14. Goodman, D.C. Unwarranted Variation in Pediatric Medical Care. *Pediatr. Clin. N. Am.* **2009**, *56*, 745–755.
15. Goodman, D.C.; Ganduglia-Cazaban, C.; Franzini, L.; Stukel, T.A.; Wasserman, J.R.; Murphy, M.A.; Kim, Y.; Mowitz, M.E.; Tyson, J.E.; Doherty, J.R.; et al. Neonatal Intensive Care Variation in Medicaid-Insured Newborns: A Population-Based Study. *J. Pediatr.* **2019**, *209*, 44–51.e2.
16. König, I.R.; Malley, J.D.; Weimar, C.; Diener, H.-C.; Ziegler, A.; on behalf of the German Stroke Study Collaboration Practical experiences on the necessity of external validation. *Stat. Med.* **2007**, *26*, 5499–5511.
17. Jensen-Doss, A.; Youngstrom, E.A.; Youngstrom, J.K.; Feeny, N.C.; Findling, R.L. Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *J. Consult. Clin. Psychol.* **2014**, *82*, 1151–1162.
18. Cherry, K.E.; Penn, D.; Matson, J.L.; Bamburg, J.W. Characteristics of schizophrenia among persons with severe or profound mental retar-dation. *Psychiatr. Serv.* **2000**, *51*, 922–924.
19. Ropacki, S.A.; Jeste, D.V. Epidemiology of and risk factors for psychosis of Alzheimer's disease: A review of 55 studies published from 1990 to 2003. *Am. J. Psychiatry* **2005**, *162*, 2022–2030.
20. Waltereit, R.; Banaschewski, T.; Meyer-Lindenberg, A.; Poustka, L. Interaction of neurodevelopmental pathways and synaptic plasticity in mental retardation, autism spectrum disorder and schizophrenia: Implications for psychiatry. *World J. Biol. Psychiatry* **2013**, *15*, 507–516.
21. Zakzanis, K.K.; Kielar, A.; Young, D.A.; Boulos, M. Neuropsychological differentiation of late onset schizophrenia and frontotemporal dementia. *Cogn. Neuropsychiatry* **2001**, *6*, 63–77.
22. Hirschfeld, R.M.; Williams, J.B.; Spitzer, R.L.; Calabrese, J.R.; Flynn, L.; Keck, P.E.; Lewis, L.; McElroy, S.L.; Post, R.M.; Rapport, D.J.; et al. Development and Validation of a Screening Instrument for Bipolar Spectrum Disorder: The Mood Disorder Questionnaire. *Am. J. Psychiatry* **2000**, *157*, 1873–1875.
23. Spitzer, R.L. Psychiatric diagnosis: Are clinicians still necessary? *Compr. Psychiatry* **1983**, *24*, 399–411.
24. Kranzler, H.R.; Kadden, R.M.; Babor, T.F.; Rounsaville, B.J. Longitudinal, Expert, All Data Procedure for Psychiatric Diagnosis in Patients with Psychoactive Substance Use Disorders. *J. Nerv. Ment. Dis.* **1994**, *182*, 277–283.
25. Organización Mundial de la Salud (OMS). WHO WMH-CIDI—The World Health Organization World Mental Health Composite International Diagnostic Interview. 2017. Available online: https://www.hcp.med.harvard.edu/wmhcidi/ (accessed on 20 November 2020).
26. Semler, G.; Witichen, H.-U.; Joschke, K.; Zaudig, M.; Geiso, T.; Kaiser, S.; Cranach, M.; Pfister, H. Test-Retest Reliability of a standardized psychiatric interview (DIS/CIDI). *Eur. Arch. Psychiatry Neurol. Sci.* **1987**, *236*, 214–222,
27. Sheehan, D.V.; Lecrubier, Y.; Sheehan, K.H.; Amorim, P.; Janavs, J.; Weiller, E.; Hergueta, T.; Baker, R.; Dunbar, G.C. The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* **1998**, *59* (Suppl. S2), 22–33.
