The LEADING Guideline

Reporting Standards for Expert Panel, Best-Estimate Diagnosis, and Longitudinal Expert All Data (LEAD) Studies

Veerle C Eijsbroek, Katarina Kjell, H Andrew Schwartz, Jan R Boehnke, Eiko I Fried, Daniel N Klein, Peik Gustafsson, Isabelle Augenstein, Patrick M M Bossuyt, and Oscar N E Kjell

V.C. Eijsbroek, PhD candidate, Department of Psychology, Lund University, Lund, Sweden
K. Kjell, PhD candidate, Department of Psychology, Lund University, Lund, Sweden
H.A. Schwartz, associate professor, Department of Computer Science, Stony Brook University, New York, the United States
J.R. Boehnke, PhD, School of Health Sciences, University of Dundee, Dundee, Scotland
E.I. Fried, associate professor, Institute of Psychology, Leiden University, Leiden, the Netherlands
D.N. Klein, professor, Department of Psychology, Stony Brook University, New York, the United States
P. Gustafsson, associate professor, Faculty of Medicine, Lund University, Lund, Sweden
I. Augenstein, professor, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark,
P.M.M. Bossuyt, professor, Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, the Netherlands
O.N.E. Kjell, associate professor, Department of Psychology, Lund University, Lund, Sweden

Correspondence to V.C. Eijsbroek: veerle.eijsbroek@psy.lu.se

Abstract

Background

Accurate assessments of symptoms and illnesses are essential for health research and clinical practice but face many challenges. The absence of a single error-free measure is currently addressed by assessment methods involving experts reviewing several sources of information to achieve a more accurate or best-estimate assessment. Three bodies of work spanning medicine, psychiatry, and psychology propose similar assessment methods: The Expert Panel, the Best-Estimate Diagnosis, and the Longitudinal Expert All Data (LEAD) method. However, the quality of such best-estimate assessments is typically very difficult to evaluate due to poor reporting of the assessment methods and when they are reported, the reporting quality varies substantially. Here, we tackle this gap by developing reporting guidelines for such best-estimate assessment studies.

Methods

The development of the reporting guidelines followed a four-stage approach: 1) drafting reporting standards accompanied by rationales and empirical evidence, which were further developed with a patient organization for depression, 2) incorporating expert feedback through a two-round Delphi procedure, 3) refining the guideline based on an expert consensus meeting, and 4) testing the guideline by i) having two researchers test it and ii) using it to examine the extent previously published studies report the standards. The last step also provides evidence for the need for the guideline: 10 to 63% (Mean = 33%) of the standards were not reported across thirty randomly selected studies.

Results

The LEADING guideline comprises 20 reporting standards related to four groups: The *Longitudinal design* (four standards); the *Appropriate data* (four standards); the *Evaluation – experts, materials, and procedures* (ten standards); and the *Validity* group (two standards).

Conclusions

We hope that the LEADING guideline will be useful in assisting researchers in planning, conducting, reporting, and evaluating research aiming to achieve best-estimate assessments.

Keywords: Expert Panel; LEAD; Best-Estimate Diagnosis; Reference standard; Criterion standard; Gold standard; Medical assessments; Psychiatric assessments; Psychological assessments.

Open data (Delphi rounds 1 and 2), code (analyses), and material (surveys): https://osf.io/fkv4b/

Introduction

Establishing valid and reliable assessments of symptoms and diagnoses is the foundation of health and clinical sciences. Given that reliable biological markers or specific objective signs for most mental health problems are lacking and many medical conditions only show objective markers in late stages, accurate assessments are difficult[1,2]. Essentially every single measure of a psychological construct has some potential source of bias (e.g., self-report and recall bias) or can be seen as fallible in some respect[3,4] – which can result in inaccurate assessments and delayed treatments.

The absence of a single error-free measure can be addressed by involving multiple experts reviewing several sources of information to form a *best-estimate assessment* or a reference standard[5–7]. To understand the quality of such an assessment, it is crucial to understand how it was reached (i.e., the quality of the specific assessment method used). However, the quality of best-estimate assessments is typically very difficult to evaluate due to poor reporting of the assessment method, and when the method is reported, the reporting quality varies substantially[7]. Here we tackle this problem by developing a guideline for how to report assessment methods that aim to achieve such a best-estimate assessments, i.e., where experts review several sources of (longitudinal) information to achieve a more accurate assessment than a single, error-prone measure.

Assessment

Assessment includes the evaluation, integration, and interpretation of several sources of information (e.g., outcomes of different measures, tests, or scans) to derive a valid and reliable decision (e.g., a best-estimate diagnosis)[8]. Accurate assessments are crucial for understanding the prevalence of clinical problems[9,10], detecting and starting early treatment[11,12], validating measurement tools[13,14], and evaluating interventions/therapies[15,16]. In clinical practice, under- or over-estimation of illnesses can have severe negative impacts on people's lives. In research, they threaten the validity of scientific results. For policy and implementation development, assessments are the basis for guideline development and the economic and societal evaluations of interventions. Furthermore, obtaining more accurate assessments has become increasingly important considering that high-accuracy assessments are needed in diverse fields such as Biological Psychiatry (e.g., to find reliable biomarkers linked to reference standard assessments[17–19]) and Artificial Intelligence (e.g., to train models to reference standard assessments[20–22]).

A methodological solution

Here, we connect three bodies of literature that have proposed similar assessment methods: The Expert Panel method in medicine[7,23,24] – as well as the Best-Estimate Diagnosis[6,25] and the Longitudinal Expert All Data (LEAD)[5] methods in psychiatry and clinical psychology. The three methods share the same goal of attaining best-estimate assessments through similar methodological approaches: All three methods use expert panels or consensus teams (e.g., clinical psychologists or medical doctors) to review several sources of information (e.g., clinical questionnaires and medical tests) to establish a more accurate assessment (e.g., for diagnostic purposes in clinical practice, or as a reference standard in statistical modeling). The three methods accentuate different parts of this methodological approach: The Best-Estimate Diagnosis method accentuates the use of informants and objective tests next to self-reported data[6,25]; the Expert Panel method focuses on the characteristics, constitution, and procedure of the panel[7,23] and only the LEAD method requires a longitudinal design[5], although longitudinal data are also used in some Expert Panel designs (≈27% of studies[7]). Herein we collectively refer to these three approaches as the *assessment methods*.

The result of the *assessment methods* is a consensually derived criterion (e.g., a best-estimate assessment) that has been used for many different applications where there is no single error-free measure. It has, for

example, been used to i) understand the accuracy of a measurement tool or marker through comparison to a best-estimate assessment[26–31]; ii) establish the prevalence of symptoms and disorders[9,10,32]; iii) establish the temporal stability or development of symptoms and disorders [33–35]; iv) improve (earlier) detection or screening of symptoms or disorders[11,12,36]; v) study genetics and family history[37–39]; and vi) examine classification systems or diagnostic criteria[40–42]. The applications span diverse fields, including medicine, psychiatry, clinical psychology, public health/epidemiology, and artificial intelligence. Box 1 provides more examples of how the assessment methods have been applied in different types of studies across fields.

Box 1 | Overview of applications of the assessment methods across different study designs and fields

The absence of a single error-free measure can be mitigated by involving multiple experts reviewing several sources of information to form a best-estimate assessment. To assist users in planning, evaluating, and reporting assessment method procedures to derive such best-estimate assessments, we developed a guideline with stakeholders from areas in which the assessment methods are used, such as medicine, psychiatry, clinical psychology, and epidemiology. The use cases below show study designs and fields where the guideline is applicable and relevant for reporting the assessment method procedure. For example, the assessment methods have been applied:

1. To evaluate a measure's accuracy against a reference standard. To understand the accuracy of a measurement tool, there is a need to compare it to a more accurate or best-estimate assessment. For example, it has been used:
- *in psychiatry*, for evaluating MINI-KID diagnoses for children and adolescents[26]
  and evaluating DSM diagnoses in patients with psychosis.[27]
- *in clinical psychology*, for evaluating Major Depression Inventory (MDI) severity scores.[28]
- *in medicine*, for evaluating deep learning models assessing liver cancer[29]
  and evaluating prediction rules for coronary artery disease.[30]
- *in public health/epidemiology*, for evaluating electronic health record algorithms for assessing asthma.[31]

2. To establish the prevalence of symptoms or disorders. For example, it has been used:
- *in public health/epidemiology*, for assessing the prevalence and familiality of pathological gambling.[9]
- *in psychiatry*, for assessing the prevalence of eating disorders in patients with personality disorders.[10]
- *in medicine*, for assessing the prevalence of clinically relevant incidental findings when diagnosing pulmonary embolism.[32]

3. To establish the temporal stability or development of symptoms or disorders. For example, it has been used:
- *in clinical psychology*, for learning about autism spectrum disorder diagnoses during childhood[33]
  and for learning about the course of bipolar disorder.[34]
- *in psychiatry*, for assessing diagnostic stability in individuals with autism spectrum disorder.[35]

