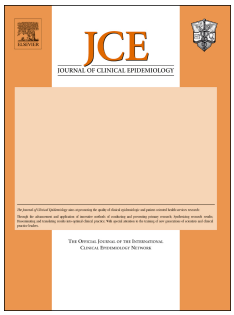


Methods for developing diagnostic criteria for conditions without objective tests, biomarkers or reference standards: a scoping review

Samuel J. White, Timothy H. Barker, Tracy Merlin, Grace Holland, Sharon Sanders, Aoife O'Mahony, Thanya Pathirana, Rebecca Theiss, Danielle Pollock, Natasha Reid, Zachary Munn

PII: S0895-4356(25)00385-3

DOI: https://doi.org/10.1016/j.jclinepi.2025.112052

Reference: JCE 112052

To appear in: *Journal of Clinical Epidemiology*

Received Date: 14 June 2025

Revised Date: 4 November 2025

Accepted Date: 10 November 2025

Please cite this article as: White SJ, Barker TH, Merlin T, Holland G, Sanders S, O'Mahony A, Pathirana T, Theiss R, Pollock D, Reid N, Munn Z, Methods for developing diagnostic criteria for conditions without objective tests, biomarkers or reference standards: a scoping review, *Journal of Clinical Epidemiology* (2025), doi: https://doi.org/10.1016/j.jclinepi.2025.112052.

TITLE

Methods for developing diagnostic criteria for conditions without objective tests, biomarkers or reference standards: a scoping review.

ARTICLE TYPE

Original research

AUTHORS

Samuel J White[1*]
Timothy H Barker[1]
Tracy Merlin[2]
Grace Holland[1]
Sharon Sanders[3]
Aoife O'Mahony[4]
Thanya Pathirana[5]
Rebecca Theiss[6]
Danielle Pollock[1]
Natasha Reid[7]
Zachary Munn[1]

AFFILIATIONS

1. Health Evidence Synthesis, Recommendations & Impact (HESRI), School of Public Health, The University of Adelaide, Adelaide, South Australia, Australia
2. Adelaide Health Technology Assessment (AHTA), School of Public Health, The University of Adelaide, Adelaide, South Australia, Australia
3. Institute for Evidence-Based Healthcare, Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Australia
4. School of Public Health, University College Cork, Cork, Ireland.
5. School of Medicine and Dentistry, Griffith University, Sunshine Coast, Queensland, Australia
6. Bloom Healthcare, Adelaide 5000, South Australia, Australia.

7. Child Health Research Centre, Faculty of Medicine, The University of Queensland, Brisbane, Queensland, Australia.

* Corresponding author

TITLE

Methods for developing diagnostic criteria for conditions without objective tests, biomarkers or reference standards: a scoping review.

ABSTRACT

Objective: Diagnostic criteria play an important role in informing clinical decision-making, particularly for conditions lacking objective tests, biomarkers or reference standards. Despite their importance, there is no established methodological guidance for developing diagnostic criteria. This scoping review aimed to identify and describe the methodological approaches used to develop diagnostic criteria in the absence of objective tests, biomarkers or reference standards.

Study design and setting: We conducted a scoping review in accordance with JBI methodology and the PRISMA-ScR reporting guideline. Studies published between 2000 and 2024 that described methods used to develop diagnostic criteria for conditions without objective tests, biomarkers or reference standards were included. A comprehensive search was performed across multiple databases and supplemented with grey literature searches and expert consultation. Data were extracted independently by two reviewers and synthesised using descriptive statistics and qualitative content analysis.

Results: We included 139 studies. Suboptimal reporting of methodology was a barrier to assessment of methodological credibility. Authors used one or more of three main approaches to develop diagnostic criteria: consensus-based, literature-based and/or primary study-based. Consensus methods were used in 98/139 (71%) of studies, with Delphi or modified Delphi approaches being the most commonly adopted. The role of evidence in diagnostic criteria development was not described in 36/139 (26%) of the included studies. In studies using consensus methodology to develop diagnostic criteria, prospective approaches to ensuring appropriate diversity amongst the diagnostic criteria development panel were employed in

only 5/98 (5%) of studies and patient/advocate consultation was performed in 18/98 (18%) of studies.

Conclusion: Methodological approaches to developing diagnostic criteria for conditions without objective tests or standards are variable, inconsistently reported and often lack a clear evidence base. This could be aided by development of specific methodological guidance.

Plain Language Summary

When lab tests or scans aren't available to confirm a diagnosis, doctors may use diagnostic criteria to help them decide what condition a patient may have. There is currently no clear way to create these criteria, which can lead to inconsistency and confusion.

We looked at why this matters – because diagnostic criteria developed without transparency or methodological rigour may lead to incorrect diagnosis, patient harm or inequitable access to care - and explored how researchers develop diagnostic criteria when an objective test doesn't exist. Most studies relied on expert meetings, but many did not explain how they chose experts, gathered evidence or involved diverse perspectives such as patient voices.

Our research shows that current practices often aren't well explained and may be biased or incomplete. By understanding what's missing and what may work, we hope this study encourages better, more transparent methods so that future diagnostic criteria are clearer, fairer and more trustworthy for both patients and healthcare providers.

Keywords

diagnostic criteria; consensus development; evidence-based practice; scoping review; guideline development; reference standards

Running Title

Methods for diagnostic criteria development

WHAT IS NEW?

Key findings

- Methodological approaches to developing diagnostic criteria for conditions without objective tests, biomarkers or reference standards were variable and inconsistently reported, acknowledging not all included studies were explicitly designed to develop diagnostic criteria
- Whilst consensus methods were employed by 71% of included studies to develop diagnostic criteria, few studies adopted strategies to ensure diverse or representative panel composition
- Within the limitations of the suboptimal reporting encountered, evidence synthesis was variably integrated into the diagnostic criteria development process and only 4% of groups used formal tools to assess the strength of evidence

What this adds to what is known?

- This scoping review introduces a conceptual map of current methodological approaches to developing diagnostic criteria, including consensus-based, literature-based and primary study-based approaches
- We identify several critical gaps relating to the integration of evidence synthesis and appraisal into diagnostic criteria development as well as methodological reporting practices
- Adequate diversity amongst panels developing diagnostic criteria using consensus methodology and patient/advocate consultation are infrequently considered, presenting barriers to health equity

What is the implication/what should change now?

- Future diagnostic criteria development initiatives should explicitly describe how evidence was used to inform each stage of development, for example by integrating systematic reviews, undertaking structured evidence appraisal and using evidence-to-decision frameworks. Specific methodological guidance is needed to ascertain the types of evidence that should be used
- Additional priorities should include proactive promotion of panel diversity when using consensus methods and inclusion of patients and advocates to ensure criteria are comprehensive and clinically relevant

INTRODUCTION

Accurate and timely diagnosis is central to informing good clinical care. Diagnosis is a complicated, dynamic and often longitudinal categorisation task. There are many potential precipitants of delayed, missed or inaccurate diagnosis including insufficient acquisition of information, inaccurate synthesis of available information and inappropriate application of existing knowledge(1, 2). To help navigate these potential pitfalls, diagnostic criteria are developed to summarise current evidence and provide (where possible) an evidence-based framework that can be used to systematically, accurately and efficiently identify individuals affected by specific diseases or conditions, enabling timely provision of appropriate care to enhance patient-centred outcomes (3). Diagnostic criteria comprise a collection of symptoms, clinical signs and readily accessible/available test results(4) that support or refute the diagnosis of a specific condition in a clinically heterogenous population. The aim being the accurate identification of *all* affected individuals including phenotypic variants and attenuated manifestations or 'formes frustes'(5), while maintaining a balance between sensitivity and the potential harms of diagnosing individuals unlikely to benefit from receiving the diagnosis or subsequent treatment based on the diagnosis(6-8). This contrasts with classification or research criteria, which aim to provide a very well-defined phenotype representative of the majority (but not all) individuals with a condition for the purposes of defining a condition, particularly in research contexts where maintaining homogeneity amongst patient cohorts is important(9). There are several different contexts in which diagnostic criteria may be established. Amongst other scenarios, diagnostic criteria may be developed as independent initiatives commissioned by specialty colleges or professional societies, during clinical guideline development or in response to emerging diseases or conditions that have not previously been defined. Despite

their considerable clinical significance, there is currently no formal methodological guidance available to support the development of evidence-based diagnostic criteria.

Objective tests, biomarkers or reference standard tests can be used to guide clinical diagnosis of many conditions. Objective tests are diagnostic assessments that yield measurable and reproducible results independent of clinician judgement. For example, polymerase chain reaction (PCR) tests analysing respiratory secretions can accurately detect influenza genetic material (ribonucleic acid (RNA)), enabling reproducible confirmation of the presence or absence of infection. Biomarkers are quantifiable characteristics that can be used to indicate the presence of disease, monitor the severity of disease or predict treatment response(10). For example, detection of elevated serum troponin over serial measurement is often used to guide diagnosis of myocardial infarction in combination with clinical assessment and other tests such as electrocardiogram (ECG). Reference standards tests represent the current best and accepted approach to establishing a diagnosis(11). For example, whilst several tests such as sonographic nuchal translucency are available to assess for the presence of trisomy 21 (Down syndrome), identification of a third chromosome 21 using chromosomal microarray is considered the reference standard test used to establish a diagnosis. However, many conditions lack objective tests, biomarkers or reference standards to guide diagnosis. This occurs for a range of reasons, including that some conditions are characterised by less clearly defined aetiology, or more heterogenous spectra of presentations, or are identified by less specific signs and symptoms. Examples include fibromyalgia syndrome(12), concussion(13), post-COVID condition(14), fetal alcohol spectrum disorder(15) and most psychiatric conditions, such as mood, anxiety, psychotic and personality disorders(16). These conditions can be particularly difficult to diagnose, which often necessitates the development of diagnostic criteria to guide diagnostic decision-making.

To support the accurate diagnosis of conditions for which objective tests, biomarkers or references standards do not exist, it is important that their diagnostic criteria are rigorously developed. However, the methodological approaches to develop diagnostic criteria for such conditions are often underreported, described with limited detail, and lack uniformity, resulting in poor generalisability to other conditions(17). Therefore, the aim of this scoping review was to describe the specific methodological approaches employed to develop diagnostic criteria for conditions without objective tests, biomarkers or reference standards.

METHODS

This scoping review was performed using the methodological approach described by Peters *et al(18)* and was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) extension for scoping reviews(19). The study protocol was prospectively registered on the Open Science Framework (https://osf.io/e5wtg/).

Eligibility Criteria

This review included all studies, guidelines and documents concerning human populations with a disease or condition for which diagnosis is being sought but for which there is no objective test, accepted reference standard or valid biomarker. Conditions for which validated blood tests, imaging tests, histopathological examinations or other objective tests are available were excluded.

Concept: the core concept of this review was the methodological processes undertaken in the development of diagnostic criteria. For inclusion, sources were required to describe a methodological approach or process of development of a set of diagnostic criteria. Sources, including guidelines, that provided only diagnostic criteria but without description of the methodological development process were excluded. Sources that described non-diagnostic criteria such as classification criteria, research diagnostic criteria, response criteria or referral criteria were also excluded. Single-author commentaries and narrative reviews/critiques without a systematic literature or consensus-based approach were also excluded.

Context: this review was not limited by context, setting, discipline or field. However, only sources published between 2000-2024 were eligible for inclusion as this period corresponds

to the uptake of evidence-based medicine in clinical practice(20, 21) and represents contemporary medicine as opposed to historical practices.

Types of sources: this review considered diagnostic/clinical practice guidelines, technical reports or protocols underpinning diagnostic criteria, and systematic reviews. Relevant grey literature such as codes of practice, manuals, handbooks or policies published on government, non-government organisation or healthcare organisation websites were also eligible for inclusion.

Search Strategy

A preliminary search of MEDLINE was performed to identify relevant sources. Following collaboration with an academic librarian, keywords within the titles and abstracts of these sources, along with relevant index terms, were used to formulate a full search strategy for MEDLINE (Supplement 1). The search strategy, including all identified keywords and index terms, was adapted for the following databases using Polyglot(22): Embase (Elsevier), Cochrane Database of Systematic Reviews, CINAHL with full text (EBSCO), PubMed and PsycInfo and performed in June 2024. A supplementary search was performed to complement the formal structured database searching. Multiple websites were searched for relevant sources using combinations of the following keywords in English: diagnostic criteria, guideline, disorder, condition, syndrome (Supplement 2). These websites were searched in incognito mode, and the first 100 results reviewed. If a result was deemed to be relevant, the next 100 results were reviewed until a judgement was made by the reviewer that the search results were no longer appropriate. Finally, the reference list of all studies that met the eligibility criteria and were included were screened for additional studies. There were no exclusions based on language or publication status (ie published, unpublished, in press, in

progress, pre-print). For documents in a language other than English, DeepL (www.DeepL.com) was used to translate the text.

Source Selection

All relevant records identified through the literature search were imported into Covidence(23) and duplicates were removed. Titles and abstracts of potentially relevant studies were screened independently by SW and GH/SS/AOM/TP/RT with disagreements resolved through consensus discussion. Full-text screening was also performed in duplicate by SW and GH/SS/AOM/TP/RT with disagreements again resolved through consensus discussion.

Data Extraction

Data extraction was performed by SW and RT independently using a data extraction template, supported through a shared vocabulary specified in a dictionary of terms developed *a priori* (Supplements 3-4). The data extracted included general characteristics such as title, year of publication and source type. Specific clinical and methodological information were also extracted including the disease or condition of interest, prior availability of diagnostic criteria for the condition of interest, the number, expertise/disciplinary background and roles of individuals involved in diagnostic criteria development, general stages/phases of criteria development, and sources of evidence considered during criteria development.