28. Wing, J.K. SCAN. *Arch. Gen. Psychiatry* **1990**, *47*, 589.
29. Aboraya, A.; Nasrallah, H.; Muvvala, S.; El-Missiry, A.; Mansour, H.; Hill, C.; Elswick, D.; Price, E.C. The Standard for Clinicians' Interview in Psychiatry (SCIP): A Clini-cian-administered Tool with Categorical, Dimensional, and Numeric Output-Conceptual Development, Design, and Descrip-tion of the SCIP. *Innov. Clin. Neurosci.* **2016**, *13*, 31–77.
30. Endicott, J.; Spitzer, R.L. A Diagnostic Interview: The Schedule for Affective Disorders and Schizophrenia. *Arch. Gen. Psychiatry* **1978**, *35*, 837–844.
31. Nurnberger JIJr Blehar, M.C.; Kaufmann, C.A.; York-Cooler, C.; Simpson, S.G.; Harkavy-Friedman, J.; Severe, J.B.; Malaspina, D.; Reich, T. Diagnostic interview for genetic studies: Rationale, unique features, and training. *Arch. Gen. Psychiatry* **1994**, *51*. https://doi.org/10.1001/archpsyc.1994.03950110009002.
32. Ghaemi, S.N.; Miller, C.J.; Berv, D.A.; Klugman, J.; Rosenquist, K.J.; Pies, R.W. Sensitivity and specificity of a new bipolar spectrum diagnostic scale. *J. Affect. Disord.* **2005**, *84*, 273–277.
33. Depue, R.A. A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: A conceptual framework and five validation studies. *J. Abnorm. Psychol.* **1981**, *90*, 381–437.
34. Andreasen, N.C.; Flaum, M.; Arndt, S. The Comprehensive Assessment of Symptoms and History (CASH). *Arch. Gen. Psychiatry* **1992**, *49*, 615.
35. Simel, D.L.; Rennie, D.; Bossuyt, P.M.M. The STARD Statement for Reporting Diagnostic Accuracy Studies: Application to the History and Physical Examination. *J. Gen. Intern. Med.* **2008**, *23*, 768–774.
36. Whiting, P.F.; Rutjes, A.W.S.; Westwood, M.E.; Mallett, S.; Deeks, J.J.; Reitsma, J.B.; Leeflang, M.M.G.; Sterne, J.A.C.; Bossuyt, P.M.M.; QUADAS-2 Group. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann. Intern. Med.* **2011**, *155*, 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
37. Life Expectancy and Healthy Life Expectancy. Available online: https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/life-expectancy-and-healthy-life-expectancy (accessed on 13 May 2022).
38. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med.* **2009**, *6*, e1000097.
39. Janse, A.; Gemke, R.; Uiterwaal, C.; van der Tweel, I.; Kimpen, J.; Sinnema, G. Quality of life: Patients and doctors don't always agree: A meta-analysis. *J. Clin. Epidemiol.* **2004**, *57*, 653–661.

40. Regier, D.A.; Narrow, W.E.; Clarke, D.E.; Kraemer, H.C.; Kuramoto, S.J.; Kuhl, E.A.; Kupfer, D.J. DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *Am. J. Psychiatry* **2013**, *170*, 59–70.
41. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **2010**, *36*, 1–48.
42. Griffin, J.W. metapoweR: An R package for Computing Meta-Analytic Statistical Power. Available online: https://cran.r-project.org/package=metapower (accessed on 20 January 2023).
43. Unenge Hallerbäck, M.; Lugnegård, T.; Gillberg, C. Is autism spectrum disorder common in schizophrenia? *Psychiatry Res.* **2012**, 198, 12–17.
44. Adelufosi, A.O.; Adebowale, T.O.; Abayomi, O.; Mosanya, J.T. Medication adherence and quality of life among Nigerian outpatients with schizophrenia. *Gen. Hosp. Psychiatry* **2012**, *34*, 72–79.
45. Rafrafi, R.; Namouri, N.; Ghaouar, M.; Hsairi, M.; Melki, W.; El Hechmi, Z. Validity and reliability of the tunisian version of section g of the composite interna-tional diagnostic interview, relative to schizophrenia. *Tunis. Med.* **2013**, *91*, 648–654.