4. To improve (earlier) detection or screening of symptoms or disorders. For example, it has been used:
- *in psychiatry*, for the assessment of personality disorders.[11]
- *in medicine*, for the early detection of heart failure.[36]
  and for early detection of injuries in physically abused older adults.[12]

5. To study genetic history and family heritability. For example, it has been used:
- *in psychiatry*, for learning about genetic risks for ADHD.[37]
- *in clinical psychology*, for studying familial aggregation and heritability of subtypes of depression[38]
  and for studying the familial transmission of mania and depression.[39]

6. To examine classification systems or diagnostic criteria. For example, it has been used:
- *in clinical psychology*, for evaluating the DSM criteria for hoarding disorder.[40]
- *in psychiatry*, for examining a DSM alternative model for personality disorders[41]
  and for comparing DSM-IV and -5 criteria of autism spectrum disorders.[42]

Description of an example study including a best-estimate assessment
Best-estimate assessments have, for example, been established to evaluate the validity of the Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL; a semi-structured diagnostic interview used in child and adolescent psychiatry)[43]. The best-estimate assessments were diagnoses of neurodevelopmental and related disorders made by five experienced child psychiatrists based on the DSM-5 criteria. To achieve best-estimate assessments, patients were followed for at least three months, and the psychiatrists had access to all available data (except for the K-SADS-PL diagnoses), including information from medical records, interviews, questionnaires, laboratory tests, as well as information provided by clinical staff, caregivers and teachers. Criterion validity of the K-SADS-PL was established as the agreement of the diagnoses with the best-estimate assessments.

Reporting issues

The assessment methods possess a high potential for achieving best-estimate reference standards in many situations. However, the quality of such proclaimed best-estimate assessments varies substantially and is typically very difficult to evaluate due to poor reporting of the method *how* they were achieved (e.g., see reviews of expert panels[7,23]). A systematic review of assessment methods and reporting of expert panels[7] has demonstrated that the methods used for panel or consensus diagnoses vary substantially across studies and that many aspects of the procedure are often unclear or not reported at all. Many recent studies fail to report central aspects of the three assessment methods, including the quality, structure, or presentation of the data[44], the training and qualifications of the experts[45], the method for avoiding biases and achieving consensus[13], and the time span of the longitudinal design-component[46]. The poor operationalization of the assessment methods jeopardizes the goal of achieving best-estimate assessments – where a vaguely described method makes it difficult to evaluate the research. Referring to an assessment as a best-estimate (and sometimes even as a gold standard) while vaguely describing or poorly operationalizing the method for achieving the assessment is alarming[47,48].

The degree of validity

These assessment methods aim to achieve high validity (i.e., the degree to which the assessment captures what it aims to measure). Typically, the assessment methods aim to achieve as high validity as possible (i.e., a "leading" assessment), or, depending on resources, at least more accurate than a single error-prone measure. Despite this central aim, research often fails to clearly describe the degree of validity of the attained assessment. Using these assessment methods does not automatically guarantee high validity – it depends on how well the method is executed.

In addition, the assessments are often described with different terms: *reference standard* is often used in medicine, and *criterion standard* or *best-estimate diagnosis* is often used in psychology. We propose that the reporting of these assessment methods benefit from more explicitly describing *what* was measured and *how* well it measures up to different standards – whether and how they relate to a state-of-the-art assessment. Whereas *reference* and *criterion standards* fail to convey an intention of "nearing" a state-of-the-art assessment, the *best-estimate diagnosis* narrowly focuses on the classification of a diagnosis and not on symptom severity. Therefore, we here use the term *best-estimate assessment* in the context of describing a "leading", state-of-the-art assessment.

Reporting standards

Previous well-established guidelines have focused on the complete reporting of specific study designs, such as the *STrengthening the Reporting of OBservational studies in Epidemiology* (STROBE)[49] for observational studies; the *Statement for Reporting for Diagnostic Accuracy* (STARD)[50] for diagnostic accuracy studies; the *Consolidated Standards of Reporting Trials* (CONSORT)[51] for randomized trials, and

the *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis* (TRIPOD)[52] for prediction model studies (see the supplementary material [SM] for other relevant guidelines). The STARD guidance is most closely related to the reporting of the assessment methods since best-estimate assessments are often used to evaluate a measure's (diagnostic) accuracy. However, none of the guidelines are sufficient for complete reporting of the assessment methods, where (multiple) experts review several sources of (longitudinal) information to form a best-estimate assessment. Although an earlier systematic review identified and structured the various choices involved in Expert Panel procedures[7], no attempt was made to develop a formal guideline for the reporting of Expert Panel assessments.

Aim

Our aim is to develop reporting standards for comprehensive reporting of these assessment methods – which can help researchers plan, carry out, and report studies employing these assessment methods, as well as help readers evaluate them. We call the reporting guideline the LEADING guideline, emphasizing the methodological components and the importance of describing *what* was assessed and *how well* (i.e., how it relates to a "leading" assessment). We further revise the original meanings of LEAD (*Longitudinal, Expert, All Data*[5]) to *Longitudinal*, *Evaluation – experts, materials and procedures*, and *Appropriate* Data). In short, the LEADING guideline aims to guide the reporting of assessment methods to improve evaluations of the assessment standard.

Methods

Development stages

We developed the reporting guideline over four stages: 1) drafting reporting standards; 2) incorporating expert feedback; 3) refining the final guideline, and 4) testing the guideline. The development method largely followed Moher and colleagues' guidance for developing reporting guidelines[53] (See Table S1 for elaborations on each recommended step). For organizational purposes, a working group (V.E., K.K., & O.K.) was set up, and a steering group (H.A.S., J.B., E.F., D.K., P.G., I.A., & P.B.) was formed to provide a wide range of expertise. The steering group included seven experts and was selected to cover a diverse range of expertise and fields related to the assessment methods (e.g., psychiatry/clinical psychology, medicine, epidemiology/public health, and Artificial Intelligence). See the declarations section for information regarding ethics.

Drafting reporting standards

The working group drafted the original reporting standards. First, the working group, with the support of the steering group members, identified relevant research using or describing the assessment methods, including the three bodies of literature: Expert Panel[7], Best-Estimate Diagnosis[6], and LEAD[5]. Second, articles using any of the three assessment methods were identified through a literature search using Google Scholar with the following search terms: ["expert panel diagnosis" OR "expert panel assessment" OR "expert panel consensus" OR "expert panel methodology" OR "expert panel standard" OR "expert panel reference"] for Expert Panel studies; ["best-estimate diagnosis" OR "best-estimate diagnostic" OR "best-estimate standard" OR "best-estimate assessment" OR "best-estimate methodology" OR "best-estimate reference"] for Best-Estimate Diagnosis studies; and ["longitudinal expert all data" OR "longitudinal evaluation all data"] for LEAD studies. Articles that clearly stated the use of one of the three assessment methods were selected. Articles stating another purpose than assessment (e.g. when an expert panel was used to reach a consensus about a treatment strategy) were excluded.

Third, relevant reporting guidelines and systematic reviews were identified, including a review of expert panels applications[7]; the STROBE statement[49]; and the STARD guidance[50] (other complementary reporting guidelines and systematic reviews are presented in the SM). The aim was for the reporting standards in the LEADING guideline to complement rather than repeat these guidelines (i.e., new standards should extend or complement existing standards rather than repeat them[53]). The use of multiple reporting guidelines may often be appropriate; for example, when reporting a randomized trial that includes best-estimate assessments, one may use CONSORT[51] to report the trial design and main results, the LEADING guideline for describing the specifics for reaching the best-estimate assessment, and the *Consolidated Health Economic Evaluation Reporting Standards* (CHEERS)[54] for reporting the economic evaluations and comparisons.

Potential standards were drafted by the working group with the objective of encompassing a comprehensive reporting of the assessment methods. The reporting standards were grouped into four groups: *Longitudinal design*, *Appropriate data*, *Evaluation – experts, materials and procedures*, and *Validity*. Empirical and theoretical inclusion rationales were stated for the groups and the individual standards (i.e., explanations and elaborations). Lastly, the standards with inclusion rationales were further developed through a workshop with a patient organization for depression, followed by receiving feedback from the steering group members to receive a wide range of perspectives early in the process.

Incorporating expert feedback

To systematically collect expert feedback from different perspectives, we used a consensus-building procedure called the *Delphi technique*[55]. We used an iterative process based on two rounds of questionnaires (i.e., Delphi surveys), enabling feedback from round 1 to feed into round 2. Delphi participants received relevant background research, the reporting guideline aims, as well as the groups and the individual standards with their inclusion rationales. They provided feedback through open- and closed-ended response formats. Through open-ended responses in Round 1, experts could propose new standards and provide feedback on the formulations of existing standards, their inclusion rationales, and evidential support. In addition, two closed-ended questions[56] about standard inclusion (*This item should be included in the reporting checklist*) and perception of study quality (*Whether this information is present or not would influence my perceptions of the quality of a study*) were answered with rating scales ranging from 1 = *Strongly disagree* to 7 = *Strongly agree*. In Round 2, the experts were asked to rate the updated reporting standards using the same two closed-ended questions as in Round 1 and to provide feedback through open-ended responses (e.g., feedback on the clarifications and reformations of the standards that resulted from Round 1).