An initial pilot of data extraction was performed for the first five studies by SW and RT. No modifications to the data extraction templates were made. Subsequently, SW and RT independently completed data extraction, resolving disagreements through consensus discussion. Critical appraisal of individual sources was not performed, as it was not relevant to the objective of this scoping review.

Data Analysis and Synthesis of Results

Data relating to the general characteristics of included studies were evaluated using descriptive statistical analysis performed using RStudio(24). Evidence regarding approaches used to develop diagnostic criteria was collated and synthesised using basic qualitative content analysis before being presented in the form of a narrative summary. We employed an inductive approach to qualitative content analysis(25) whereby an initial set of methodological categories were developed based on the data extraction template, including formal consensus methods (e.g. Delphi, RAND/UCLA appropriateness method, NIH consensus development workshop), informal consensus methods (e.g. other workshops, unstructured meetings), systematic reviews, narrative reviews, randomised controlled trials, cohort studies, case control studies, case series/studies and validation studies. Each included article was coded into one or more of these categories, after which it became apparent that several categories overlapped in function. Through iterative refinement, these categories were consolidated into broader themes. For example, formal and informal consensus methods were merged into a single category of 'consensus-based approaches. During qualitative content analysis, we inductively identified additional methodological considerations (such as atypical case bias and practicality) that were explicitly reported in some studies. Atypical case bias was defined as the risk of developing diagnostic criteria based on non-representative or unusual patient cases. Practicality was defined as considerations of how well proposed diagnostic criteria could be implemented across different clinical or resource settings. The final number and content of methodological themes was agreed upon by all authors.

RESULTS

Search Results

The literature search identified 16,265 unique records. Screening of titles and abstracts yielded 1301 potentially relevant reports of studies. A total of 139 studies were included in our final analysis following full-text review. The PRISMA flow diagram summarising the multilevel study screening process is presented in Figure 1.

Characteristics of Sources

The 139 studies included 115 studies publishing new diagnostic criteria in the format of original research articles, 17 clinical practice guidelines, 4 narrative reviews, 1 randomised controlled trial, 1 case series and 1 study protocol (Table 1). The most common place of publication (defined as the country of the first author's first-listed affiliation) was North America (65, 47%) followed by Europe (56, 40%) and Asia (9, 6%). Many studies (107, 77%) involved collaboration between five or more institutions.

Diagnostic criteria were most commonly developed for neurological conditions (40, 29%) followed by psychological/psychiatric (36, 26%) and rheumatological (17, 12%) conditions. There were 41 studies (29%) that evaluated conditions associated with preceding insults or injuries.

Of the 139 included studies, 77/139 (55%) studies addressed a disease or condition for which diagnostic criteria had previously been published. Justification for developing new criteria was provided in 74/77 (96%) of these studies. Amongst this group of studies, 42/77 (55%) revised the existing diagnostic criteria whereas 35/77 (45%) developed a new set of

diagnostic criteria. Within the group of 42 studies revising previous diagnostic criteria, 38/42 (90%) involved one or more of the authors of the previously published diagnostic criteria.

Evidence obtained through literature review, primary research or a combination of both was presented or considered during diagnostic criteria development in 103/139 studies(74%). The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach was used to inform the strength of recommendations was performed in three studies(26-28). Additionally Höglinger and colleagues(29) evaluated the strength of evidence in accordance with the Scottish Intercollegiate Guidelines Network (SIGN) guidance, which itself endorses the GRADE approach(30). Similarly, Tinuper *et al(31)* followed American Academy of Neurology guidance(32), which endorses a modified version of the GRADE approach whereby confidence level is anchored to the designation of risk of bias of included studies with other factors only able to upgrade confidence by a maximum of one level. Conversely, Landgraf and colleagues(33) determined strength of evidence using the Oxford Centre for Evidence-based Medicine's classification for diagnostic studies(34), which involves assigning a level of evidence to each informative study and then providing an overall grade of recommendation based on this information. The majority of studies (83/139, 60%) did not consider any specific measures of diagnostic criteria performance. When specific measures of performance were considered, sensitivity (36/139, 26%), specificity (41/139, 29%) and area under receiver operating characteristic (ROC) curve (10/139, 73%) were most commonly used. One study detailed a planned diagnostic criteria updating process: Wenning *et al* specified the intended use of horizon scanning to inform future revisions to their diagnostic criteria(35). External validation of developed diagnostic criteria was planned or performed in 4 out of 139 (3%) of studies(26, 36-38).

Methodological Approaches to Developing Diagnostic Criteria

We identified three main categories of methodology used to develop diagnostic criteria: 1) consensus-based approaches, 2) literature-based approaches and 3) primary study-based approaches (Table 2; Figure 2; Supplement 5). Most included studies adopted a single approach: 37 studies (27%) adopted a consensus-based approach, 10 (7%) adopted a literature-based approach and 29 (21%) adopted a primary study-based approach. Two or three methodological approaches were used in 63 (45%) studies.

1. Consensus-based approaches

Consensus methodology was used to develop diagnostic criteria in 98/139 studies (71%). Among these, specific consensus approaches included Delphi or modified Delphi methods 32/98 (33%), nominal group exercise groups (2%) and NIH consensus development conferences (1%). The consensus processes employed in the remaining 63/98 studies (64%) were variable, including 1) a core author group or steering committee drafting criteria and then disseminating a document containing the draft criteria to a larger group of multidisciplinary experts who provided iterative feedback and amendments across multiple rounds remotely(39), 2) holding in-person(40-47) or online(48) meetings or workshops amongst members of a topic-specific special interest or working group or group of invited experts and 3) structuring discussions amongst invited experts around major topic- or field-specific conferences. Several studies (12%) indicated that some form of consensus methodology was adopted without reporting who was involved or what processes were performed.

Expert recruitment and consideration of diversity amongst the panel- A range of approaches to recruiting expert panellists were adopted. Several groups employed an informal approach

to recruitment. For example, Collins *et al* convened group discussions amongst senior authors to nominate experts for potential recruitment(49). Other groups prospectively developed specific definitions of 'expert' prior to recruitment. For instance, prior to developing diagnostic criteria for caesarean scar disorder, Klein Meuleman and colleagues defined an expert in the field as an obstetrician or gynaecologist who consults on 30 or more uterine niche-related cases per year and has participated in related research(50). Similarly, in their study examining gaming disorder diagnostic criteria, Castro-Calvo *et al* defined an expert as an individual with a minimum of 1 year of clinical experience and at least two peer-reviewed publications in the field(51). Muller *et al* identified potential experts by performing a literature search involving multiple electronic databases and using topic-specific keywords(52). A list of all first and last authors from the search results was compiled, and these individuals were invited via email to participate in the study.

Specific methodology was employed to improve diversity amongst the groups involved in developing diagnostic criteria in 5/98 (5%) of studies that adopted consensus methodology. Amongst these studies, 4/5 (80%) adopted strategies to improve racial, ethnic or geographic diversity(51, 53-55). The most common approach was to define geographic regions and then recruit individuals from each region to achieve truly global representation(53-55). In one study, one of the criteria used to define an 'expert' in the context of panellist recruitment was publication of 20 or more studies in the subject area. However, to improve racial/ethnic diversity amongst the panel, a less stringent criterion for involvement was applied for researchers based in underrepresented countries(51). In 1/5 (20%) studies, the editorial board developed pre-set criteria that included gender-based quotas to identify individuals for subcommittee chair and co-chair positions to improve gender diversity(53). One study devising criteria for 'irremediable psychiatric suffering' in the context of medical assistance

in dying sought to improve diversity of opinion by purposefully inviting psychiatrists who were both vocal proponents and opponents of medical assistance in dying (as well as moderates) to participate(56). Notably, none of the included studies defined panel diversity, reported whether or not First Nations people were included or discussed issues regarding the diversity of their panel.

Patient, client, consumer and advocate consultation

Eighteen studies (18%) that adopted consensus approaches to develop diagnostic criteria employed deliberate strategies to involve patients, clients, consumers or advocates. The timing and level of engagement of interest holders within the diagnostic criteria development process was variable. A total of 6/18 (33%) studies performed patient consultation prior to the consensus process(15, 50, 57-60) whereas 13/18 (72%) included patients in the consensus process itself(27, 33, 57-67), and 3/18 (17%) sought patient feedback on outcomes derived from the completed consensus process(50, 68, 69). Furthermore, whilst some authors elected to interview patient groups in isolation, other groups integrated patients into multi-interest holder discussions and forums involving clinicians, researchers and other invited individuals. For example, the questions presented to experts during the Delphi process in Klein Meuleman and colleagues' study were derived from an initial focus group discussion with patients living with symptomatic disease(50). Following completion of the Delphi process, a group of patient representatives were consulted on the proposed diagnostic criteria prior to their finalisation and publication. In contrast, Mease *et al* integrated patient focus groups along with clinicians and researchers in a nominal group exercise facilitated by qualitative research experts to generate a vocabulary and set of concepts that would form the basis for the subsequent Delphi process designed to derive diagnostic criteria(58).

Question development- The statements and questions that formed the basis for discussion during the consensus processes were developed through a range of different approaches. A common approach involved a core group of authors or steering group defining study objectives and performing a preliminary literature review (using either a pre-defined search strategy or informal methods) to identify important or relevant publications relating to the topic and then drafting a set of questions, statements or diagnostic criteria. These drafts were then voted on by a larger group of invited experts in a subsequent consensus process(39, 63). For example, Healy and colleagues aimed to develop diagnostic criteria for persistent sexual dysfunction following treatment with a range of medications including antidepressants, finasteride and isotretinoin(39). The authors reviewed data from the two largest case series published on the topic and identified common features to draft diagnostic criteria (specific prevalence thresholds to define 'common' were not provided). This draft was iteratively reviewed and revised by a multidisciplinary group of experts until a finalised set of criteria were agreed upon.

Cummings and colleagues(62) aimed to revise diagnostic criteria for psychosis in neurocognitive disorders. The authors performed an initial survey of specialist clinicians to identify whether participants consistently indicated necessary changes to the pre-existing Jeste and Finkel diagnostic criteria(70) to enhance clinical utility. Concepts derived from this survey were then used to prompt discussion at a subsequent in-person expert consensus meeting.

Mease *et al* published a protocol detailing their planned multi-stage approach to developing diagnostic criteria to distinguish inflammatory from non-inflammatory arthritis, enthesitis, dactylitis and spondylitis(58). An initial nominal group exercise involving clinicians,

researchers and patient focus groups (and facilitated by experts in qualitative research) was planned to generate a list of keywords, phrases and concepts that may assist in defining and/or distinguishing inflammatory from non-inflammatory arthritis. In acknowledgement of potential differences in expression between clinicians/researchers and patients themselves, a language harmonisation process was undertaken by a group of experts and patient researchers. The elements obtained from the nominal group exercise (and refined during the language harmonisation process) were then used to develop the questionnaire used in a subsequent Delphi process.

Defining and measuring consensus- Multiple approaches to defining and measuring consensus were employed. Brunner and colleagues asked study participants to rank items for importance on a scale of 1 to 10 and defined agreement as parameters with a median $\geq$ 7 and an interquartile range $\leq$ 3(71). Two groups considered the level of panellist expertise. Klein Meuleman *et al* asked study participants to indicate whether or not a topic or question related to their area of expertise prior to voting(50). Consensus for a given question was defined as an agreement level of 70% amongst all participants excluding those who indicated "not my expertise." Muller and colleagues defined consensus as being reached when $\geq$ 75% of the panel and $\geq$ 75% of a subsample of self-identified master experts on a specific topic were in agreement(52). Rinella and colleagues employed different consensus cut-offs for different topics or questions. A general consensus threshold of $\geq$ 67% was adopted for most topics except for considering the presence of stigmatising language and deciding whether or not to retain the existing disease definition(63).

2. *Literature-based approaches*

Seventeen groups (12%) performed narrative reviews that considered factors such as historical perspectives, diagnostic feature prevalence and clinical practicality and utility of specific criteria, with synthesis of this information used to inform development of diagnostic criteria(45, 72-74). Another approach involved evaluating the prevalence of specific diagnostic features in the published literature to inform development of diagnostic criteria. Brandi and colleagues performed a systematic literature search across multiple electronic databases to identify studies that addressed diagnosis of hypophosphatasia in adults(75). A group of experts in the field identified 17 potential diagnostic features. The prevalence of these characteristics across the included studies was pooled, and all features with prevalence $\geq$ 50% were considered for inclusion as diagnostic criteria during subsequent working group discussions.

Reid *et al* adapted the GRADE approach for prognostic factors as part of their revision of Australian diagnostic criteria for fetal alcohol spectrum disorder(66), which involves prenatal alcohol exposure (PAE) as a preceding insult. Multiple systematic reviews and meta-analyses examining the association between PAE and possible diagnostic features were initially performed using the GRADE approach for prognostic factors. Modified evidence-to-decision frameworks (EtDFs) were subsequently generated for potential diagnostic features, accompanied by recommendations to include or exclude the feature. These EtDFs were drafted by the authors, piloted using available evidence for diagnostic features and then modified based on feedback from the Guidelines Development Group. The final iteration included domains related to strength of association, certainty of evidence, values, resources required, certainty of the resources required, equity, acceptability, feasibility and diagnostic accuracy. Following this, a separate overarching EtDF was created for the proposed diagnostic criteria as a whole, which included similar domains with additional consideration

of adoption implications to facilitate assessment of whether there was a clear net benefit of implementing the revised criteria over the previous criteria.