46. Yazici, E.; Cimen, Z.; Akyollu, I.I.U.; Yazici, A.B.; Turkmen, B.A.; Erol, A. Depressive Temperament in Relatives of Patients with Schizophrenia Is Associated with Suicidality in Patients with Schizophrenia. *Clin. Psychopharmacol. Neurosci.* **2018**, *16*, 302–309.
47. Nordgaard, J.; Revsbech, R.; Saebye, D.; Parnas, J. Assessing the diagnostic validity of a structured psychiatric interview in a first-admission hospital sample. *World Psychiatry* **2012**, *11*, 181–185.
48. Stewart, C.; El-mallakh, R.S. Is bipolar disorder overdiagnosed among patients with substance abuse? *Bipolar Disord* **2007**, *9*, 646–648.
49. Zimmerman M, Ruggero CJ, Chelminski I; et al. Is bipolar disorder Overdiagnosed? *J. Clin. Psychiatry* **2008**, *69*, 935–940.
50. Jon, D.-I.; Hong, N.; Yoon, B.-H.; Jung, H.Y.; Ha, K.; Shin, Y.C.; Bahk, W.-M. Validity and reliability of the Korean version of the Mood Disorder Questionnaire. *Compr. Psychiatry* **2009**, *50*, 286–291.
51. Vázquez, G.H.; Romero, E.; Fabregues, F.; Pies, R.; Ghaemi, N.; Mota-Castillo, M. Screening for bipolar disorders in Spanish-speaking Populations:Sensitivity and specificity of the Bipolar Spectrum Diagnostic Scale-Spanish Version. *Compr. Psychiatry* **2010**, *51*, 552–556.
52. Jiménez, E.; Arias, B.; Castellví, P.; Goikolea, J.; Rosa, A.; Fañanás, L.; Vieta, E.; Benabarre, A. Impulsivity and functional impairment in bipolar disorder. *J. Affect. Disord.* **2011**, *136*, 491–497.
53. Imamura, K.; Kawakami, N.; Naganuma, Y.; Igarashi, Y. Development of screening inventories for bipolar disorder at workplace: A diagnostic accuracy study. *J. Affect. Disord.* **2015**, *178*, 32–38.
54. Lee, D.; Cha, B.; Park, C.-S.; Kim, B.-J.; Lee, C.-S.; Lee, S. Usefulness of the combined application of the Mood Disorder Questionnaire and Bipolar Spectrum Diagnostic Scale in screening for bipolar disorder. *Compr. Psychiatry* **2012**, *54*, 334–340.
55. Suresh, K.S.; Roberts, R.J.; El-Mallakh, R.S. The Sensitivity and Specificity of the Mood Disorders Questionnaire Varies with the Intensity of Mood Illness. *Psychiatr. Q.* **2012**, *84*, 337–341.
56. Asaad, T.; Okasha, T.; Ramy, H.; Fekry, M.; Zaki, N.; Azzam, H.; Rabie, M.A.; Elghoneimy, S.; Sultan, M.; Hamed, H.; et al. Correlates of psychiatric co-morbidity in a sample of Egyptian patients with bipolar disorder. *J. Affect. Disord.* **2014**, *166*, 347–352.
57. Verhoeven, F.; Swaab, L.; Carlier, I.; van Hemert, A.; Zitman, F.; Ruhé, H.; Schoevers, R.; Giltay, E. Agreement between clinical and MINI diagnoses in outpatients with mood and anxiety disorders. *J. Affect. Disord.* **2017**, *221*, 268–274.
58. Ince, B.; Cansiz, A.; Ulusoy, S.; Yavuz, K.F.; Kurt, E.; Altinbas, K. Reliability and Validity Study of the Turkish Version of Bipolar Spectrum Diagnostic Scale. *Turk. J. Psychiatry* **2019**, *30*, 272–278.