The first and/or last authors of articles since 2013 using any of the three assessment methods were identified using the search terms described above (n = 87 articles, n = 124 authors; see the SM for more details). These authors and the seven steering group members were invited via email to participate in the Delphi Round 1 (n = 131 participants emailed). In total, 27 participants completed the survey (response rate 21%). Only participants from Round 1 who provided their contact details were invited to Round 2 (n = 25). In total, 20 participants completed the survey (response rate 80%). All participants provided their informed consent. Figure 1 presents the research experiences and demographics of the Delphi participants. Participants reported a wide range of academic backgrounds (e.g., Medicine, Psychiatry or Clinical Psychology,

Artificial Intelligence, Journal Editors), and an extensive variety of relevant methodological experiences (e.g., Biological Markers, Ecological Momentary Assessments, and Expert Panels; Figure 1), with an age range of 30 – 70 years (*M* = 51.54, *SD* = 12.40).

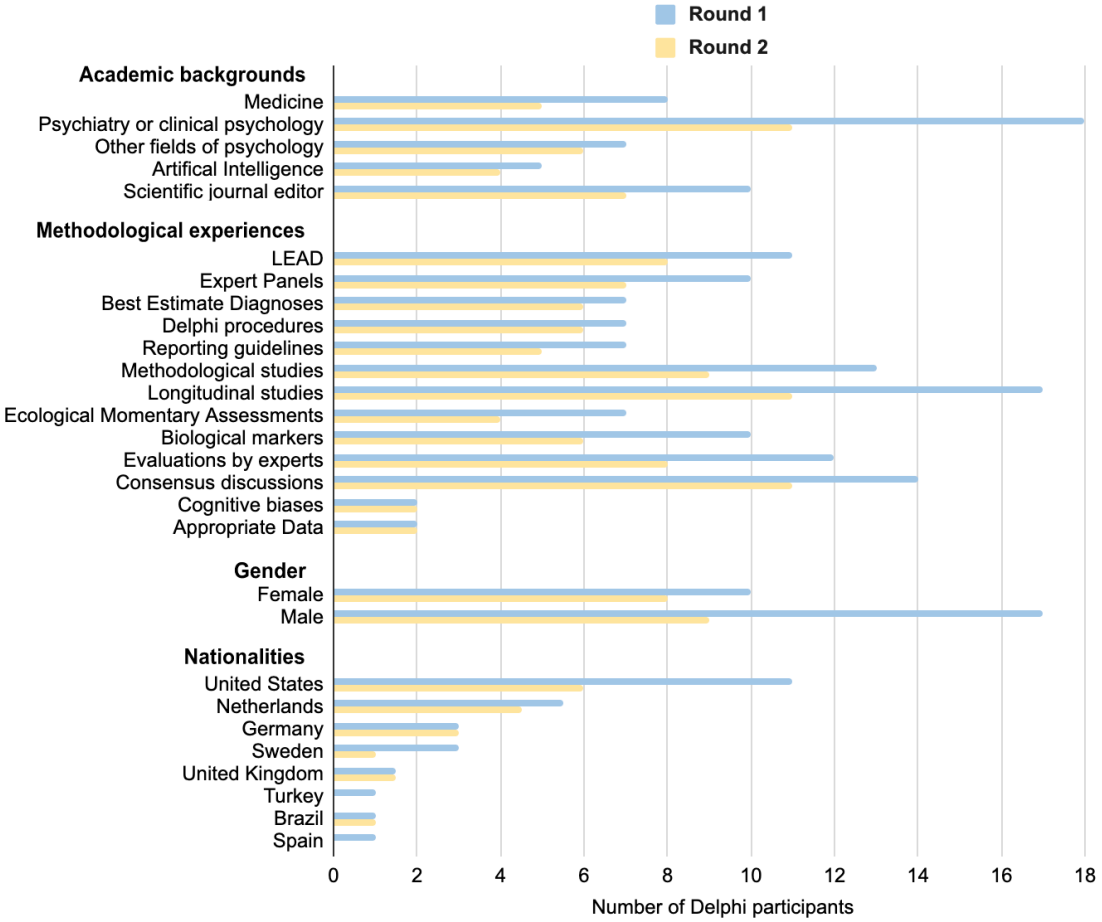

Figure 1 | *Research experiences and demographics of the Delphi participants.*
Notes. The answer options for Academic backgrounds and Methodological experiences were not mutually exclusive (i.e., multiple backgrounds and experiences could be reported by the participants). In Round 2, the demographics and reported experiences are known for 17 of the 20 participants.

Delphi survey results. The criteria for including a reporting standard was that the median of Delphi expert responses was at least *6 = Agree* on the question about its inclusion. In Round 1, the mean ratings for the *item inclusion* scale ranged from 5.37 - 6.67 (*M* = 6.06; *SD* = 0.31; Table S2) with a median agreement ranging from *6 = Agree* to *7 = Strongly Agree.* No new standards were suggested. The feedback resulted in the removal of one reporting standard and the clarification and reformulation of 20 standards. The standard *on Transparency and replicability* was rated as relevant but removed because it is achieved by reporting the other reporting standards. Standard *4.2 Validity and Standard* needed a major clarification about the meaning of validity as well as standard. Minor clarifications and reformations, such as grammar, or word changes, were made for 19 standards (see open material). The mean ratings in Round 2 ranged from 5.47 – 6.70 (*M* = 6.20; *SD* = 0.37; Table S3), with open feedback resulting in minor clarifications and reformulations of nine standards.

Refining the final guideline

The authors finalized the guidelines in an expert consensus meeting. The meeting was held online with nine working and steering group members. The content and structure of the consensus meeting were prepared by the working group, and the meeting was led by the last author (O.K.). Participants had access to the guidelines, elaboration and explanation (inclusion) rationales, and the drafted paper before the meeting, where they also had the option to provide comments and feedback in writing. The meeting included going through the Delphi Rounds 1 and 2 findings and discussing the paper draft, including the individual reporting standards and groups. We decided not to carry out another Delphi round since i) the median agreement for each reporting standard in both Delphi Rounds 1 and 2 ranged from *6 = Agree* to *7 = Strongly Agree*, ii) no new standards were suggested, and iii) only minor changes were needed after Round 2, which taken together suggest consensus.

Testing the guideline

To pilot the applicability of the guideline, the guideline was tested i) by independent researchers with experience of each method piloting the reporting of each standard and ii) by the authors (V.E., K.K.) using it to evaluate published articles. The two test procedures resulted in adding minor clarifications to three standards (*2.4 The access to the index measure, 3.3 Blindness and conflict of interest,* and *3.4 Instructions and training*). Also, a concrete example of how to report the items was added to the general guideline instructions.

Incorporating test-user feedback. Two test users (Ph.D., with experience using the LEAD and Expert Panel method) who had not been involved in the development of the guideline (e.g., in the Delphi procedure) were recruited to test the guideline (see the SM for more details). They were asked to report each standard based on a finished, ongoing or planned study using one of the asssessment methods and/or provide feedback about the formulation of the standards.

Reports of the standards in 2022 and 2023. Three separate targeted searches (LEAD, Expert-panel, Best-estimate) were conducted using the search terms described above, and the first author (V.E.) examined which standards were reported in 30 randomly selected articles applying the assessment methods in 2022 and 2023 (i.e., five from each method from each year; see the SM for the selection process). Each reporting standard was rated using four categories: standard *not reported*; standard *reported vaguely or insufficiently* ; standard *(minimally) sufficiently reported*; or standard *not applicable to the study*. Out of the 30 articles, six were randomly selected (i.e., one from each method from each year) and examined by the second author (K.K.) to get insight into the accuracy of the ratings of the first author. Discussing their disagreements to reach consensus resulted in changing 23 ratings (19%) of the first author. This testing procedure also provided information about the strengths and shortcomings of contemporary reporting of published articles using these methods (see Results section).

Results

The reporting guideline is presented in Table 1 (see Figure 2 for an overview). It comprises 20 standards for comprehensive reporting of the assessment methods divided into four groups: 1. *The Longitudinal design* group (4 standards), 2. *The Appropriate data* group (4 standards), 3. *The Evaluation – experts, materials, and procedures* group (10 standards), and 4. *The Validity* group (2 standards). The reporting standards encourage researchers to elaborate on what was done and why – whilst avoiding normative standards, such as a minimum number of experts. Each standard description in Table 1 is accompanied by an example. Further *Explanations and Elaborations* regarding the individual reporting standards and the four groups are presented in the SM, including Tables S4 and S5.