3. Primary study-based approaches

Seventeen groups (12%) recruited a cohort of patients with the condition of interest, measured the prevalence of certain features among the group, and then used statistical techniques to categorise features and thus derive diagnostic criteria. The methods used to establish whether potential study participants had the condition of interest varied between studies. Where available, existing commonly used diagnostic tools/aids were frequently employed. For example, Parker and colleagues used clinical interviews and DSM-V criteria to establish whether potential study participants had bipolar disorder(36). However, for other conditions that lacked existing purpose-built diagnostic tools, determination of whether potential study participants had the condition of interest was most commonly based on clinician diagnosis. For example, Matias-Guiu and colleagues identified patients with post-COVID cognitive dysfunction based on a combination of clinician diagnosis and results from general neuropsychological test data(76).

Specific statistical approaches included factor analysis(36), cluster analysis and random forest algorithm development(76). Another approach used in two studies (1%) was to build predictive models comprising varying combinations of potential diagnostic features(77, 78), construct ROC curves and then select the model with the highest area under the curve (AUC) value using clinical outcomes to discriminate true positive from false positive cases.

Modifying existing diagnostic criteria- Two studies (1%) revised existing diagnostic criteria to enhance their utility for specific clinical purposes. Berg *et al* sought to improve the

specificity of existing International Parkinson and Movement Disorder Society diagnostic criteria for Parkinson's disease(38) in early disease(79). The group removed components of the existing criteria that related to disease duration and adjusted 'red flag' features to absolute exclusions. They then evaluated the modified criteria in a cohort of 364 case and control patients from a larger validation study. The sensitivity and specificity of the modified criteria for 'clinically established early Parkinson's disease' were then reported. Ohayon and colleagues modified symptom frequency and duration criteria from the existing Diagnostic and Statistical Manual of Mental Disorders 5th Edition (DSM-V)(16) criteria for restless legs syndrome to identify patients most likely to benefit from treatment(80). A telephone survey involving 19,136 individuals assessed the prevalence and chronicity of specific features of restless legs syndrome. This informed adjustments to the frequency and duration components of the existing DSM-V criteria, resulting in reduced prevalence of the disorder.

Measures to Optimise the Clinical Utility of Diagnostic Criteria

Establishing levels of diagnostic certainty

Thirty-seven studies (27%) developed several sets of diagnostic criteria for the same condition corresponding to different levels of certainty. Many of these studies established 2-3 different diagnostic certainty levels. For example, Armstrong established criteria for 'probable' and 'possible' corticobasal degeneration(46) whereas Engelman defined 'confirmed,' 'clinical' and 'suspected' scabies(81) and Adam defined 'definitive,' 'probable,' and 'possible' criteria for Kabuki syndrome(82). There was a general trend towards lower diagnostic certainty criteria offering increased sensitivity at the cost of reduced specificity. Unlike the 'probable' corticobasal degeneration criteria developed by Armstrong *et al*, the 'possible' corticobasal degeneration criteria did not feature restrictions on age or family

history, permitted the presence of tau mutations and allowed for less rigorous phenotypic fulfillment(46). This resulted in the 'possible' criteria being more sensitive and less specific than the 'probable' criteria, increasing the potential for false positive diagnosis of a different tauopathy as corticobasal degeneration.

Consideration of language and terminology

Two author groups considered the implications of specific definitions, phrases and terminology on the clinical practicality of diagnostic criteria. To facilitate clinical implementation of the diagnostic criteria for paediatric restless legs syndrome that had been derived, Picchietti and colleagues provided a set of recommendations regarding real-world application of the criteria including consideration of language typically used by children to describe common symptoms(40). The potential contribution of language to stigmatisation of affected patients was considered in one study (1%). Rinella *et al* asked Delphi participants whether the terms 'non-alcoholic' and 'fatty' were potentially stigmatising in the context of liver disease. The affirmative sentiment amongst respondents subsequently informed revision of the disease name 'non-alcoholic fatty liver disease' to 'metabolic dysfunction-associated steatotic liver disease.'

Consideration of practicality

One study considered the impact of clinical or geographic setting on the practicality of derived diagnostic criteria. Engelman *et al* conducted a four-round Delphi process to develop diagnostic criteria for scabies. The panel included clinicians from fields including dermatology, paediatrics, infectious diseases and public health. During the first round of voting, panellists were asked to nominate (using open-ended questions) potential features that could be considered for inclusion in the diagnostic criteria during subsequent rounds of

voting. Participants were asked to consider diagnosis in both resource-limited/field settings as well as high income or office-based settings.

Minimisation of atypical case bias

Amongst the group of studies that developed diagnostic criteria based on the prevalence of certain features in the published literature, the potential for introduction of bias through use of atypical cases was considered in one study. Armstrong and colleagues developed diagnostic criteria for corticobasal degeneration through a combination of an international consortium of neurologists and neuropsychologists, data derived from a brain bank and review of published cases, as well as a systematic review evaluating clinicopathologic correlation studies(46). Prevalent clinical features described in brain banks or included studies were considered for inclusion in the diagnostic criteria. To address potential bias arising from the inclusion of atypical cases, the authors instituted an inclusion criterion of a minimum of five or more pathologically proven corticobasal degeneration cases when screening articles for their literature review.

DISCUSSION

This scoping review identified 139 studies that developed new or revised existing diagnostic criteria for conditions without objective tests, biomarkers or reference standards. A diverse range of approaches to diagnostic criteria development were reported, with consensus methodology (specifically Delphi or modified Delphi process) being the most commonly adopted. This diversity reflects the varied group of conditions captured and should be an important consideration in any future attempts to establish methodological guidance to support development of diagnostic criteria. Assessment of the credibility of specific methodologies employed in the different consensus processes adopted was limited by the quality of reporting. This finding is consistent with previous work examining methods used by expert panels to define reference standards(17) and stands to benefit from the recent introduction of specific guidance to support reporting of consensus methods(83, 84).

Our findings suggest the role of evidence in diagnostic criteria development could be significantly strengthened. Presentation of evidence during diagnostic criteria development was not reported in 26% of studies. Amongst the 74% of studies that did incorporate evidence, there was significant variation in the types of evidence considered and the use of formal or systematic methods to examine this evidence. Only a few studies conducted primary empirical research to develop diagnostic criteria and used statistically rigorous approaches. In the secondary research that was undertaken, the reviews were often not comprehensive or systematic, and risk of bias assessment or assessment of methodological quality was not performed. Tools such as the GRADE approach were typically not employed to assess the strength of recommendations. The appropriate use of evidence is essential to informing clinical decision-making, particularly during the diagnostic process, and also facilitates patient decision-making consistent with individual circumstances and values(85).

Therefore, systematic literature review and assessment of risk of bias/methodological quality should therefore feature in any diagnostic criteria development process to identify pre-existing diagnostic criteria or features and identify evidence gaps that could be feasibly addressed with primary research rather than consensus decision-making alone. Measures that may assist in improving the role of evidence in diagnostic criteria development include development of specific methodological guidance and mandating reporting of the level of evidence or strength of recommendations. Furthermore, the integration of living systematic reviews could provide an effective means of updating diagnostic criteria as new data emerge(86).

Equitable representation of affected interest holders amongst groups developing diagnostic criteria using consensus methodology is essential in ensuring recommendations are credible and applicable to the entire population with the condition(87). Our finding that just 5/98 (5%) of groups using consensus methodology adopted specific approaches to improve gender or racial/ethnic diversity suggests there is considerable room for improvement. The lack of a validated approach to selecting individuals for diagnostic criteria development panels(88) as well as the absence of actionable methodological guidance or a formal reporting guideline for this study type are both likely contributing factors and present directions for future research. There are several measures that may facilitate greater diversity amongst panels involved in developing diagnostic criteria using consensus methodology. For example, funding agencies could require detailed prospective explanation of how equity will be addressed in the panellist recruitment process prior to project initiation. The National Health and Medical Research Council (NHMRC) in Australia supports this by requiring diversity in panel membership(89), having a gender equity strategy(90) and having a statement on sex and gender in health research(91). Secondly, during study planning, dedicated time and resources

667 could be allocated to provide financial support, mentorship and training for individuals from
668 underrepresented groups to enable participation. There also needs to be greater transparency
669 with respect to reporting the composition of panels in publications, which could be supported
670 through the development of a reporting guideline specific to studies publishing new or
671 revised diagnostic criteria. Sufficient panel diversity should be defined prospectively. Both 1)
672 the methods and eligibility criteria or thresholds used to recruit expert panellists as well as 2)
673 the final panel composition should be clearly reported. Specific measures taken to achieve
674 sufficient panel diversity should also be made clear. Ultimately, diversity must no longer be
675 considered an optional 'tickbox'- it must become foundational and considered at all stages of
676 diagnostic criteria development.
677
678 There are several limitations to the current review to consider. The analysis was contingent
679 upon complete reporting of study methodology. Several studies either failed to report
680 important methodological details or provided very brief descriptions, which limited our
681 assessment. Secondly, while this scoping review described methodological approaches to
682 diagnostic criteria development, it did not comprehensively explore the role of other
683 important and potentially influential contextual factors such as cultural, geographic or
684 healthcare system differences. Thirdly, not all included studies were explicitly designed to
685 develop diagnostic criteria (despite employing methodological approaches that are relevant to
686 diagnostic criteria development). Inclusion of such studies therefore represents a limitation
687 with respect to the study objective, and the conclusions of our study should be interpreted
688 with this in mind.
689
690 In conclusion, this scoping review identified considerable variability in the methodological
691 approaches used to develop new or revised diagnostic criteria for conditions without

objective tests, biomarkers or reference standards. The diversity in methods underscores the complexity and challenges inherent in diagnosing such conditions. The findings highlight the critical need for more robust integration of evidence (including critical appraisal of evidence) in diagnostic criteria development. Currently, the inconsistent use of evidence and the varied approaches to appraisal suggest that many diagnostic criteria may not be as reliable or applicable as they could be. Strengthening the role of evidence in this process will ensure that diagnostic criteria are both scientifically sound and clinically useful. To address these challenges, specific methodological guidance could standardise the process of diagnostic criteria development. By prioritising evidence-based methodologies, we can enhance the reliability and utility of diagnostic criteria, ultimately improving patient care and clinical outcomes.

REFERENCES

1. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of Internal Medicine. 2005;165(13):1493-9.

2. Schiff GD, Volodarskaya M, Ruan E, Lim A, Wright A, Singh H, Reyes Nieva H. Characteristics of Disease-Specific and Generic Diagnostic Pitfalls: A Qualitative Study. JAMA Netw Open. 2022;5(1):e2144531.

3. Nair R, Aggarwal R, Khanna D. Methods of formal consensus in classification/diagnostic criteria and guideline development. Semin Arthritis Rheum. 2011;41(2):95-105.

4. Mestre TA, Fabbri M, Luo S, Stebbins GT, Goetz CG, Sampaio C. A Unified Framework for Evidence-Based Diagnostic Criteria Programs in Movement Disorders. Mov Disord. 2023;38(7):1156-62.

5. Gelfand T. Tuesdays at the Salpetriere. "Charcot the clinician: the Tuesday lessons, excerpts from nine case presentations on general neurology delivered at the Salpetriere Hospital in 1887-88." By Jean-Martin Charcot. Essay review. Bull Hist Med. 1989;63(1):132-6.

6. Doust JA, Bell KJL, Glasziou PP. Potential Consequences of Changing Disease Classifications. JAMA. 2020.

7. Carter SM, Degeling C, Doust J, Barratt A. A definition and ethical evaluation of overdiagnosis. J Med Ethics. 2016;42(11):705-14.

8. White S, Gong H, Zhu L, Doust J, Loh TP, Lord S, et al. Simulations found within-subject measurement variation in glycaemic measures may cause overdiagnosis of prediabetes and diabetes. J Clin Epidemiol. 2022;145:20-8.

9. Aggarwal R, Ringold S, Khanna D, Neogi T, Johnson SR, Miller A, et al. Distinctions between diagnostic and classification criteria? Arthritis Care Res (Hoboken). 2015;67(7):891-7.

10. Zoccali C, Tripepi G, Stel V, Fu EL, Mallamaci F, Dekker F, Jager KJ. Biomarkers in clinical epidemiology studies. Clin Kidney J. 2024;17(6):sfae130.

11. Schunemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020;122:129-41.

12. Hauser W, Brahler E, Ablin J, Wolfe F. Modified 2016 American College of Rheumatology Fibromyalgia Criteria, the Analgesic, Anesthetic, and Addiction Clinical Trial Translations Innovations Opportunities and Networks-American Pain Society Pain Taxonomy, and the Prevalence of Fibromyalgia. Arthritis Care Res (Hoboken). 2021;73(5):617-25.

13. Patricios JS, Schneider KJ, Dvorak J, Ahmed OH, Blauwet C, Cantu RC, et al. Consensus statement on concussion in sport: the 6th International Conference on Concussion in Sport-Amsterdam, October 2022. Brit J Sport Med. 2023;57(11):695-711.

14. Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV, Definition WCC. A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infectious Diseases. 2022;22(4):E102-E7.

15. Bower C, Elliott EJ, on behalf of the Steering Group. Report to the Australian Government Department of Health:"Australian Guide to the diagnosis of Fetal Alcohol Spectrum Disorder (FASD)". 2020.

16. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Arlington, United States; 2013.

17. Bertens LC, Broekhuizen BD, 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;10(10):e1001531.

18. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. Jbi Evid Implement. 2021;19(1):3-10.

19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169(7):467-73.

20. Montori VM, Guyatt GH. Progress in evidence-based medicine. JAMA. 2008;300(15):1814-6.

21. Lehane E, Leahy-Warren P, O'Riordan C, Savage E, Drennan J, O'Tuathaigh C, et al. Evidence-based practice education for healthcare professions: an expert view. BMJ Evid Based Med. 2019;24(3):103-8.