59. Rajkumar, R.P. Recurrent unipolar mania: A comparative, cross-sectional study. *Compr. Psychiatry* **2016**, *65*, 136–140.
60. Wesley, M.S.; Manjula, M.; Thirthalli, J. Interepisodic Functioning in Patients with Bipolar Disorder in Remission. *Indian J. Psychol. Med.* **2018**, *40*, 52–60.
61. Hebbrecht, K.; Stuivenga, M.; Birkenhäger, T.; Van Der Mast, R.C.; Sabbe, B.; Giltay, E.J. Symptom Profile and Clinical Course of Inpatients with Unipolar versus Bipolar Depression. *Neuropsychobiology* **2019**, *79*, 313–323. https://doi.org/10.1159/000503686.
62. Hong, N.; Bahk, W.-M.; Yoon, B.-H.; Shin, Y.C.; Min, K.J.; Jon, D.-I. Characteristics of bipolar symptoms in psychiatric patients: Pattern of responses to the Korean version of the Mood Disorder Questionnaire. *Asia-Pacific Psychiatry* **2012**, *6*, 120–126/
63. Kung, S.; Palmer, B.A.; Lapid, M.I.; Poppe, K.A.; Alarcon, R.D.; Frye, M.A. Screening for bipolar disorders: Clinical utilization of the Mood Disorders Questionnaire on an inpatient mood disorders unit. *J. Affect. Disord.* **2015**, *188*, 97–100.
64. Nordgaard, J.; Sass, L.; Parnas, J. The psychiatric interview: Validity, structure, and subjectivity. *Eur. Arch. Psychiatry Clin. Neurosci* **2013**, *263*, 353–364.
65. Telles Correia, D.; Stoyanov, D.; Rocha Neto, H.G. How to define today a medical disorder? Biological and psychosocial disadvantages as the paramount criteria. *J. Eval. Clin. Pract.* **2021**, *00*, 2021–2022.
66. E Verdon, M.; Siemens, K. Yield of review of systems in a self-administered questionnaire. *J. Am. Board Fam. Pr.* **1997**, *10*, 20–27.
67. Collen, M.F.; Cutler, J.L.; Siegelaub, A.B.; Cella, R.L. Reliability of a Self-Administered Medical Questionnaire. *Arch. Intern. Med.* **1969**, *123*, 664–681.
68. Oliveira IC de, Nascimento I, Coutinho ESF; et al. Clinical stability, diagnosis and catchment area: The patients of a universi-ty-based psychiatric outpatient clinic. *J. Bras. Psiquiatr.* **2018**, *67*, 213–222.
69. Kraemer, H.C. Measurement of reliability for categorical data in medical research. *Stat. Methods Med. Res.* **1992**, *1*, 183–199.
70. Mitchell, P.B. Bipolar Disorder: The Shift to Overdiagnosis. *Can. J. Psychiatry* **2012**, *57*, 659–665.
71. Cantor N, Smith EE, French RD; et al. Psychiatric diagnosis as prototype categorization. *J Abnorm Psychol* **1980**, *89*, 181–193.

72. Reed, G.M.; Sharan, P.; Rebello, T.J.; Keeley, J.W.; Medina-Mora, M.E.; Gureje, O.; Ayuso-Mateos, J.L.; Kanba, S.; Khoury, B.; Kogan, C.S.; et al. The ICD-11 developmental field study of reliability of diagnoses of high-burden mental disorders: Results among adult patients in mental health settings of 13 countries. *World Psychiatry* **2018**, *17*, 174–186.
73. Kotwicki, R.; Harvey, P.D. Systematic Study of Structured Diagnostic Procedures in Outpatient Psychiatric Rehabilitation: A Three-year, Three-cohort Study of the Stability of Psychiatric Diagnoses. *Innov. Clin. Neurosci.* **2013**, *10*, 14–19.
74. Ash, P. The reliability of psychiatric diagnoses. *J. Abnorm. Soc. Psychol.* **1949**, *44*, 272–276.
75. Parnas, J. Differential diagnosis and current polythetic classification. *World Psychiatry* **2015**, *14*, 284–287.