Table 1 | The LEADING guideline reporting standards

Group	#	Reporting standards
Longitudinal Design *Report the longitudinal design, by describing:*	1.1	**The time period**. The data collection period covered for each participant (i.e., start and end of the data collection) and to what extent the length is sufficient for capturing the targeted symptoms. *For example, the weeks/months a participant is followed and how this matches the criteria for the targeted disease/disorder.*
	1.2	**The number of time points**. Whether and how data were collected on multiple occasions between the start and the end of the time period, the sufficiency of the data collection, and of its frequency and intensity for capturing the target. *For example, report the number of check-ins with the participants and the included measures for each assessment.*
	1.3	**History or lifetime information**. Whether and which data from before the start of the data collection were taken into account and how these data are relevant for the assessment of the target. *History or lifetime data may include self-report of medical history, childhood memory accounts, or other-than-self information such as from relatives or from medical records.*
	1.4	**The targeted time point(s) of the experts' assessment**. The time point(s) for which the experts provide their assessment, on which time period the data of the assessments are based (i.e., past data, future data, or both), and justifications for the targeted time point(s). *For example, the experts can assess the presence of a diagnosis at the start of the study and thus base their assessment on future data from that reference point; or in the middle of the study time period and thus have access to both past and future data from that reference point.*
Appropriate data *Report the appropriate-ness of the data, by describing:*	2.1	**The type and quality of the data.** The type, quality, and relevance of the data and why these data sources are sufficient and suitable for capturing the target. *For example, describe the validity and reliability of the data – and how it relates to capturing the targeted construct.*
	2.2	**The data triangulation.** Whether and why the data come from different methodological approaches and the degree to which these approaches complement each other. *For example, how self-reported data is complemented by objective/physical tests and/or other informant data.*
	2.3	**The data presentation**. How the data were structured and presented to the experts for their assessments and why. *For example, were the data presented in a case report; and was the information presented with or without any interpretation?*
	2.4	**The access to the index measure**. For an assessment accuracy study, the extent the experts had access to the index measure and why (i.e., an assessment that is being compared to the best-estimate assessment), and how its information was weighted in their assessment. *For example, were experts blind to the measure (its outcome and/or its raw data) that is being validated?*

	3.1	**The expert and panel characteristics.** The characteristics of the experts and the panel, as well as how these characteristics are relevant for assessing the target. *Relevant characteristics may include clinical and research experiences, professions, education, and demographics.*
	3.2	**The number of experts and panels.** The total number of experts and panels, and how many experts/panels were assessing each case and why. *For example, how many and are the same expert(s)/expertise(s) present in every assessment?*
	3.3	**Blindness and conflicts of interest.** Whether and to what extent the experts are blind to the research aims and/or have any conflicts of interest. *This may include experts' study authorship or the experts' relationship to the index measure or any other assessment method. If the study examines an index measure (i.e., an assessment that is being compared with the best-estimate assessment), declare the authors' as well as the experts' relationship to it.*
	3.4	**Instructions and training.** The instructions, training, and/or preparation that the experts specifically received for this assessment task and why they did or did not receive this. *For example, provide information regarding 1) whether the assessment method and procedure are kept standardized across the individual assessments, 2) the methods to ensure experts' preparedness for the assessment, or 3) any specific measures to limit biases.*
Evaluation – experts, materials and procedures *Report the evaluation experts, materials and procedures, by describing:*	3.5	**The assessment procedure.** The procedure that the experts followed for their assessment. *For example, describe whether there was a standardized procedure and what this procedure included (such as following clear diagnostic criteria).*
	3.6	**The assessment response format.** The response format used by the experts for their individual assessments, what it included, and how it was structured. *For example, describe any assessment sheet, including assessment questions and answer options.*
	3.7	**The data combination method.** The method or guidelines for how the data should be weighted, judged, and combined by the individual experts to reach a conclusion in their individual assessment. *For example, should any data sources be evaluated first or weighted more strongly; or are the experts asked to assess certain diagnostic criteria/symptoms first, before forming a final diagnosis?*
	3.8	**Independent expert assessments.** Whether and how the experts first evaluated the data individually and made their first individual assessments independently. *For example, how it was ensured the experts first reviewed the data individually/independently before discussing their assessment outcome with the other panel members.*
	3.9	**The inter-rater and inter-panel reliability.** The inter-rater/inter-panel reliability, how it was calculated and evaluated, or why it was not possible to calculate it. *For example, which reliability metric was used and over how many experts/panels and cases the reliability was calculated.*
	3.10	**The solution to disagreements.** The approach for solving (any) disagreements between the individual expert assessments, the rationale for the chosen approach, and potential problems that may have occurred and how these were assessed. *Methods may include reaching a consensus, taking the average, or majority vote. Potential problems may, for example, include power imbalances in the expert panel.*
Validity *Report what was assessed and how well, by describing:*	4.1	**The assessment description.** Description of *what* the assessment actually is. *For example, is the assessment a diagnosis, symptom severity assessment, course of illness assessment, or treatment response assessment?*
	4.2	**The validity and standard.** Reflect on the degree of validity and describe the standard that the method aims to achieve, *how* well the assessment method measures up to that degree, and how it compares with current standards.

For example, reflect on evidence supporting or against validity aspects such as construct, face, and criterion validity; and state whether the assessment should be seen as a best-estimate assessment standard or an accepted reference standard (see Table S2 for more examples).

Instructions. The LEADING guideline comprises these 20 reporting standards for comprehensive reporting of assessment methods involving expert(s) reviewing several sources of information (over time) to achieve a more accurate assessment (e.g., see Expert Panel, Best-Estimate diagnosis, and Longitudinal Expert All Data methods). The standards aim to help researchers plan, carry out, and report studies employing these assessment methods, as well as help readers evaluate them. As such, avoid simply answering yes or no to the standards when you instead can (succinctly) *describe* justifications and courses of action. Ensure the reports of the standards are clear, specific, and justified. To exemplify, standard *1.1 The time period* could be reported as '*The time span was six weeks, which covers more than the two weeks a person should have the symptoms for meeting the criteria for Major Depressive Disorder according to the DSM-5.*'

Not all of the reporting standards will be applicable to all types of studies – however, it is typically better to describe how a standard is not applicable than to leave the information out. Since the guideline covers the reporting of the assessment method, the method section would suit the reporting of most standards in most cases. However, the reporting guideline does *not* standardize where standards should be reported. When standards are considered less relevant or not applicable to a specific study, they can, for example, be described in an Appendix. Since the guideline focuses specifically on the reporting of the assessment method, it is recommended to use a complementary guideline for the reporting of the other study components: Which complementary guideline is dependent on the study type in which the assessment method is employed (e.g., see STARD for diagnostic accuracy studies; STROBE for observational studies; and CONSORT for randomized trials).

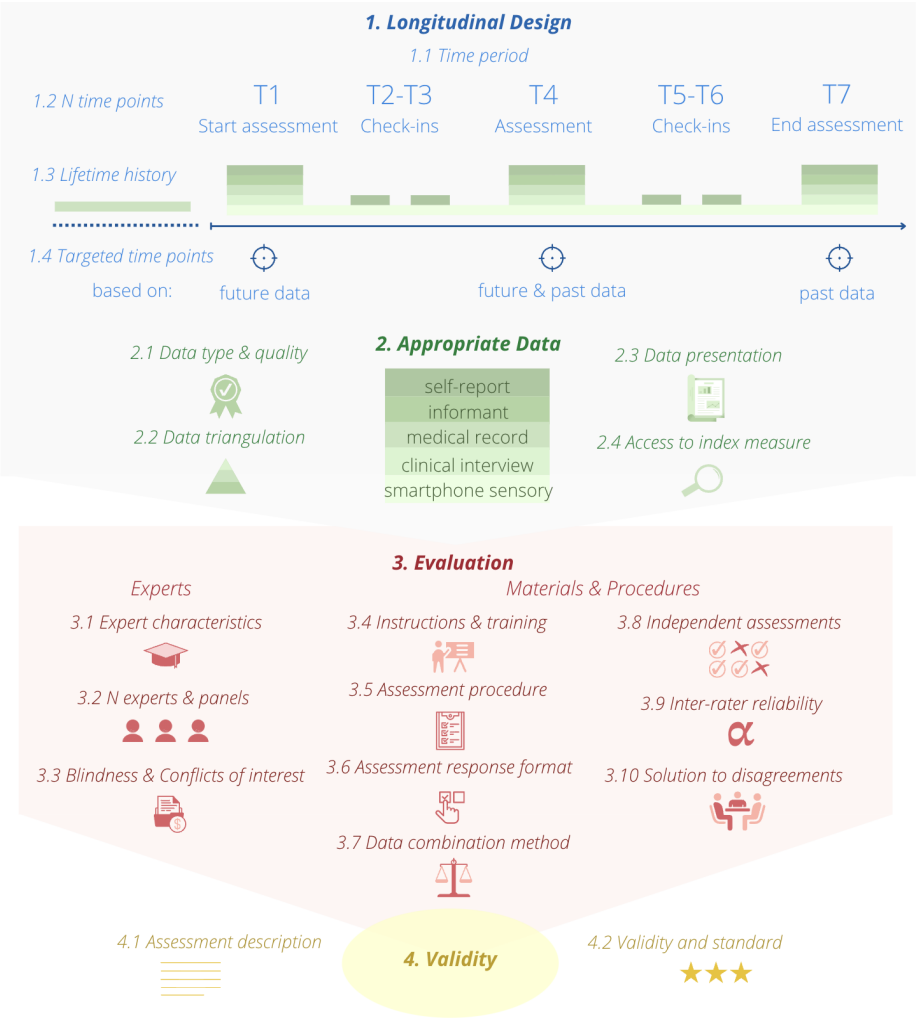

Figure 2 | *Overview of the LEADING guideline reporting standards.* For more details about each standard, see Table 1.