22. Clark JM, Sanders S, Carter M, Honeyman D, Cleo G, Auld Y, et al. Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. J Med Libr Assoc. 2020;108(2):195-207.

23. Veritas Health Innovation. Covidence systematic review software. Melbourne, Australia: Veritas Health Innovation; 2021.

24. RStudio, Team. RStudio: Integrated Development for R. RStudio. Boston, MA: PBC; 2020.

25. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth. 2023;21(3):520-32.

788 26. Di Nardo G, Barbara G, Borrelli O, Cremon C, Giorgio V, Greco L, et al. Italian
789 guidelines for the management of irritable bowel syndrome in children and adolescents : Joint
790 Consensus from the Italian Societies of: Gastroenterology, Hepatology and Pediatric
791 Nutrition (SIGENP), Pediatrics (SIP), Gastroenterology and Endoscopy (SIGE) and
792 Neurogastroenterology and Motility (SINGEM). Ital J Pediatr. 2024;50(1):51.
793 27. Moore A, Patterson C, Lee L, Vedel I, Bergman H, Canadian Consensus Conference
794 on the D, Treatment of D. Fourth Canadian Consensus Conference on the Diagnosis and
795 Treatment of Dementia: recommendations for family physicians. Can Fam Physician.
796 2014;60(5):433-8.
797 28. Khan AA, Brandi ML, Rush ET, Ali DS, Al-Alwani H, Almonaei K, et al.
798 Hypophosphatasia diagnosis: current state of the art and proposed diagnostic criteria for
799 children and adults. Osteoporos Int. 2024;35(3):431-8.
800 29. Hoglinger GU, Respondek G, Stamelou M, Kurz C, Josephs KA, Lang AE, et al.
801 Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria.
802 Mov Disord. 2017;32(6):853-64.
803 30. SIGN SIGN. A Guideline Developer's Handbook. Edinburgh, UK: SIGN; 2008.
804 31. Tinuper P, Bisulli F, Cross JH, Hesdorffer D, Kahane P, Nobili L, et al. Definition
805 and diagnostic criteria of sleep-related hypermotor epilepsy. Neurology. 2016;86(19):1834-
806 42.
807 32. Gronseth GS, Woodroffe LM, Getchius TSD. Clinical practice guideline process
808 manual. Minnesota, United States: American Academy of Neurology; 2011.
809 33. Landgraf MN, Nothacker M, Kopp IB, Heinen F. The diagnosis of fetal alcohol
810 syndrome. Dtsch Arztebl Int. 2013;110(42):703-10.

811 34. CEBM CfE-bM. Oxford Centre for Evidence-based Medicine: levels of evidence
812 Oxford, UK: Nuffield Department of Primary Care Health Sciences, University of Oxford;
813 2009 [
814 35. Wenning GK, Stankovic I, Vignatelli L, Fanciulli A, Calandra-Buonaura G, Seppi K,
815 et al. The Movement Disorder Society Criteria for the Diagnosis of Multiple System Atrophy.
816 Mov Disord. 2022;37(6):1131-48.
817 36. Parker G, Tavella G, Ricciardi T, Hadzi-Pavlovic D, Alda M, Hajek T, et al. Refined
818 diagnostic criteria for the bipolar disorders: phase two of the AREDOC project. Acta
819 Psychiatr Scand. 2020;142(3):193-202.
820 37. van de Merwe JP, Nordling J, Bouchelouche P, Bouchelouche K, Cervigni M, Daha
821 LK, et al. Diagnostic criteria, classification, and nomenclature for painful bladder
822 syndrome/interstitial cystitis: an ESSIC proposal. Eur Urol. 2008;53(1):60-7.
823 38. Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, et al. MDS clinical
824 diagnostic criteria for Parkinson's disease. Mov Disord. 2015;30(12):1591-601.
825 39. Healy D, Bahrick A, Bak M, Barbato A, Calabro RS, Chubak BM, et al. Diagnostic
826 criteria for enduring sexual dysfunction after treatment with antidepressants, finasteride and
827 isotretinoin. Int J Risk Saf Med. 2022;33(1):65-76.
828 40. Picchietti DL, Bruni O, de Weerd A, Durmer JS, Kotagal S, Owens JA, et al. Pediatric
829 restless legs syndrome diagnostic criteria: an update by the International Restless Legs
830 Syndrome Study Group. Sleep Med. 2013;14(12):1253-9.
831 41. Arnold LM, Bennett RM, Crofford LJ, Dean LE, Clauw DJ, Goldenberg DL, et al.
832 AAPT Diagnostic Criteria for Fibromyalgia. J Pain. 2019;20(6):611-28.
833 42. Field JJ, Ballas SK, Campbell CM, Crosby LE, Dampier C, Darbari DS, et al.
834 AAAPT Diagnostic Criteria for Acute Sickle Cell Disease Pain. J Pain. 2019;20(7):746-59.

43. Litvan I, Goldman JG, Troster AI, Schmand BA, Weintraub D, Petersen RC, et al. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Mov Disord. 2012;27(3):349-56.

44. Cha YH, Baloh RW, Cho C, Magnusson M, Song JJ, Strupp M, et al. Mal de debarquement syndrome diagnostic criteria: Consensus document of the Classification Committee of the Barany Society. J Vestib Res. 2020;30(5):285-93.

45. Jacobs S, Mazure C, Prigerson H. Diagnostic criteria for traumatic grief. Death Stud. 2000;24(3):185-99.

46. Armstrong MJ, Litvan I, Lang AE, Bak TH, Bhatia KP, Borroni B, et al. Criteria for the diagnosis of corticobasal degeneration. Neurology. 2013;80(5):496-503.

47. Ciron J, Gueguen A, Al Khedr A, Bourre B, Clavelou P, Defer G, et al. Secondary progressive multiple sclerosis: A national consensus paper on diagnostic criteria. Rev Neurol (Paris). 2022;178(10):1098-104.

48. Albanese A, Bhatia KP, Cardoso F, Comella C, Defazio G, Fung VSC, et al. Isolated Cervical Dystonia: Diagnosis and Classification. Mov Disord. 2023;38(8):1367-78.

49. Collins DW, Rehak D, Dawes A, Collins DP, Daly C, Wagner ER, Gottschalk MB. Cubital Tunnel Syndrome: Does a Consensus Exist for Diagnosis? J Hand Surg Am. 2025;50(2):230 e1- e7.

50. Klein Meuleman SJM, Murji A, van den Bosch T, Donnez O, Grimbizis G, Saridogan E, et al. Definition and Criteria for Diagnosing Cesarean Scar Disorder. JAMA Netw Open. 2023;6(3):e235321.

51. Castro-Calvo J, King DL, Stein DJ, Brand M, Carmi L, Chamberlain SR, et al. Expert appraisal of criteria for assessing gaming disorder: an international Delphi study. Addiction. 2021;116(9):2463-75.

52. Muller A, Laskowski NM, Trotzke P, Ali K, Fassnacht DB, de Zwaan M, et al. Proposed diagnostic criteria for compulsive buying-shopping disorder: A Delphi expert consensus study. J Behav Addict. 2021;10(2):208-22.

53. Drossman DA. The functional gastrointestinal disorders and the Rome III process. Gastroenterology. 2006;130(5):1377-90.

54. Allen RP, Picchietti DL, Garcia-Borreguero D, Ondo WG, Walters AS, Winkelman JW, et al. Restless legs syndrome/Willis-Ekbom disease diagnostic criteria: updated International Restless Legs Syndrome Study Group (IRLSSG) consensus criteria--history, rationale, description, and significance. Sleep Med. 2014;15(8):860-73.

55. Choon SE, van de Kerkhof P, Gudjonsson JE, de la Cruz C, Barker J, Morita A, et al. International Consensus Definition and Diagnostic Criteria for Generalized Pustular Psoriasis From the International Psoriasis Council. JAMA Dermatol. 2024;160(7):758-68.

56. van Veen SMP, Evans N, Ruissen AM, Vandenberghe J, Beekman ATF, Widdershoven GAM. Irremediable Psychiatric Suffering in The Context of Medical Assistance in Dying: A Delphi-Study. Can J Psychiatry. 2022;67(10):758-67.

57. Jagasia MH, Greinix HT, Arora M, Williams KM, Wolff D, Cowen EW, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: I. The 2014 Diagnosis and Staging Working Group report. Biol Blood Marrow Transplant. 2015;21(3):389-401 e1.

58. Mease PJ, Garg A, Helliwell PS, Park JJ, Gladman DD. Development of criteria to distinguish inflammatory from noninflammatory arthritis, enthesitis, dactylitis, and spondylitis: a report from the GRAPPA 2013 Annual Meeting. J Rheumatol. 2014;41(6):1249-51.

59. Smith ME, Hardman JC, Mehta N, Jones GH, Mandavia R, Anderson C, et al. Acute otitis externa: Consensus definition, diagnostic criteria and core outcome set development. PLoS One. 2021;16(5):e0251395.

60. Branagan A, Hurley T, Quirke F, Devane D, Taneri PE, Badawi N, et al. Consensus definition and diagnostic criteria for neonatal encephalopathy-study protocol for a real-time modified delphi study. Pediatr Res. 2025;97(1):430-6.

61. Carruthers BM, van de Sande MI, De Meirleir KL, Klimas NG, Broderick G, Mitchell T, et al. Myalgic encephalomyelitis: International Consensus Criteria. J Intern Med. 2011;270(4):327-38.

62. Cummings J, Pinto LC, Cruz M, Fischer CE, Gerritsen DL, Grossberg GT, et al. Criteria for Psychosis in Major and Mild Neurocognitive Disorders: International Psychogeriatric Association (IPA) Consensus Clinical and Research Definition. Am J Geriatr Psychiatry. 2020;28(12):1256-69.

63. Rinella ME, Sookoian S. From NAFLD to MASLD: updated naming and diagnosis criteria for fatty liver disease. J Lipid Res. 2024;65(1):100485.

64. McKeith IG, Boeve BF, Dickson DW, Halliday G, Taylor JP, Weintraub D, et al. Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. Neurology. 2017;89(1):88-100.

65. Kalra S, Silman A, Akman-Demir G, Bohlega S, Borhani-Haghighi A, Constantinescu CS, et al. Diagnosis and management of Neuro-Behcet's disease: international consensus recommendations. J Neurol. 2014;261(9):1662-76.

66. Reid N, Hayes N, Akison LK, Vanderpeet C, Middleton P, Munn Z, and the AFGC. GRADE notes: using GRADE informed approaches when developing diagnostic criteria: example and reflections from the Australian guidelines for assessment and diagnosis of fetal alcohol spectrum disorder. Preprint. 2024.

67. Katz DI, Bernick C, Dodick DW, Mez J, Mariani ML, Adler CH, et al. National Institute of Neurological Disorders and Stroke Consensus Diagnostic Criteria for Traumatic Encephalopathy Syndrome. Neurology. 2021;96(18):848-63.

68. Leong P, Vertigan AE, Hew M, Baxter M, Phyland D, Hull JH, et al. Diagnosis of vocal cord dysfunction/inducible laryngeal obstruction: An International Delphi Consensus Study. J Allergy Clin Immunol. 2023;152(4):899-906.

69. Parker G, Spoelma MJ, Tavella G. The AREDOC project and its implications for the definition and measurement of the bipolar disorders: A summary report. Aust N Z J Psychiatry. 2022;56(11):1389-97.

70. Jeste DV, Finkel SI. Psychosis of Alzheimer's disease and related dementias. Diagnostic criteria for a distinct syndrome. Am J Geriatr Psychiatry. 2000;8(1):29-34.

71. Brunner F, Lienhardt SB, Kissling RO, Bachmann LM, Weber U. Diagnostic criteria and follow-up parameters in complex regional pain syndrome type I--a Delphi survey. Eur J Pain. 2008;12(1):48-52.

72. Freeman R, Edwards R, Baron R, Bruehl S, Cruccu G, Dworkin RH, Haroutounian S. AAPT Diagnostic Criteria for Peripheral Neuropathic Pain: Focal and Segmental Disorders. J Pain. 2019;20(4):369-93.

73. Mysliwiec V, Brock MS, Creamer JL, O'Reilly BM, Germain A, Roth BJ. Trauma associated sleep disorder: A parasomnia induced by trauma. Sleep Med Rev. 2018;37:94-104.

74. Wetselaar P, Lobbezoo F, de Vries R, Mehta SB, Opdam NJM, Loomans BAC. Developing diagnostic criteria for tooth wear, a preliminary beta version based on expert opinion, and a narrative literature review. J Oral Rehabil. 2023;50(10):1030-42.

75. Brandi ML, Khan AA, Rush ET, Ali DS, Al-Alwani H, Almonaei K, et al. The challenge of hypophosphatasia diagnosis in adults: results from the HPP International Working Group Literature Surveillance. Osteoporos Int. 2024;35(3):439-49.

76. Matias-Guiu JA, Herrera E, Gonzalez-Nosti M, Krishnan K, Delgado-Alonso C, Diez-Cirarda M, et al. Development of criteria for cognitive dysfunction in post-COVID syndrome: the IC-CoDi-COVID approach. Psychiatry Res. 2023;319:115006.

77. Esteba-Castillo S, Garcia-Alba J, Rodriguez-Hildago E, Vaquero L, Novell R, Moldenhauer F, Castellanos MA. Proposed diagnostic criteria for mild cognitive impairment in Down syndrome population. J Appl Res Intellect Disabil. 2022;35(2):495-505.

78. Lin YH, Chiang CL, Lin PH, Chang LR, Ko CH, Lee YH, Lin SH. Proposed Diagnostic Criteria for Smartphone Addiction. PLoS One. 2016;11(11):e0163010.