Using the LEADING guideline to evaluate published studies

Evaluating a random selection of 30 articles indicates severe heterogeneity in *what* of the methods is reported and *how*, suggesting the need for a guideline that enables comprehensive reporting of these assessment methods (Table 2; see the SM for the search strategy). Across the 30 studies, 10 to 63% (Mean = 33%) of the standards were *not* reported. Regarding the reporting standards, the type and quality of the data (2.1), the access to the index measure (2.4), the expert and panel characteristics (3.1), the number of experts and panels (3.2), the assessment procedure (3.5), and the assessment description (4.1) were mostly reported (i.e., green in more than 50% of the studies). However, the data presentation (2.3), the instructions and training (3.4), the data combination method (3.7), the inter-rater and inter-panel reliability (3.9), and the validity and standard (4.2) were *not* reported at all in the majority of the studies (i.e., red in more than 50% of the studies). Considering that most changes that resulted from discussing disagreements between the raters (V.E. and K.K.) were from green to orange, and that green refers to a (minimally) sufficiently reported and orange to insufficiently reported, this suggests that the table is conservative in regards to the severity of the current state of poor reporting (i.e., potentially showing a more positive picture; for more information see the SM).

Table 2 | Reports across the LEADING guideline standards in 30 randomly selected articles published in 2022 and 2023

	LEAD*										Expert Panel*										Best-Estimate Diagnosis*											
	2022						2023				2022						2023				2022						2023					
LEADING reporting standards	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**green %**	**red %**
1.1 The time period																															43	23
1.2 The number of time points																															40	37
1.3 History or lifetime information																															43	20
1.4 The targeted time point(s)																															47	13
2.1 The type and quality of data																															67	7
2.2 The data triangulation																															50	10
2.3 The data presentation																															10	70
2.4 The access to index measure																															75	0
3.1 The expert and panel characteristics																															70	3
3.2 The number of experts and panels																															67	30
3.3 Blindness and conflicts of interest																															20	23
3.4 Instructions and training																															13	73
3.5 The assessment procedure																															70	7
3.6 The assessment response format																															40	37
3.7 The data combination method																															0	83
3.8 Independent expert assessments																															33	47
3.9 The inter-rater and inter-panel reliability																															31	52
3.10 The solution to disagreements																															46	36
4.1 The assessment description																															97	0
4.2 The validity and standard																															13	63
green %	60	26	16	16	25	37	58	26	53	47	50	70	45	40	70	45	50	45	55	50	42	65	42	40	16	35	37	26	70	37		32[1]
red %	30	47	58	53	45	26	22	63	26	32	20	15	30	35	15	15	35	45	20	20	32	18	32	30	37	35	42	53	10	47	33[2]	

Notes. **red** = not reported; **orange** = insufficiently reported; **green** = (minimally) sufficiently reported; **gray** = not applicable to report. [1] Mean of reporting standards; [2] Mean of studies.
* Ten articles for each assessment method were randomly selected (see SM for selection process): 1 = Morrison et al. (2022)[14]; 2 = Mackenhauer et al. (2022)[57]; 3 = Hendriks et al. (2022)[58]; 4 = Paap et al. (2022)[59]; 5 = Aydin et al. (2022)[60]; 6 = Rocha Neto et al. (2023)[61]; 7 = Spangenberg et al. (2023)[62]; 8 = Muris et al. (2023)[63] 9 = Sveen et al. (2023)[64]; 10 = Pedersen et al. (2023)[65]; 11 = Sadleir et al. (2022)[66]; 12 = Khan et al. (2022)[67]; 13 = Blackmore et al. (2022)[13]; 14 = Loots et al. (2022)[68]; 15 = Leroux et al. (2022)[69]; 16 = Kocks et al. (2023)[70]; 17 = Lacroix et al. (2023)[71]; 18 = Nienhuis et al. (2023)[72]; 19 = Himmelreich et al. (2023)[73]; 20 = de la Matta et al. (2023)[74]; 21 = Peterson et al. (2022)[75]; 22 = Reiersen et al. (2022)[76]; 23 = Bradshaw et al. (2022)[77]; 24 = Hesam-Shariati et al. (2022)[78]; 25 = Shima et al. (2022)[79]; 26 = Kvig et al. (2023)[80]; 27 = Detera-Wadleigh et al. (2023)[81]; 28 = Hill et al. (2023)[82]; 29 = Jones et al. (2023)[83]; 30 = Dimian et al. (2023)[84]. The articles cover different disciplines and fields, including medicine[14,66–74,81], psychiatry or clinical psychology[13,57–65,75–84], and Artificial Intelligence[68,70,73].

Discussion

Our objective was to develop a guideline that supports comprehensive reporting of studies collecting longitudinal, appropriate data that experts evaluate to achieve an assessment that is more accurate than a single error-prone measure. The aim is to help researchers plan, report, and evaluate the assessment method-related elements of this study design.

The LEADING reporting standards were established through an open process, incorporating relevant empirical evidence and methodological work, complementary reporting guidelines, and comprehensive iterations of expert feedback and patients' perspectives. As this guideline focuses on the assessment methods, we recommend that researchers also rely on established guidelines for other parts of their research, such as sampling and other epidemiological aspects (e.g., STROBE[49], CONSORT[51], and STARD[50]). We encourage knowledge about and adherence to the LEADING guideline via scientific journals, editorials, and the EQUATOR network, as well as inclusion in research method courses in clinical studies.

Limitations

We have connected three assessment methods with similar approaches from related fields and drafted applicable reporting standards. We presented the rationale for selecting these three methods and each reporting standard with supporting evidence in the Delphi survey for review, which did not bring up additional methods or reporting standards. However, as we did not carry out a systematic literature review of the three identified literature bodies or for each of the reporting standards, we cannot exclude the existence of other assessment methods with similar approaches: We welcome any suggestions about similar methods to which the guideline is applicable.

Although the Delphi survey participants and the author group had a wide range of experiences and backgrounds, psychiatry and clinical psychology (n = 18) were overrepresented as compared to, for example, other areas of medicine (n = 8) in the Delphi and author group. Geographically, Europe and North America were the most common in the Delphi and author group, whereas several areas were not represented. The Delphi participants were the first or last authors of studies employing the assessment methods. However, the quality of the articles, and the education or experience of the authors, were not taken into account as selection criteria (although it was self-reported as presented in Figure 1). Finally, the number of Delphi participants (27 in Round 1, 20 in Round 2) is relatively small compared to some other standard developments (e.g., 73 in the development of STARD[85]), but it is comparable to others (e.g., 24 for development of the TRIPOD statement[52]). Even though the response rate in Round 1 (21%) can be considered low, the number of participants was sufficient to cover a broad range of academic backgrounds, methodological experiences, and demographics (Figure 1). The same limitation is applicable to the size of the steering group (n = 7) as well as the test-user group (n = 2). The LEADING guideline should be regarded as an evolving reporting guideline requiring ongoing evaluation, refinement, and revision. Suggestions and recommendations for improvements are welcomed by emailing the corresponding authors.

Conclusions

The LEADING guideline emphasizes the transparent reporting of the methodological components of the assessment method and the importance of reporting *what* was assessed and *how* well. Considering the increasing need for high-accuracy assessments in diverse fields, we hope that the LEADING guideline will

be useful in assisting researchers in planning, carrying out, reporting, and evaluating research that aims to achieve accurate assessments.

Declarations

Ethics approval

Swedish law (2003:460) and the Swedish Ethical Review Authority state that only research that includes i) collecting personal information, and/or ii) that involves "obvious" (uppenbar) risk for physical or psychological harm, or iii) involves manipulating or deceiving individuals, should undergo an external ethical review. Considering that the current research involves asking participants to rate and comment on the reporting recommendations, this research does not fall within these criteria. The participants provided their informed consent and their individual open-ended comments and closed-ended ratings in the open material are anonymized. The closed-ended ratings per reporting standard in the Supplementary Material are presented on group level. The study is deemed exempt from requiring ethical approval according to Swedish Law (see §3-4 of the Act [2003:460] on ethical review of research involving humans in Sweden). Hence the research should not be reviewed by the Swedish Ethical Review Authority. More information can be found at the Swedish Ethical Review Authority (https://etikprovningsmyndigheten.se/).

Availability of data and materials

Open data (Delphi rounds 1 and 2), code (analyses), and material (surveys) can be found on the Open Science Framework: https://osf.io/fkv4b/

Competing interests

All authors have completed the ICMJE uniform disclosure form and declare: O. Kjell and K. Kjell have co-founded and hold shares in a start-up using computational language assessments to diagnose mental health problems based on best-estimate assessments and J.R. Boehnke is as editor part of the International Society for Quality of Life Research.

Funding

V.C. Eijsbroek, K. Kjell, and O. Kjell received funding from FORTE (2022-01022) and H.A. Schwartz from the National Institutes of Health (Grant R01 AA028032-01).

Authors' contributions

The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Contributor roles: E = Equal; S = Supporting; L = Lead
Working group: V.C. Eijsbroek, K. Kjell, and O. Kjell.
Steering group: H.A. Schwartz, J.R. Boehnke, E.I. Fried, D.N. Klein, P. Gustafsson, I. Augenstein, and P.M.M. Bossuyt.

Conceptualization	Working group
Data curation	V.E.
Formal analysis	V.E., O.K. (S)

Funding acquisition	O.K. (E) & K.K. (E)
Investigation	Working group
Methodology	Working group (L), Steering group
Project administration	Working group
Resources	Working group (L), Steering group
Software	-
Supervision	O.K.
Validation	-
Visualization	V.E.
Writing – original draft	Working group
Writing – review & editing	Working group (L), Steering group
Other	-

Acknowledgments

Patient organization members

E. Sellberg (Chairman of Libra Balans Skåne, board member of Balans National) and
I. Odenbrand (Vice Chairman of Libra Balans Skåne)

Participants in Delphi Round 1

I. Augenstein (Professor, Computer Science, University of Copenhagen),
S. Aydin (PhD, Developmental and Educational Psychology, Leiden University),
E. Billstedt (Professor, Neuroscience and Physiology, University of Gothenburg),
J.R. Boehnke (PhD, School of Health Sciences, University of Dundee),
D.W. Black (MD, Professor, Medicine, University of Iowa Carver College of Medicine),
B. Cannell (Associate Professor, Public Health, University of Texas Health Science Center at Houston),
G.A. Carlson (Professor, Psychiatry, Stony Brook University),
K.A.S. Davis (Researcher, Psychiatry Psychology and Neuroscience, King's College London),
F. Dereboy (MD, Psychiatry, Aydın Adnan Menderes University),
E.I. Fried (Associate Professor, Clinical Psychology, Leiden University),
P. Gustafsson (Associate Professor, Child and Adolescent Psychiatry, Lund University),
R. Handels (Assistant Professor, Psychiatry and Neuropsychology, Maastricht University),
K. Jenniskens (Assistant Professor, Clinical Epidemiology, Utrecht University),
C. Klaiman (Associate Professor, Pediatrics, Emory University),
D.N. Klein (Professor, Clinical Psychology, Stony Brook University),
M. McCloskey (Professor, Cognitive Science and Psychology, Johns Hopkins University),
A.C. Miers (Associate Professor, Developmental and Educational Psychology, Leiden University),
K.G.M. Moons (Professor, Clinical Epidemiology, Utrecht University),
L. Mosqueda (Professor, Medicine, University of Southern California),
H.A. Schwartz (Associate Professor, Computer Science, Stony Brook University),
M. Stein (Professor, Psychiatry and Behavioral Sciences, University of Washington),
J.G. Tillman (PhD, Clinical Psychology, Yale School of Medicine),
Y.P. Wang (MD, PhD, Medicine, University of Sao Paulo Medical School), and
J. Yonashiro-Cho (PhD, Medicine, University of Southern California).

Participants in Delphi Round 2
I. Augenstein (Professor, Computer Science, University of Copenhagen),
S. Aydin (PhD, Developmental and Educational Psychology, Leiden University),
J.R. Boehnke (PhD, School of Health Sciences, University of Dundee),
G.A. Carlson (Professor, Psychiatry, Stony Brook University),
K.A.S. Davis (Researcher, Psychiatry Psychology and Neuroscience, King's College London),
E.I. Fried (Associate Professor, Clinical Psychology, Leiden University),
P. Gustafsson (Associate Professor, Child and adolescent psychiatry, Lund University),
R. Handels (Assistant Professor, Psychiatry and Neuropsychology, Maastricht University),
K. Jenniskens (Assistant Professor, Clinical Epidemiology, Utrecht University),
C. Klaiman (Associate Professor, Pediatrics, Emory University),
D.N. Klein (Professor, Clinical Psychology, Stony Brook University),
M. McCloskey (Professor, Cognitive Science and Psychology, Johns Hopkins University),
A.C. Miers (Associate Professor, Developmental and Educational Psychology, Leiden University),
K.G.M. Moons (Professor, Clinical Epidemiology, Utrecht University),
L. Mosqueda (Professor, Medicine, University of Southern California),
Y.P. Wang (MD, PhD, Medicine, University of Sao Paulo Medical School), and
J. Yonashiro-Cho (PhD, Medicine, University of Southern California).

Test-users
T. Ivarsson (Associate Professor, Neuroscience and Physiology, University of Gothenburg) and
P.J. Snelling (PhD, Medicine, Gold Coast University Hospital and Griffith University)

Patient and public involvement
Prior to the Delphi procedure and test procedures, the reporting standards with inclusion rationales were discussed in an online workshop with a patient organization for depression (the chairman and vice chairman from Libra Balans Skåne), followed by receiving feedback from the steering group members to receive a wide range of perspectives early in the process.

References

1. Hirschtritt ME, Insel TR. Digital Technologies in Psychiatry: Present and Future. Focus Am Psychiatr Publ. 2018 Jul;16(3):251–8.
2. Venkatasubramanian G, Keshavan MS. Biomarkers in Psychiatry - A Critique. Ann Neurosci. 2016 Mar;23(1):3–5.
3. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955 Jul;52(4):281–302.
4. Scott O Lilienfeld, Katheryn eSauvigne, Steven Jay Lynn, Robert D Latzman, Robin eCautin, Irwin D. Waldman. Fifty Psychological and Psychiatric Terms to Avoid: A List of Inaccurate, Misleading, Misused, Ambiguous, and Logically Confused Words and Phrases. Front Psychol. 2015 Aug 1;6.
5. Spitzer RL. Psychiatric diagnosis: Are clinicians still necessary? Compr Psychiatry. 1983 Sep;24(5):399–411.
6. Leckman JF, Sholomskas D, Thompson WD, Belanger A, Weissman MM. Best estimate of lifetime psychiatric diagnosis: a methodological study. Arch Gen Psychiatry. 1982 Aug;39(8):879–83.
7. Bertens LCM, Broekhuizen BDL, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of Expert Panels to Define the Reference Standard in Diagnostic Research: A Systematic Review of Published Methods and Reporting. PLoS Med. 2013 Oct 15;10(10):e1001531.
8. Hunsley J, Mash EJ. Evidence-based assessment. Annu Rev Clin Psychol. 2007;3:29–51.
9. Black DW, Coryell WH, Crowe RR, McCormick B, Shaw MC, Allen J. A Direct, Controlled, Blind Family Study of DSM-IV Pathological Gambling. J Clin Psychiatry. 2014 Mar;75(3):215–21.
10. Reas D l., Rø O, Karterud S, Hummelen B, Pedersen G. Eating disorders in a large clinical sample of men and women with personality disorders. Int J Eat Disord. 2013 Dec 1;46(8):801–9.
11. Pedersen G, Karterud S, Wilberg T, Hummelen B. The impact of extended longitudinal observation on the assessment of personality disorders. Personal Ment Health. 2013 Nov 1;7(4):277–87.
12. Yonashiro-Cho J m. f., Gassoumis Z d., Homeier D c., Wilber K h. Improving forensics: Characterizing injuries among community-dwelling physically abused older adults. J Am Geriatr Soc. 2021 Aug 1;69(8):2252–61.
13. Blackmore R, Gray KM, Melvin GA, Newman L, Boyle JA, Gibson-Helm M. Identifying post-traumatic stress disorder in women of refugee background at a public antenatal clinic. Arch Womens Ment Health. 2022 Feb;25(1):191–8.
14. Morrison EH, Sorkin D, Mosqueda L, Ayutyanont N. Validity and Reliability of the Scale to Report Emotional Stress Signs-Multiple Sclerosis (STRESS-MS) in Assessing Abuse and Neglect of Adults With Multiple Sclerosis. Int J MS Care. 2022;(1).
15. Davis MAC, Spriggs A, Rodgers A, Campbell J. The Effects of a Peer-Delivered Social Skills Intervention for Adults with Comorbid Down Syndrome and Autism Spectrum Disorder. J Autism Dev Disord. 2018 Jun;48(6):1869–85.
16. Feder KM, Rahr HB, Lautrup MD, Egebæk HK, Christensen R, Ingwersen KG. Effectiveness of an expert assessment and individualised treatment compared with a minimal home-based exercise program in women with late-term shoulder impairments after primary breast cancer surgery: study protocol for a randomised controlled trial. Trials. 2022 Aug 20;23(1):701.
17. Niculescu AB, Le-Niculescu H. Precision medicine in psychiatry: biomarkers to the forefront. Neuropsychopharmacol Intersect Brain Behav Ther. 2022 Jan 1;47(1):422–3.
18. IMAGEN Consortium, Quinlan EB, Banaschewski T, Barker GJ, Bokde ALW, Bromberg U, et al. Identifying biological markers for improved precision medicine in psychiatry. Mol Psychiatry. 2020 Feb;25(2):243–53.
19. Maria Salud García-Gutiérrez, Francisco Navarrete, Francisco Sala, Ani Gasparyan, Amaya Austrich-Olivares, Jorge Manzanares. Biomarkers in Psychiatry: Concept, Definition, Types and Relevance to the Clinical Reality. Front Psychiatry. 2020 May 1;11.
20. Giovanni Briganti, Olivier Le Moine. Artificial Intelligence in Medicine: Today and Tomorrow. Front Med. 2020 Feb 1;7.
21. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31–