79. Berg D, Adler CH, Bloem BR, Chan P, Gasser T, Goetz CG, et al. Movement disorder society criteria for clinically established early Parkinson's disease. Mov Disord. 2018;33(10):1643-6.

80. Ohayon MM, Bagai K, Roberts LW, Walters AS, Milesi C. Refining duration and frequency thresholds of restless legs syndrome diagnosis criteria. Neurology. 2016;87(24):2546-53.

81. Engelman D, Fuller LC, Steer AC, International Alliance for the Control of Scabies Delphi p. Consensus criteria for the diagnosis of scabies: A Delphi study of international experts. PLoS Negl Trop Dis. 2018;12(5):e0006549.

82. Adam MP, Banka S, Bjornsson HT, Bodamer O, Chudley AE, Harris J, et al. Kabuki syndrome: international consensus diagnostic criteria. J Med Genet. 2019;56(2):89-95.

83. Logullo P, van Zuuren EJ, Winchester CC, Tovey D, Gattrell WT, Price A, et al. ACcurate COnsensus Reporting Document (ACCORD) explanation and elaboration: Guidance and examples to support reporting consensus methods. PLoS Med. 2024;21(5):e1004390.

84. Gattrell WT, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, et al. ACCORD (ACcurate COnsensus Reporting Document): A reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med. 2024;21(1):e1004326.

85. Djulbegovic B, Guyatt G. Evidence vs Consensus in Clinical Practice Guidelines. JAMA. 2019;322(8):725-6.

86. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, Gruen RL. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med. 2014;11(2):e1001603.

87. Welch VA, Akl EA, Guyatt G, Pottie K, Eslava-Schmalbach J, Ansari MT, et al. GRADE equity guidelines 1: considering health equity in GRADE guideline development: introduction and rationale. J Clin Epidemiol. 2017;90:59-67.

88. Weiss DJ, Shanteau J. Who's the best? A relativistic view of expertise. Appl Cognit Psychol. 2014;28(4):447-57.

89. NHMRC. Guidelines for guidelines: guideline development group 2019 [Available from: https://nhmrc.gov.au/guidelinesforguidelines/plan/guideline-development-group.

90. NHMRC. NHMRC gender equity strategy 2022-2025 2022 [Available from: https://www.nhmrc.gov.au/research-policy/gender-equity/nhmrc-gender-equity-strategy-2022-2025?utm_source=chatgpt.com.

91. NHMRC. Statement on sex, gender, variations of sex characteristics and sexual orientation in health and medical research 2024 [Available from: https://www.nhmrc.gov.au/research-policy/gender-equity/statement-sex-and-gender-health-and-medical-research.

92. Peterson D, Keeley JW. Syndrome, disorder, and disease. In: Cautin RL, Lilienfeld SO, editors. The Encyclopedia of Clinical Psychology. 1st ed. Hoboken, New Jersey: John Wiley & Sons Inc; 2015.

93. Calvo F, Karras BT, Phillips R, Kimball AM, Wolf F. Diagnoses, syndromes, and diseases: a knowledge representation problem. AMIA Annu Symp Proc. 2003;2003:802.

94. Field MJ, Lohr KN. Clinical practice guidelines: directions for a new program. Washington, US: National Academy Press; 1990.

95. Swisher LL, Arsalanian LE, Davis CM. The realm-individual-process-situation (RIPS) model of ethical decision-making. HPA Resource. 2005;5(3):3-8.

96. Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. BMJ. 2016;353:i2089.

97. ICER. Value Assessment Framework. Institute for Clinical and Economic Review,; 2023.

98. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med. 1999;18(22):2987-3003.

99. Reitsma JB, Rutjes AW, 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;62(8):797-806.

100. Pepe MS, Janes H. Insights into latent class analysis.; 2005. Contract No.: paper 236.

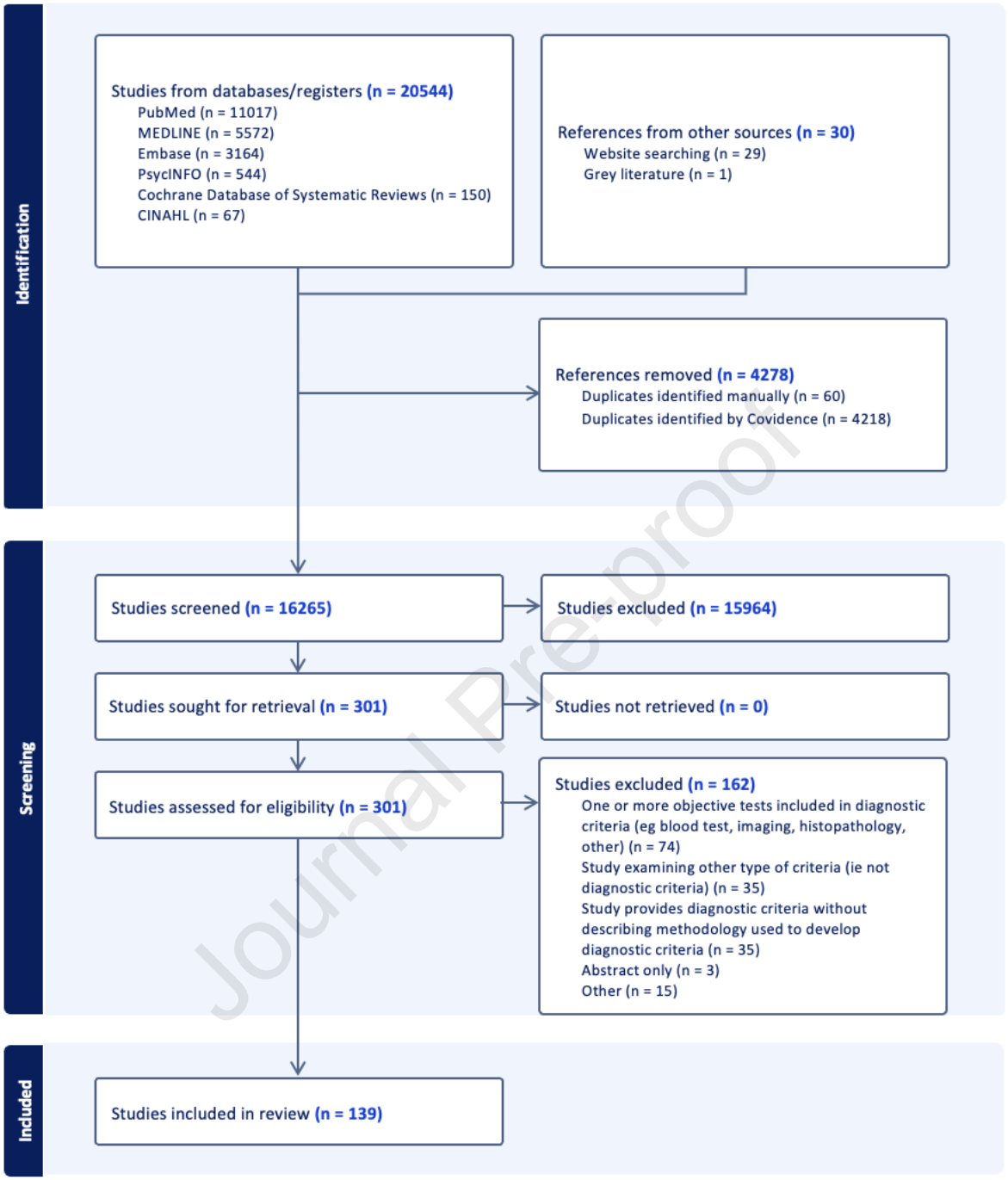
FUNDING

No specific funding was obtained to support this work.

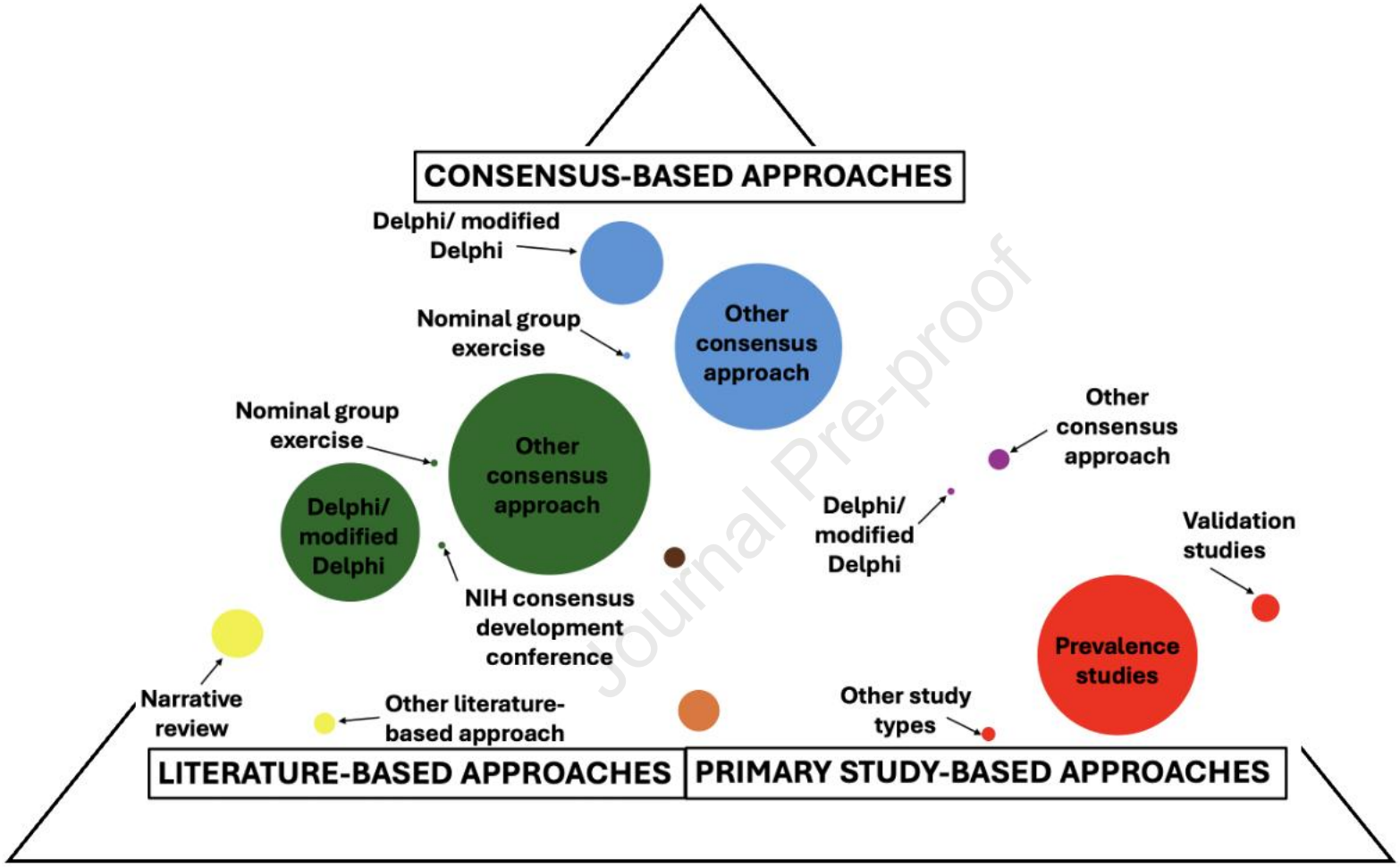
FIGURE LEGENDS

Figure 1. PRISMA flowchart summarising study selection process.

Figure 2. Bubble chart illustrating different methodological approaches used to develop new or revised diagnostic criteria amongst included studies. Bubble location and colour reflects the specific methodological approach (or combination of methodological approaches adopted): blue reflects consensus-based approaches, yellow literature-based approaches, red primary study-based approaches, green a combination of consensus- and literature-based approaches, orange a combination of literature- and primary study-based approaches, purple a combination of consensus- and primary study-based approaches and brown a combination of consensus-, literature- and primary study-based approaches. Bubble size is proportionate to the number of studies.



1050



1051

1052

1053

TABLES

Table 1. Characteristics of included studies.

Characteristic		*n* (%)
General information		
Publication type		
	Original research article	115 (83)
	Clinical practice guideline	17 (12)
	Narrative review	4 (3)
	Other	3 (2)
Place of publication		
	North America	65 (47)
	Europe	56 (40)
	Asia	9 (6)
	Australia	8 (6)
	South America	1 (1)
Number of institutions involved		
	≤ 4	32 (23)
	≥ 5	107 (77)
New vs revised diagnostic criteria developed		

	New	59 (42)
	Revised	80 (58)
Disease/condition information		
Disease/condition category		
	Neurological	40 (29)
	Psychological/psychiatric	36 (26)
	Rheumatological	17 (12)
	Multisystem/syndromic	11 (8)
	Other	35 (25)
Disease/condition associated with preceding insult		
	Yes	41 (29)
	No	98 (71)
Methodological information		
Formal level/strength of evidence assessment		
	Yes	6 (4)
	No	133 (96)
Patient/consumer advocate involvement		
	Yes	18 (13)
	No	121 (87)

Levels of diagnostic certainty		
	Single	102 (73)
	Multiple	37 (27)
External validation of diagnostic criteria planned or performed		
	Yes	4 (3)
	No	135 (97)
Diagnostic criteria updating processes specified		
	Yes	1 (1)
	No	138 (99)

1064 **Table 2.** Exemplar studies for each main methodological approach to developing diagnostic criteria.