8.
22. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. Annu Rev Public Health. 2020;41(1):21–36.
23. Handels RLH, Wolfs CAG, Aalten P, Bossuyt PMM, Joore MA, Leentjens AFG, et al. Optimizing the use of expert panel reference diagnoses in diagnostic studies of multidimensional syndromes. BMC Neurol. 2014 Oct 4;14:190.
24. Reitsma JB, Rutjes AWS, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009 Jan 1;62(8):797–806.
25. Klein DN, Ouimette PC, Kelly HS, Ferro T, Riso LP. Test-retest reliability of team consensus best-estimate diagnoses of axis I and II disorders in a family study. Am J Psychiatry. 1994 Jul;151(7):1043–7.
26. Hogberg C, Billstedt E, Bjorck C, Bjorck P, Ehlers S, Gustle L, et al. Diagnostic validity of the MINI-KID disorder classifications in specialized child and adolescent psychiatric outpatient clinics in Sweden. BMC Psychiatry. 2019 Jan 1;19(1):142.
27. Anglin DM, Malaspina D. Ethnicity effects on clinical diagnoses compared to best-estimate research diagnoses in patients with psychosis: a retrospective medical chart review. J Clin Psychiatry. 2008 Jun;69(6):941–5.
28. Bech P, Timmerby N, Martiny K, Lunde M, Soendergaard S. Psychometric evaluation of the Major Depression Inventory (MDI) as depression severity scale using the LEAD (Longitudinal Expert Assessment of All Data) as index of validity. BMC Psychiatry. 2015 Aug 5;15:190.
29. Ruitian Gao, Shuai Zhao, Kedeerya Aishanjiang, Hao Cai, Ting Wei, Yichi Zhang, et al. Deep learning for differential diagnosis of malignant hepatic tumors based on multi-phase contrast-enhanced CT and clinical data. J Hematol OncolJ Hematol Oncol. 2021 Sep 1;14(1):1–7.
30. Bösner S, Haasenritter J, Becker A, Heinzel-Gutenbrunner M, Hani M a., Keller H, et al. Ruling out coronary artery disease in primary care: Development and validation of a simple prediction rule. CMAJ Can Med Assoc J. 2010 Sep 7;182(12):1295–300.
31. Cowan KJ, Tandias A, Arndt B, Hanrahan L, Mundt M, Guilbert TW. Defining Asthma: Validating Automated Electronic Health Record Algorithm With Expert Panel Diagnosis. Am J Respir Crit CARE Med. 2014 Jan 1;189.
32. Hall WB, Truitt SG, Scheunemann LP, Shah SA, Rivera MP, Parker LA, et al. The prevalence of clinically relevant incidental findings on chest computed tomographic angiograms ordered to diagnose pulmonary embolism. Arch Intern Med. 2009 Nov 23;169(21):1961–5.
33. Brian J, Roberts W, Szatmari P, Bryson S e., Smith I m., Roncadin C, et al. Stability and change in autism spectrum disorder diagnosis from age 3 to middle childhood in a high-risk sibling cohort. Autism. 2016 Oct 1;20(7):888–92.
34. Duffy A, Grof P, Goodday S, Keown-Stoneman C. The emergent course of bipolar disorder: Observations over two decades from the Canadian high-risk offspring cohort. Am J Psychiatry. 2019 Jan 1;176(9):720–9.
35. Elias R, Lord C. Diagnostic stability in individuals with autism spectrum disorder: insights from a longitudinal follow-up study. J Child Psychol Psychiatry. 2022 Sep 1;63(9):973–83.
36. Oudejans I, Mosterd A, Bloemen JA, Valk MJ, van Velzen E, Wielders JP, et al. Clinical evaluation of geriatric outpatients with suspected heart failure: value of symptoms, signs, and additional tests. Eur J Heart Fail. 2011;13(5):518–27.
37. Mooney M a., Wilmot B, Bhatt P, Nigg J t., Hermosillo R j. m., Fair D a., et al. Smaller total brain volume but not subcortical structure volume related to common genetic risk for ADHD. Psychol Med. 2021 Jun 1;51(8):1279–88.
38. Lamers F, Cui L, Hickie IB, Roca C, Machado-Vieira R, Zarate JrCA, et al. Familial aggregation and heritability of the melancholic and atypical subtypes of depression. J Affect Disord. 2016 Nov 1;204:241–6.
39. Merikangas KR, Cui L, Heaton L, Nakamura E, Roca C, Ding J, et al. Independence of familial

transmission of mania and depression: results of the NIMH family study of affective spectrum disorders. Mol Psychiatry. 2014 Feb 1;19(2):214–9.
40. Mataix-Cols D, Billotti D, Fernández De La Cruz L, Nordsletten AE. The London field trial for hoarding disorder. Psychol Med. 2013 Apr;43(4):837–47.
41. Dereboy F, Dereboy Ç, Eskin M. Validation of the DSM–5 alternative model personality disorder diagnoses in Turkey, Part 1: LEAD validity and reliability of the personality functioning ratings. J Pers Assess. 2018 Nov;100(6):603–11.
42. Sung M, Goh TJ, Tan BLJ, Chan JS, Liew HSA. Comparison of DSM-IV-TR and DSM-5 Criteria in Diagnosing Autism Spectrum Disorders in Singapore. J Autism Dev Disord. 2018 Oct;48(10):3273–81.
43. Takeshi Nishiyama, Satoshi Sumi, Hiroto Watanabe, Futoshi Suzuki, Yukiko Kuru, Tomoko Shiino, et al. The Kiddie Schedule for Affective Disorders and Schizophrenia Present and Lifetime Version (K-SADS-PL) for DSM-5: A validation for neurodevelopmental disorders in Japanese outpatients. Compr Psychiatry. 2020 Jan 1;96.
44. Gerdner A, Kestenberg J, Mattias E. Validity of the Swedish SCID and ADDIS diagnostic interviews for substance use disorders: Sensitivity and specificity compared with a LEAD golden standard. Nord J Psychiatry. 2015 Jan 1;69(1):48–56.
45. North CS, Simic Z, Burruss J. Design, Implementation, and Assessment of a Public Comprehensive Specialty Care Program for Early Psychosis. J Psychiatr Pract. 2019 Mar;25(2):91–102.
46. Osório FL, Loureiro SR, Hallak JEC, Machado-de-Sousa JP, Ushirohira JM, Baes CVW, et al. Clinical validity and intrarater and test-retest reliability of the Structured Clinical Interview for DSM-5 - Clinician Version (SCID-5-CV). Psychiatry Clin Neurosci. 2019 Dec;73(12):754–60.
47. Flake J k., Fried E i. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. Adv Methods Pract Psychol Sci. 2020 Dec 1;3(4):456–65.
48. Aguinis H, Ramani RS, Alabduljader N. What You See Is What You Get? Enhancing Methodological Transparency in Management Research. Acad Manag Ann. 2018 Jan;12(1):83–110.
49. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. Int J Surg. 2014 Dec 1;12(12):1495–9.
50. STARD Group, Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. STARD 2015 : an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015 Oct 28;351.
51. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010 Mar 23;340:c332.
52. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015 Jan 7;350:g7594.
53. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. 2010 Feb 16;7(2):e1000217.
54. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. BMC Med. 2022 Jan 12;20(1):23.
55. Hsu CC, Sandford B a. The Delphi technique: Making sense of consensus. Pract Assess Res Eval. 2007 Jan 1;12(10):1–8.
56. Robin W M Vernooij, Pablo Alonso-Coello, Melissa Brouwers, Laura Martínez García, CheckUp Panel. Reporting Items for Updated Clinical Guidelines: Checklist for the Reporting of Updated Guidelines (CheckUp). PLoS Med. 2017 Jan 1;14(1):e1002207–e1002207.
57. Mackenhauer J, Winsløv JH, Holmskov J, Brødsgaard I, Larsen TG, Mainz J. Analysis of Suicides Reported as Adverse Events in Psychiatry Resulted in Nine Quality Improvement Initiatives. Crisis. 2022 Jul;43(4):307–14.
58. Hendriks E, Muris P, Meesters C, Houben K. Childhood Disorder: Dysregulated Self-Conscious