Methodological Approach	Author, Year	Disease/Condition	Summary of Methods
Consensus-based			
	Choon, 2024	Pustular psoriasis	The International Psoriasis Council convened a Pustular Psoriasis Working Group of global experts who conducted a modified Delphi process. Expert panelists reviewed 64 challenging cases to generate 43 diagnostic statements which were refined through two rounds of virtual consensus meetings. Statements achieving ≥80% agreement informed the final definition and diagnostic criteria, with features selected through iterative voting and structured expert feedback
	Edmond, 2023	Prescription opioid dependence syndrome	Three-round Delphi study with 38 expert participants in pain, addiction medicine, and opioid use disorder (OUD). Participants recruited from a Veterans Health Administration State-of-the-Art (SOTA) conference. In round 1, qualitative responses were analysed to identify themes and potential diagnostic criteria. In rounds 2 and 3, participants rated the criteria using Likert scales to reach consensus (defined as a median score >5), and a follow-up expert workgroup refined the final criteria
Literature-based			
	Lehman, 2008	Schinzel-Giedion syndrome	Review of 46 published cases performed and frequencies of craniofacial, skeletal, visceral, and neurological abnormalities calculated. Criteria formulated based on prevalence and specificity of certain clinical signs

	Rothman, 2014	Circumferential skin folds syndrome	Literature review including 42 cases with characteristic phenotype. Reported associated anomalies were compiled, categorized and sorted by frequency. Major and minor diagnostic criteria were developed based on frequency across reported cases
Primary study-based			
	Esteba-Castillo, 2022	Mild cognitive impairment in Down syndrome	Longitudinal design over three years, assessing two groups of adults with Down syndrome (control and prodromal). Cognitive and behavioral data were collected at three time points. Linear and generalized linear mixed models were used to identify predictors of prodromal decline. Diagnostic features were selected based on statistically significant differences between groups
	Matias-Guiu, 2022	Cognitive dysfunction in post-COVID condition	The study adapted the IC-CoDE framework, originally used in epilepsy, to develop the IC-CoDi-COVID criteria. 404 patients with post-COVID condition were evaluated with neuropsychological batteries. Cognitive impairment was defined based on performance in five cognitive domains using thresholds of 1 and 1.5 standard deviations below the normative mean. Impairment in each domain required both representative tests to fall below the threshold. Machine learning (hierarchical clustering and random forest classification) was used to validate domain impairment thresholds and phenotypic classification
Multiple approaches			
	Brandi, 2023	Hypophosphatasia	Hypophosphatasia (HPP) International Working Group conducted a systematic review to identify diagnostic features associated with adult hypophosphatasia. Diagnostic characteristics were initially proposed by clinical experts with ≥5 years of experience. These were analyzed for pooled prevalence and diagnostic relevance. Decisions on inclusion as diagnostic features were informed by both prevalence in the literature and expert judgment regarding specificity. Features with pooled prevalence >50% were considered for major criteria; those <50% for minor. However, non-specific high-prevalence findings (eg musculoskeletal pain) were

			excluded or downgraded based on their diagnostic utility. No objective thresholds or scoring tools were used. Decisions were based on pooled proportions and group consensus
	Di Nardo, 2024	Irritable bowel syndrome	Delphi consensus process involving experts from four Italian pediatric and gastroenterology societies. The panel identified 22 key clinical questions and used the PICO framework to guide systematic literature reviews. Statements were rated using the GRADE approach, and consensus was defined as ≥80% agreement during one or two rounds of blinded voting. External reviewers also validated the final set of recommendations

SUPPLEMENTARY FILES

Supplement 1. Search strategy for Medline database.

1. (criteria adj2 (develop* or method* or consensus or evidence based))ti,ab,kf
2. (disorder or condition or syndrome or disease).ti,ab,kf
3. 1 and 2
4. limit 3 to yr="2000–2024"

Supplement 2. List of websites used in supplementary literature search.

- Agency for Healthcare Research and Quality (https://www.ahrq.gov/)
- Medical Services Advisory Committee (http://www.msac.gov.au/)
- National Health and Medical Research Council (https://www.nhmrc.gov.au/)
- National Health Service (https://www.nhs.uk/)
- National Institute for Health and Care Excellence (https://www.nice.org.uk/guidance)
- National Institutes of Health (https://www.nih.gov/)
- US Centres for Disease Control and Prevention (https://www.cdc.gov/index.html)
- World Health Organisation (https://www.who.int/)

Supplement 3. Data extraction template for studies describing new/revised diagnostic criteria.

General Characteristics

1) Full title of source:
2) First author (last name):
3) Year of publication:
4) Place of publication
    a) Country of first author's first-listed affiliation:
    b) Organisational affiliation:
5) Type of evidence source
    a) Source type (diagnostic criteria, clinical guideline, methodological study, other- specify):
    b) Site of publication (peer-review journal, website, conference proceeding/presentation, other- specify)
        i) Grey literature? (yes/no)

Clinical Context

1) Population
    a) Age group (paediatric, adolescent/young adult (12-30 years), middle-aged adult (30-60 years), older adult (> 60 years), all adult, not specified)
    b) Geographic location in which diagnostic criteria are intended for use (Africa, Asia, Australia, Europe, North America, South America, global, not specified):
1) Condition

a) Name of condition:
b) Is the condition associated with a preceding insult/injury (yes/no):
c) Have diagnostic criteria for this condition previously been published? (yes/no/not specified):
    i) If YES, provide link to previous criteria
    ii) If YES, has one or more of the authors/affiliated organisations been involved in development of the previous diagnostic criteria (yes/no)?

2) Rationale for formulating/revising diagnostic criteria provided? (yes/no):
3) Target audience as described by the authors (surgeons, specialist physicians, general practitioners, public health clinicians, allied health professionals, researchers, members of the general public, patients/clients, other- specify, not specified):

Methodological Details

1) Specific clinical/non-clinical roles of individuals involved in diagnostic criteria development (surgeons, specialist physicians, general practitioners, researchers, public health clinicians, allied health professionals, members of the general public, patients/clients, other- specify):
    a) Did the clinical/non-clinical role of at least 1 member of the panel match that of the target audience? (yes/no):
2) Methodology used to ensure appropriate diversity within the panel involved in diagnostic criteria development (yes/no):
    a) If YES, specify methodology:
3) Number of affiliated organisations involved in diagnostic criteria development:
4) General stages/phases of diagnostic criteria development
    a) Planning of diagnostic criteria development methodology *a priori* (yes/no):

i) If YES, specify how this was planned (based on previous methodology, based on evidence, based on subjective judgement, other- specify:

ii) If YES, specify whether rationale for specific methodology was provided (yes/no):

b) Consideration of ethical issues (yes/no):

i) If YES, how were ethical issues considered? (panel discussion, consultation with non-panel member interest holders, other- specify)

c) Consideration/presentation of evidence (yes/no):

i) If YES, specify the sources of evidence considered (source type- case report, case series, DTA study, case control study, prospective cohort study, retrospective cohort study, randomised controlled trial, scoping review, narrative review, systematic review without meta-analysis, systematic review with meta-analysis, other- specify):

(1) If SYSTEMATIC REVIEW was considered, specify whether it used GRADE (yes/no):

ii) If YES, specify the year of publication of these sources:

iii) If YES, specify whether quality assessment of the sources was performed (yes/no):

(1) If YES, specify how quality assessment was performed (subjective discussion, use of tool- specify, other- specify)

iv) If YES, specify whether any specific sign/symptom/test performance measures were considered (sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, negative likelihood ratio, other-specify, no):

v) If YES, specify whether an evidence assessment framework was adopted (yes/no):

(1) If YES, specify whether this was a previously published framework (eg GRADE evidence to decision framework) (yes/no)

(a) If YES, specify the name of the framework

(2) If YES, specify the framework elements/criteria that were considered

d) Was modification of disease/condition definition considered or performed? (yes/no):

e) Revision of previously published diagnostic criteria in cases where previously published diagnostic criteria were available (yes/no/not applicable):

i) If YES, which of the following factors have been discussed/considered?

(1) Number of people affected (yes/no):

(2) Prognostic ability compared to previous iteration of diagnostic criteria (yes/no/not applicable)

(a) If YES, specify:

(3) Incremental benefit for patients diagnosed using revised criteria (yes/no/not applicable)

(a) If YES, specify:

(4) Incremental harm for patients diagnosed using the revised criteria (yes/no)

(a) If YES, specify:

f) Use of consensus methodology (yes/no)

i) If YES, specify type of consensus methodology used (Delphi, nominal group technique, RAND/UCLA Appropriateness Method, NIH consensus development conference, other- specify)

g) Use of non-consensus methodology (yes/no)

i) If YES, specify type of non-consensus methodology used

h) Consideration of minority opinions and dissenting views (yes/no):

i) If YES, specify how this was done (discussion of anonymised feedback, discussion of non-anonymised feedback, other- specify)

i) End user engagement- was 1 or more members of the target audience consulted on the appropriateness/application of the diagnostic criteria developed? (yes/no):

i) If YES, how many members of the target audience were consulted?

(1) If YES, how was this consultation performed? (inclusion in panel, face-to-face discussion with panel after criteria had been developed, conference presentation, online or virtual interaction, other-specify):

j) Updating processes planned (yes/no):

i) If YES, specify processes planned:

(1) If YES, specify rationale provided for planned processes:

5) Sources of funding (government, non-government organisation, industry/commercial, other- specify):

6) Has a formal external review process of the diagnostic criteria by a separate group/organisation been planned or performed? (yes/no):

a) If YES, specify (planned/performed):

i) If PERFORMED, specify how:

1228 **Supplement 4. Dictionary of terms.**

Term	Definition
Diagnostic criteria	Collection of symptoms, clinical signs and test results that support or refute the diagnosis of a specific disease, syndrome or disorder.
Nomenclature for abnormal/pathological states	
Condition	Umbrella term that can be used to describe disease, syndrome and disorder.
Disease	Distinct biological process with a defined cause(92, 93).
Disorder	An interruption or disturbance to normal function(92, 93).
Syndrome	Characteristic group of signs and symptoms that co-occur(92, 93).
Source types eligible for inclusion	
Clinical guideline	Systematically developed statement designed to assist clinician and/or patient decision-making with respect to specific clinical scenarios(94).
No objective test, biomarker or reference standard	Diagnostic criteria do not feature any objective tests such as blood tests, imaging studies or histopathological examination
Data extraction items	
Allied health professional	University-qualified health care professionals such as psychologists, physiotherapists, occupational therapists, speech pathologists, dieticians, podiatrists and exercise physiologists.
Affiliated organisation	Group or organisation who has commissioned development of diagnostic criteria or provided funding and/or other resources to support development of diagnostic criteria.
Consensus methodology	Formal method by which a group of individuals representing a broad range of knowledge and experience is assembled to interact with a view to identify a central tendency and grade the level of agreement reached(3). Examples include Delphi, nominal group technique, RAND/UCLA Appropriateness Method and NIH consensus development conference.
End user engagement	Formal planned consultation with one or more members of the target audience of a set of diagnostic criteria or clinical guideline.

Ethical issue	A situation that may compromise one or more moral values deemed to be socially legitimate and therefore worthy of respect(95).
Evidence assessment framework	Structured approach to evidence synthesis/review that involves addressing multiple elements/criteria in order to reach a formal decision or recommendation. An example is the GRADE Evidence to Decision framework(96) and the ICER Value Assessment Framework(97).
External review processes	Formal planned process by which key interest holders such as members of the target audience and patients/clients are consulted regarding the appropriateness, application or implications of new/revised diagnostic criteria.
General practitioner	Medical doctor with qualification obtained from an internationally recognised college/board of general practice and/or family medicine.
Grey literature	Information that is either unpublished or informally/non-commercially published. Examples include websites and conference proceedings.
Non-consensus methodology	Formal technique or strategy used to establish diagnostic criteria without involving consensus methodology. Examples include establishing a composite reference standard by combining multiple tests(98, 99) or performing latent class analysis whereby a condition is defined as a statistical model-based entity using multiple tests(99, 100).
Preceding insult/injury	Clearly defined exposure that precipitates a condition and is included within the diagnostic criteria for that condition.
Previously published diagnostic criteria	Diagnostic criteria that were publicly available prior to the date of publication of a new/revised set of diagnostic criteria.
Public health clinician	Medical doctor with specialist qualification obtained from an internationally recognised college/board of public health medicine.
Researcher	University-qualified individual who performs health-related research as part of their role within a university/college, independent institute or for a commercial organisation.
Separate group/organisation	Group or organisation comprising individuals who were not involved in

	development of the diagnostic criteria. The 'separate organisation' must not be listed as one of the 'affiliated organisations'
Specialist physician	Medical doctor with specialist qualification obtained from an internationally recognised college/board of physicians.
Surgeon	Medical doctor with specialist qualification obtained from an internationally recognised college/board of surgery.
Updating process	Formal planned process by which diagnostic criteria will be iteratively updated in the future.

Supplement 5. Key methodological characteristics of studies publishing new or revised diagnostic criteria.