Emotions? Psychopathological Correlates of Implicit and Explicit Shame and Guilt in Clinical and Non-clinical Children and Adolescents. Front Psychol. 2022;13:822725.
59. Paap MCS, Heltne A, Pedersen G, Germans Selvik S, Frans N, Wilberg T, et al. More is more: Evidence for the incremental value of the SCID-II/SCID-5-PD specific factors over and above a general personality disorder factor. Personal Disord. 2022 Mar;13(2):108–18.
60. Aydin S, Siebelink BM, Crone MR, van Ginkel JR, Numans ME, Vermeiren RRJM, et al. The diagnostic process from primary care to child and adolescent mental healthcare services: the incremental value of information conveyed through referral letters, screening questionnaires and structured multi-informant assessment. BJPsych Open. 2022 Apr 7;8(3):e81.
61. Rocha Neto HG, Lessa JLM, Koiller LM, Pereira AM, de Souza Gomes BM, Veloso Filho CL, et al. Non-standard diagnostic assessment reliability in psychiatry: a study in a Brazilian outpatient setting using Kappa. Eur Arch Psychiatry Clin Neurosci [Internet]. 2023 Dec 12 [cited 2024 Aug 30]; Available from: https://doi.org/10.1007/s00406-023-01730-7
62. Spangenberg H, Ramklint M, Ramirez A. A long-term follow-up study of labor market marginalization in psychiatric patients with and without personality disorder. Ups J Med Sci. 2023 Jul 31;128:10.48101/ujms.v128.9014.
63. Muris P, Büttgens L, Koolen M, Manniën C, Scholtes N, van Dooren-Theunissen W. Symptoms of Selective Mutism in Middle Childhood: Psychopathological and Temperament Correlates in Non-clinical and Clinically Referred 6- to 12-year-old Children. Child Psychiatry Hum Dev. 2023 Feb 28;
64. Sveen CA, Pedersen G, Hummelen B, Kvarstein EH. Personality disorders: The impact of severity on societal costs. Eur Arch Psychiatry Clin Neurosci. 2023;No Pagination Specified-No Pagination Specified.
65. Pedersen G, Kvarstein EH, Wilberg T, Folmo EJ, Burlingame GM, Lorentzen S. The Group Questionnaire (GQ)—Psychometric properties among outpatients with personality disorders. Group Dyn Theory Res Pract. 2023;27(2):81–98.
66. Sadleir PHM, Clarke RC, Goddard CE, Mickle P, Platt PR. Agreement of a clinical scoring system with allergic anaphylaxis in suspected perioperative hypersensitivity reactions: prospective validation of a new tool. Br J Anaesth. 2022 Nov;129(5):670–8.
67. Khan AM, Ahmed S, Chowdhury NH, Islam MS, McCollum ED, King C, et al. Developing a video expert panel as a reference standard to evaluate respiratory rate counting in paediatric pneumonia diagnosis: protocol for a cross-sectional study. BMJ Open. 2022 Nov 15;12(11):e067389.
68. Loots FJ, Smits M, Hopstaken RM, Jenniskens K, Schroeten FH, van den Bruel A, et al. New clinical prediction model for early recognition of sepsis in adult primary care patients: a prospective diagnostic cohort study of development and external validation. Br J Gen Pract J R Coll Gen Pract. 2022 Jun;72(719):e437–45.
69. Leroux A, Frey KP, Crainiceanu CM, Obremskey WT, Stinner DJ, Bosse MJ, et al. Defining Incidence of Acute Compartment Syndrome in the Research Setting: A Proposed Method From the PACS Study. J Orthop Trauma. 2022 Jan 1;36(Suppl 1):S26–32.
70. Kocks JWH, Cao H, Holzhauer B, Kaplan A, FitzGerald JM, Kostikas K, et al. Diagnostic Performance of a Machine Learning Algorithm (Asthma/Chronic Obstructive Pulmonary Disease [COPD] Differentiation Classification) Tool Versus Primary Care Physicians and Pulmonologists in Asthma, COPD, and Asthma/COPD Overlap. J Allergy Clin Immunol Pract. 2023 May 1;11(5):1463-1474.e3.
71. Lacroix L, Papis S, Mardegan C, Luterbacher F, L'Huillier A, Sahyoun C, et al. Host biomarkers and combinatorial scores for the detection of serious and invasive bacterial infection in pediatric patients with fever without source. PloS One. 2023;18(11):e0294032.
72. Nienhuis PH, van Nieuwland M, van Praagh GD, Markusiewicz K, Colin EM, van der Geest KSM, et al. Comparing Diagnostic Performance of Short and Long [18F]FDG-PET Acquisition Times in Giant Cell Arteritis. Diagn Basel Switz. 2023 Dec 27;14(1):62.
73. Himmelreich JCL, Harskamp RE. Diagnostic accuracy of the PMcardio smartphone application for artificial intelligence–based interpretation of electrocardiograms in primary care (AMSTELHEART-

1). Cardiovasc Digit Health J. 2023 Jun 1;4(3):80–90.
74. Matta M de la, Alonso-González M, García-Santigosa M, Arance-García M, Sánchez-Peña J, Castro-Liñán LM, et al. Accuracy and Comprehensiveness in Recording Information of a Web-Based Application for Preoperative Assessment: A Prospective Observational Study. J Perianesth Nurs. 2023 Jun 1;38(3):440–7.
75. Peterson BS, Kaur T, Baez MA, Whiteman RC, Sawardekar S, Sanchez-Peña J, et al. Morphological Biomarkers in the Amygdala and Hippocampus of Children and Adults at High Familial Risk for Depression. Diagnostics. 2022 May;12(5):1218.
76. Reiersen AM, Noel JS, Doty T, Sinkre RA, Narayanan A, Hershey T. Psychiatric Diagnoses and Medications in Wolfram Syndrome. Scand J Child Adolesc Psychiatry Psychol. 2022 Jan;10(1):163–74.
77. Bradshaw J, Shi D, Hendrix CL, Saulnier C, Klaiman C. Neonatal neurobehavior in infants with autism spectrum disorder. Dev Med Child Neurol. 2022 May;64(5):600–7.
78. Hesam-Shariati S, Overs BJ, Roberts G, Toma C, Watkeys OJ, Green MJ, et al. Epigenetic signatures relating to disease-associated genotypic burden in familial risk of bipolar disorder. Transl Psychiatry. 2022 Aug 3;12(1):310.
79. Shima C, Lee R, Coccaro EF. Associations of agression and use of caffeine, alcohol and nicotine in healthy and aggressive individuals. J Psychiatr Res. 2022 Feb;146:21–7.
80. Kvig EI, Nilssen S. Does method matter? Assessing the validity and clinical utility of structured diagnostic interviews among a clinical sample of first-admitted patients with psychosis: A replication study. Front Psychiatry. 2023;14:1076299.
81. Detera-Wadleigh SD, Kassem L, Besancon E, Lopes F, Akula N, Sung H, et al. A resource of induced pluripotent stem cell (iPSC) lines including clinical, genomic, and cellular data from genetically isolated families with mood and psychotic disorders. Transl Psychiatry. 2023 Dec 16;13(1):397.
82. Hill SY, Hostyk J. A whole exome sequencing study to identify rare variants in multiplex families with alcohol use disorder. Front Psychiatry [Internet]. 2023 Oct 17 [cited 2024 Aug 15];14. Available from: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1216493/full
83. Jones W, Klaiman C, Richardson S, Lambha M, Reid M, Hamner T, et al. Development and Replication of Objective Measurements of Social Visual Engagement to Aid in Early Diagnosis and Assessment of Autism. JAMA Netw Open. 2023 Sep 5;6(9):e2330145.
84. Dimian AF, Estes AM, Dager S, Piven J, Wolff JJ, Network for the I. Predicting self-injurious behavior at age three among infant siblings of children with autism. Autism Res. 2023;16(9):1670–80.
85. Korevaar DA, Cohen JF, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. Updating standards for reporting diagnostic accuracy: the development of STARD 2015. Res Integr Peer Rev. 2016 Jun 7;1(1):7.