Author, Year	Condition	Mandatory preceding insult?	New or revised criteria	Sum	Formal level/strength vidence assessment	Patient, consumer or other advocate involvement	Multiple levels of diagnostic certainty?
Albanese, 2023	Isolated cervical dystonia	No	New	Consensus committee formed by the International Parkinson and Movement Disorder Society Dystonia Study Group conducted a structured literature review and held four online meetings and multiple mail rounds between 2021 and 2022. Using a consensus development methodology, the group developed diagnostic criteria based on clinical phenomenology, supportive and exclusion criteria- the manuscript was drafted based on both findings of the literature review and panel discussions; however the specific role of evidence was not described in detail. Criteria were subsequently refined through structured discussion rounds and reviewed until full consensus was reached among expert contributors	No	No	Yes 1) definite idiopathic isolated cervical dystonia and 2) probable idiopathic isolated cervical dystonia
Albert, 2011	Mild cognitive impairment due to Alzheimer's disease	No	New	Workgroup convened by the National Institute on Aging and the Alzheimer's Association developed criteria over a structured review and consensus process. Inclusion of features was guided by expert judgement and whilst reference is made to literature, the specific role of evidence in informing decision-making is not described in detail	No	No	Yes 1) MCI due to AD high likelihood, 2) MCI due to AD intermediate likelihood and 3) MCI unlikely due to AD
Allen, 2003	Restless legs syndrome	No	Revised	Criteria developed through a workshop hosted by the NIH in collaboration with the IRLSSG. Participants included restless legs syndrome experts and authorities on epidemiology and scale design. Consensus was reached during structured in-person sessions and follow-up exchanges. The revised criteria were approved by the IRLSSG executive committee and include adaptations for special populations and augmentation	No	No	Yes 1) definite restless legs syndrome, 2) probable restless legs syndrome and 3) possible restless legs syndrome
Allen, 2014	Restless legs syndrome	No	Revised	Four-year international, interdisciplinary consensus process was used, initiated by a 2008 clinical standards workshop. IRLSSG members were organized into workgroups on differential diagnosis, clinical significance, and testing, which developed proposals through teleconferences and a workshop. The criteria were reviewed and approved by the IRLSSG membership and Executive Committee	No	No	No
Allison, 2010	Night eating syndrome	No	New	Diagnostic criteria were developed at the first International Night Eating Symposium in 2008. Criteria were developed based on expert clinical experience. Whilst there is reference to literature including studies performing psychometric validation of diagnostic features, the specific role of evidence in informing decision-making is not described in detail	No	No	No
Arai, 2023	Cachexia	Yes	New	The Asian Working Group for Cachexia (AWGC) initially performed a systematic review examining potential diagnostic features before employing a structured Delphi consensus process comprising three survey rounds and five meetings to vote on criteria. Experts were recruited from across Asia with clinical and research expertise in cachexia. Items were rated using a 10-point Likert scale, and consensus was defined as ≥70% of respondents scoring items 7 or higher	No	No	No
Armstrong, 2013	Corticobasal degeneration	No	New	International panel of experts in behavioral neurology, neuropsychology, and movement disorders was convened and conducted a systematic review of published cases as well as analysis of brain bank data. Data from 267 nonoverlapping cases was obtained. Criteria were developed based on results of the evidence review as well as panel member opinion	No	No	No
Arnold, 2019	Fibromyalgia	No	New	The ACTTION-APS Pain Taxonomy (AAPT) initiative formed an international fibromyalgia working group composed of clinicians and researchers with expertise in fibromyalgia. This group conducted literature reviews, data analyses from large population-based UK studies, and consensus meetings (in-person, teleconferences, emails) to develop the criteria	No	No	No

Artuso, 2010	Early-onset seizure variant of Rett syndrome	No	New	9 patients (comprising new and previously reported) were evaluated using a 22-item clinical severity score and compared to patients with classic and Zappella variant Rett syndrome. Statistical comparisons were performed to delineate characteristic features of the variant	No	No	No
Baguley, 2014	Paroxysmal sympathetic hyperactivity (after acquired brain injury)	Yes	Revised	Consensus group of international experts formed through invitations based on literature contributions and clinical expertise. Using a Delphi process, participants rated the importance of 16 potential diagnostic features using a visual analogue scale, with items below a threshold median score removed. Two rounds of scoring were conducted, and consensus was achieved on 11 core items, which were operationalized into a diagnostic tool using a probabilistic approach	No	No	Yes 1) probable PSH, 2) possible PSH, and 3) unlikely PSH
Baker, 2012	Tobacco use disorder	No	Revised	Authors provide a critique of the existing DSM-IV criteria based on conceptual clarity, psychometric performance, and predictive validity. They propose revisions informed by empirical data on smoking behaviors, including craving intensity, withdrawal symptoms, heaviness of smoking, and latency to first cigarette. Recommendations are based on comparisons with alternative measures, prioritising constructs with demonstrated predictive validity and clinical utility	No	No	No
Barros, 2020	Disorders of arousal	No	New	Retrospective case–control study analyzing N3 sleep interruptions recorded via video polysomnography in adults with DOA and control participants. Researchers manually scored behavioral features (eg eye opening, head raising, fear expressions) during N3 interruptions and identified a reproducible cutoff (≥2 N3 interruptions with eye opening per night) to distinguish DOA. Statistical tests (including ROC analysis, Rasch modeling, and interrater agreement) were used to assess the sensitivity and specificity of proposed behavioral markers	Not applicable	No	No
Bennett, 2014	Fibromyalgia	No	Revised	Eight clinicians from the US assessed 321 patients with fibromyalgia or other chronic pain disorders using structured questionnaires. Validation of the modified 2010 ACR criteria was conducted against the 1990 criteria. Alternative criteria were then derived from results of several of the structured questionnaires used	Not applicable	No	No
Berg, 2015	Dietary fructose intolerance in irritable bowel syndrome	No	New	Multicentre, open-label randomized controlled trial involving IBS patients, who underwent a 2-week habitual diet followed by 12 weeks with or without a fructose-reduced diet (FRD) and concluded with a fructose-rich provocation test. Diagnostic criteria were developed based on the combination of symptom relief (≥25 mm on VAS) during FRD and symptom aggravation (≥25 mm worsening on VAS) after provocation	Not applicable	No	No
Berg, 2018	Early Parkinson's disease	No	Revised	The 2015 MDS clinical diagnostic criteria were revised to enhance specificity for early Parkinson's disease by eliminating duration-related criteria and redefining red flags as absolute exclusions. The modified criteria were validated using data from a previously published 626-patient multicentre study, with 366 patients with disease duration <5 years selected. Feature inclusion was based on expert consensus and empirical performance (sensitivity/specificity) in the validation study	No	No	No
Black, 2000	Selective serotonin reuptake inhibitor discontinuation syndrome	Yes	New	Authors reviewed 46 published case reports and three additional studies of SSRI discontinuation symptoms, and extracted data on symptom type, timing, and frequency and synthesised these findings to propose diagnostic criteria. Decisions on feature inclusion were based on symptom frequency (≥10%), and distinctiveness	No	No	No
Bower, 2020	Fetal alcohol spectrum disorder	Yes	Revised	Systematic review used to identify existing diagnostic guidelines. Potential diagnostic features subsequently voted on during Delphi process involving FASD experts. 'Community Conversations' were used to engage patients, caregivers and advocates	No	Yes- series of 'Community Conversations' performed	No
Branagan, 2024	Neonatal encephalopathy	No	New	Five-phase approach based on the ACCORD guideline and using real-time Delphi methodology. An international multidisciplinary steering group including clinicians, researchers, and patient representatives was formed to oversee the process. Diagnostic criteria developed through a systematic review and structured online Real-time Delphi survey, with participants rating definitional components	No	Yes- parents and public representatives included in the	No

				across domains. Consensus defined as ≥80% agreement. Statements not reaching consensus will be reviewed at an expert consensus meeting chaired by James Lind Alliance facilitators		steering committee and will participate in consensus meetings	
Brandi, 2023	Hypophosphatasia	No	New	Hypophosphatasia (HPP) International Working Group conducted a systematic review to identify diagnostic features associated with adult hypophosphatasia. Diagnostic characteristics were initially proposed by clinical experts with ≥5 years of experience. These were analyzed for pooled prevalence and diagnostic relevance. Decisions on inclusion as diagnostic features were informed by both prevalence in the literature and expert judgment regarding specificity. Features with pooled prevalence >50% were considered for major criteria; those <50% for minor. However, non-specific high-prevalence findings (eg musculoskeletal pain) were excluded or downgraded based on their diagnostic utility. No objective thresholds or scoring tools were used. Decisions were based on pooled proportions and group consensus	No	No	No
Brunner, 2008	Complex regional pain syndrome type I	Yes	New	Two-round Delphi survey involving 13 international experts from five countries. In round one, experts listed diagnostic and follow-up parameters for CRPS I; in round two, they rated each parameter on a scale from 1 to 10. Parameters with a median score ≥7 and interquartile range ≤3 were considered both important and agreed upon	No	No	No
Carruthers, 2011	Chronic fatigue syndrome	No	New	International Consensus Panel composed of clinicians, researchers, educators, and a patient advocate from 13 countries developed the criteria using a Delphi-type process. Criteria were developed based on expert opinion with reference made to literature, however the specific role of evidence in informing decision-making is not described in detail. The final criteria reflect symptom clusters grounded in neuroimmune dysfunction with operational notes provided to clarify application and threshold severity	No	Yes- patient advocate was included as a member of the expert panel	No
Castro-Calvo, 2021	Gaming disorder	No	Revised	Three-round Delphi study conducted involving 29 international experts in clinical and research fields related to gaming disorder. Experts selected based on experience in treatment and publication history. Experts rated the diagnostic validity, clinical utility, and prognostic value of the DSM-5 and ICD-11 criteria, and suggested additional candidate criteria. Agreement was defined as ≥80% rating a criterion as "very important" or "extremely important," and the process included structured feedback between rounds to facilitate convergence of expert opinion	No	No	No
Cederholm, 2018	Malnutrition	No	New	The Global Leadership Initiative on Malnutrition (GLIM) convened a Core Leadership Committee and a Working Group with global representation. Diagnostic criteria were developed through a structured consensus process involving literature review, surveys, face-to-face meetings, teleconferences, and email communication. The specific role of evidence in informing decision-making was not described	No	No	No
Cha, 2020	Mal de debarquement syndrome	No	New	Bárány Society Classification Committee convened a multidisciplinary subcommittee of international experts in vestibular science. Draft criteria were developed through a series of meetings, discussions, and revisions. Decisions about feature inclusion were based on review of clinical reports, comparison with related disorders, expert consensus, and evaluation of symptom specificity. Criteria were refined through iterative feedback and harmonization with the International Classification of Vestibular Disorders (ICVD)	No	No	No
Cha, 2021	Motion sickness	Yes	New	Subcommittee of the Bárány Society Classification Committee was convened with international experts from multiple specialties. Criteria were developed through multiple meetings, literature review, email communications, and revisions based on internal discussions and feedback from Bárány Society members. Symptom domains were selected based on cross-referencing with the most cited motion sickness questionnaires and expert consensus	No	No	Yes 1) motion sickness disorder, 2) probable motion sickness disorder
Chafetz, 2020	Factitious disorder	No	New	The authors developed proposed diagnostic criteria for Factitious Disorder based on extensive comparison with malingering literature, emphasising the use of psychometric methods. They defined a tiered evidentiary approach modelled after frameworks for malingered neurocognitive dysfunction, integrating objective indicators with clinical context. Decisions about feature inclusion were based on	No	No	Yes 1) definite (compelling) evidence of factitious disorder, 2) probable

				previous heuristics used in malingering research, but tailored to the unique features of factitious presentations in clinical settings			evidence of factitious disorder and 3) possible evidence of factitious disorder
Chandler, 2003	Cohen syndrome	No	Revised	33 patients from 22 UK-based families were clinically assessed through detailed history, physical and ophthalmologic examination and neuropsychological testing. The authors reviewed the practicality of the previous Finnish diagnostic criteria to non-Finnish patients and proposed modified criteria based on the prevalence of different features in the study cohort	Not applicable	No	No
Chapman, 2010	Alzheimer's disease	No	New	Criteria for early-stage Alzheimer's disease developed using a two-step multivariate approach: principal components analysis (PCA) followed by discriminant analysis. PCA was applied to scores from 49 neuropsychological test measures to extract 13 cognitive components. These components were then used to build a discriminant function to classify participants as having AD or being cognitively normal. Decisions on inclusion of features were based on their statistical contribution (via stepwise selection) to discrimination accuracy in the model	Not applicable	No	No
Choon, 2024	Generalised pustular psoriasis	No	New	The International Psoriasis Council convened a Pustular Psoriasis Working Group of global experts who conducted a modified Delphi process. Expert panelists reviewed 64 challenging cases to generate 43 diagnostic statements which were refined through two rounds of virtual consensus meetings. Statements achieving ≥80% agreement informed the final definition and diagnostic criteria, with features selected through iterative voting and structured expert feedback	No	No	No
Ciron, 2022	Secondary progressive multiple sclerosis	No	New	Two-stage consensus development approach in France: eight regional board meetings involving 56 MS neurologists, followed by a national board of 13 experts (with representation of each geographical region). Potential diagnostic features were listed prior to discussion, however the role of evidence in informing decision-making was not described in detail. Diagnostic criteria were classified as "retained," "case-by-case," or "rejected" based on consensus	No	No	No
Coccaro, 2011	Intermittent explosive disorder	No	Revised	The study integrated existing DSM-IV and prior research-based criteria for IED to form an "Integrated Research" (IED-IR) criteria set. Diagnostic determinations were made using structured interviews and a best-estimate consensus method involving at least two psychiatrists and three clinical psychologists. Features were included if they improved sensitivity and construct validity for impulsive aggression, with thresholds such as frequency and severity used to operationalize inclusion criteria	Not applicable	No	No
Collins, 2023	Cubital tunnel syndrome	No	New	Delphi process involving 12 hand and upper-extremity surgeons. Recent literature was reviewed by authors to inform selection of panellists. Potential diagnostic features were identified using literature review and textbook chapters. Panellists rated the clinical importance of 55 diagnostic items on a scale from 1 to 10. The six most highly rated features were identified as consensus diagnostic criteria	No	No	No
Cummings, 2020	Psychosis in major and mild neurocognitive disorders	Yes	Revised	Multi-stage consensus process that included two global surveys of clinicians and interest holders, a literature review, and an in-person international consensus meeting with representatives from academia, clinical practice, regulatory agencies, advocacy groups, and industry. Feedback from these steps informed the development of revised diagnostic criteria. The final draft was reviewed and revised in an iterative manner, leading to a consensus definition	No	Yes- advocates present in consensus meetings	No
Decq, 2021	Sport-related concussion	Yes	New	Expert working group convened by the French Ministry of Sports with representation from national societies in neurotrauma, neurology, rehabilitation, and sports medicine. The group reviewed existing definitions of concussion and mild traumatic brain injury (mTBI) from international consensus statements, the American Congress of Rehabilitation Medicine and WHO guidelines. The group synthesised these definitions and their clinical elements to create operational diagnostic criteria applicable in sporting contexts. The final criteria were based on structured discussion and integration of practical field observations	No	No	No
DelRosso, 2020	Restless sleep disorder	No	New	The International Restless Legs Syndrome Study Group (IRLSSG) convened a task force of 10 international paediatric sleep experts. The group developed a list of 16 consensus questions before	No	No	No

				performing a literature review to address these questions. Criteria were subsequently developed based on the review findings (published separately) and group consensus			
Di Nardo, 2024	Irritable bowel syndrome	No	Revised	Delphi consensus process involving experts from four Italian pediatric and gastroenterology societies. The panel identified 22 key clinical questions and used the PICO framework to guide systematic literature reviews. Statements were rated using the GRADE approach, and consensus was defined as ≥80% agreement during one or two rounds of blinded voting. External reviewers also validated the final set of recommendations	Yes- quality of evidence and strength of recommendations assessed using GRADE approach	No	No
Doshi, 2021	Acute neuropathic pain	Yes	New	Working group of experts convened under the ACTTION-APS-AAPM collaboration reviewed existing literature and arranged webinars covering the AAAPT framework and prior chronic pain models. Consensus was developed through group meetings and iterative communication to identify core features of acute neuropathic pain and specific examples, though no objective definitions of consensus were used	No	No	Yes 1) definite acute neuropathic pain, 2) probable acute neuropathic pain and 3) possible acute neuropathic pain
Drossman, 2006	Functional gastrointestinal disorders	No	Revised	Multi-step consensus process involving 14 subcommittees composed of international experts. Chairs and co-chairs coordinated literature synthesis and iterative draft revisions. Consensus was achieved via face-to-face meetings, iterative reviews, and Delphi methodology to update diagnostic criteria and classification frameworks based on new evidence	No	No	No
Edmond, 2023	Prescription opioid dependence syndrome	Yes	New	Three-round Delphi study with 38 expert participants in pain, addiction medicine, and opioid use disorder (OUD). Participants recruited from a Veterans Health Administration State-of-the-Art (SOTA) conference. In round 1, qualitative responses were analysed to identify themes and potential diagnostic criteria. In rounds 2 and 3, participants rated the criteria using Likert scales to reach consensus (defined as a median score >5), and a follow-up expert workgroup refined the final criteria	No	No	No
Emre, 2007	Dementia associated with Parkinson's disease	Yes	New	Movement Disorder Society Task Force of 23 experts was convened. Members were selected based on expertise in epidemiology, clinical neurology, neuropsychology, geriatric psychiatry, and pathology. The group was divided into five subcommittees, each tasked with conducting systematic literature reviews based on pre-defined inclusion/exclusion criteria in specific areas (eg epidemiology, cognitive and neuropsychiatric features, motor and other clinical features). Diagnostic features were synthesised from the literature and consensus-based criteria for probable and possible PD-D were derived based on group discussion	No	No	Yes 1) probable PD-D and 2) possible PD-D
Engelman, 2020	Scabies	No	New	Modified Delphi study involving 34 international experts. Feedback from the panel was combined with a literature review as well as the opinion of an expanded group of experts to create definitions and criteria. Consensus was determined by iterative rounds of expert input and used to finalize three levels of diagnostic certainty	No	No	Yes 1) confirmed scabies, 2) clinical scabies and 3) suspected scabies
Esteba-Castillo, 2022	Mild cognitive impairment in Down syndrome	No	New	Longitudinal design over three years, assessing two groups of adults with Down syndrome (control and prodromal). Cognitive and behavioral data were collected at three time points. Linear and generalized linear mixed models were used to identify predictors of prodromal decline. Diagnostic features were selected based on statistically significant differences between groups	Not applicable	No	Yes 1) suspicion of MCI, 2) uncertain diagnosis and 3) no suspicion of MCI
Freeman, 2019	Peripheral neuropathic pain	No	New	A multidisciplinary international working group was convened to develop evidence-based diagnostic criteria for four focal or segmental peripheral neuropathic pain disorders. Experts conducted narrative reviews for their assigned condition, presented findings using the AAPT multidimensional taxonomy framework, and refined the criteria through iterative in-person and virtual discussions. Decisions on inclusion of features were based on symptom prevalence, specificity, and compatibility with established neuropathic pain grading schemes	No	No	No
Frota, 2011	Alzheimer's disease	No	Revised	Recommendations were developed by a working group of the Scientific Department of Cognitive Neurology and Aging of the Brazilian Academy of Neurology. The group reviewed international	No	No	Yes 1) definite AD, 2) probable

				consensus criteria (eg DSM-IV, NINCDS-ADRDA, NIA-AA) and incorporated emerging concepts like preclinical AD based on group consensus. Whilst reference is made to literature, the specific role of evidence in informing decision-making is not described in detail			AD and 3) possible AD
Giacino, 2002	Minimally conscious state	No	New	Nine consensus meetings were held between 1995 and 2000 by the Aspen Neurobehavioral Conference Workgroup, which included interdisciplinary experts in neurology, neuropsychology, rehabilitation, and ethics. Selected group members performed independent MEDLINE searches, however only five relevant reports were identified which was deemed insufficient evidence from which to develop guidance. The group subsequently developed diagnostic criteria through group consensus	No	No	No
Godefroy, 2010	Dysexecutive syndrome	No	New	GREFEX group developed separate diagnostic criteria for behavioural and cognitive dysexecutive syndromes through a consensus conference. An initial literature review was conducted to identify potential diagnostic features. A battery of standardized neuropsychological tests and a behavioural inventory was designed, validated in 718 healthy controls, and tested in 461 patients with diverse neurological conditions. Criteria were operationalized using specific performance thresholds (≤5th percentile in ≥3 domains), and their validity was assessed based on ability to discriminate between patients and controls and their association with loss of autonomy	Yes- level of evidence was evaluated using a novel tool to stratify potential diagnostic features as either "highly suggestive" or "supportive." "Highly suggestive" features were defined as impairment demonstrated in at least 2 studies showing a significant relation between the impairment and the lesion of the frontal subcortical network, and "supportive" deficits as impairment demonstrated in a group (or subgroup) of patients compared to healthy controls or controversial results across studies or limited number of studies	No	No
Graham, 2006	Carpal tunnel syndrome	No	New	Authors developed diagnostic criteria in three phases: item generation (via literature review, key informant interviews and focus groups), item reduction (via Delphi consensus among experts), and statistical modelling. Top 8 criteria were used to construct 256 case vignettes evaluated by expert panels. A logistic regression model based on 6 key criteria was developed and validated by comparing predicted probabilities with expert ratings on a visual analogue scale	No	No	No
Harden, 2007	Complex regional pain syndrome	Yes	Revised	International consensus workshop involving 35 pain experts from 7 countries. The group endorsed statistically derived revisions of the IASP diagnostic criteria based on factor analysis of symptoms in 123 complex regional pain syndrome patients	No	No	No
Healy, 2022	Enduring sexual dysfunction after treatment with antidepressants,	Yes	New	Criteria were developed from two large case series and additional patient reports submitted via the RxISK global pharmacovigilance platform. A multidisciplinary panel of experts refined the criteria through an iterative consensus using the published data as well as their own clinical experience. Agreement was reached through repeated circulation and revision of draft criteria	No	No	No

	finasteride and isotretinoin						
Hertzman, 2001	Eosinophilia-myalgia syndrome	No	New	Multi-stage, pattern-based approach including informal testing and formal validation. The coordinating committee constructed a set of preliminary criteria through consensus methodology before informally testing them by applying them to a set of case scenarios. An external panel of experts then generated and validated a gold standard set of EMS and non-EMS cases. The final criteria were tested against this set for accuracy, with elements defined in a glossary and interobserver variability assessed to ensure reproducibility	No	No	No
Höglinger, 2017	Progressive supranuclear palsy	No	Revised	Criteria were developed through a three-step process involving systematic literature review, analysis of a large autopsy-confirmed clinical dataset from nine international brain banks, and expert consensus. A modified Delphi method was used including multiple rounds of written feedback, a 2-day consensus meeting, and further refinements. Feature inclusion was guided by review findings, data from clinicopathological correlation and group consensus with clinical features categorised by diagnostic weight and stratified into four core functional domains	Yes- authors report evaluating the evidence in accordance with the Scottish Intercollegiate Guidelines Network (SIGN) guidance; however results of evidence appraisal are not provided	No	Yes 1) definitive PSP, 2) probable PSP, 3) possible PSP and 4) suggestive of PSP
Hoyme, 2016	Fetal alcohol spectrum disorder	Yes	Revised	1500 children who were prenatally exposed to alcohol were identified through active case-ascertainment methods. The children were assessed using standardised multidisciplinary evaluations including dysmorphology exams, neuropsychological testing, and structured maternal interviews. The authors formulated more precise clarifications of the previous diagnostic criteria based on the data obtained from the study cohort. Diagnostic revisions were informed by detailed analysis of 164 cases. Diagnostic categories were confirmed via case conferences, ensuring consensus when initial assessments diverged	No	No	No
Imazio, 2013	Post-pericardiotomy syndrome	Yes	New	Authors proposed new diagnostic criteria for PPS based on a synthesis of findings from the largest published randomized controlled trials investigating pharmacological prevention strategies. The specific role of the evidence in informing decision-making was not described in detail	No	No	No
Jacobs, 2000	Traumatic grief	Yes	New	Panel of bereavement, trauma, and psychiatric nosology experts convened to develop diagnostic criteria literature review. Based on narrative review and expert opinion, the group identified a distinct symptom cluster shown to predict impairment and morbidity	No	No	No
Jagasia, 2015	Chronic graft-versus-host disease	Yes	Revised	International consensus process conducted by a multidisciplinary NIH working group. Members were divided into organ-specific subgroups which reviewed evidence published since the 2005 consensus to address controversies and unanswered questions. Recommendations were developed collaboratively through multiple discussions and consensus meetings. Decisions regarding inclusion of diagnostic features were based on whether features were considered diagnostic, distinctive, common, or non-specific, with clarified rules for organ scoring and attribution of abnormalities to chronic GVHD. The updated criteria aimed to improve specificity and prognostic value	No	Yes- Meredith Cowden GVHD Foundation and various patient advocacy groups were recognized for their engagement and contributions during the planning and consensus development phases	No
Jensen, 2018	Malnutrition	No	New	Structured international consensus process involving leaders and representatives from global clinical nutrition societies (ASPEN, ESPEN, FELANPE, PENSA). The process included multiple face-to-face meetings, telephone conferences, and formal ballots. Diagnostic criteria were selected based on consensus agreement and empirical support from previous malnutrition screening and assessment tools. A two-step approach was adopted involving initial risk screening followed by diagnosis using at least one phenotypic and one etiologic criterion. Severity thresholds for diagnosis were established based on existing data and expert judgement	No	No	No

Kalra, 2014	Neuro-Behçet's disease	No	New	Four-round Delphi process was conducted involving 52 international experts across multiple disciplines. Experts were recruited based on clinical and academic expertise in Behçet's disease, with representation from 22 countries. Draft recommendations and diagnostic criteria were iteratively refined and rated using a 9-point Likert scale. Consensus was defined as ≥75% of respondents rating agreement ≥7	Yes	Yes- patient representative included in the advisory group	Yes 1) definite Neuro-Behçet's disease and 2) probable Neuro-Behçet's disease
Katz, 2021	Traumatic encephalopathy syndrome	Yes	Revised	Modified Delphi process conducted involving 20 clinician-scientists from multiple disciplines who participated in four rounds of voting following a consensus workshop. Experts reviewed literature on confirmed chronic traumatic encephalopathy cases and unpublished clinicopathologic data from the UNITE study. Inclusion of features in the criteria was based on evidence of frequency in pathologically confirmed cases, predictive validity data, and expert consensus, with an 80% agreement threshold set for consensus on each criterion	No	Yes- representatives of patient/family advocacy groups included in Delphi process	No
Khan, 2023	Hypophosphatasia	No	New	An International Working Group (IWG) was formed consisting of experts in HPP from Europe and North America, selected based on clinical expertise and publication history. Methodologists from McMaster University conducted systematic reviews using the GRADE approach. The criteria were developed through review of literature (including expert opinions, case series, and observational studies) and synthesis of features commonly present in HPP. Criteria were designated as major if present in >50% of cases or recommended by >50% of studies; otherwise, they were considered minor. Consensus was reached via virtual meetings	Yes- GRADE approach was used to evaluate the certainty of evidence	No	No
Kim, 2019	Haemodynamic orthostatic dizziness/vertigo	No	New	The Barany Society's Committee for the Classification of Vestibular Disorders formed a working group that iteratively developed the diagnostic criteria through international expert discussion (mainly through electronic communications supplemented by several in-person meetings) over a two-year period. Criteria were refined to reflect both historical definitions and modern clinical usage	No	No	Yes 1) haemodynamic orthostatic dizziness/vertigo and 2) probable haemodynamic orthostatic dizziness/vertigo
King, 2005	Chronic fatigue syndrome	No	Revised	Study recruited three groups (CFS, major depressive disorder, and healthy controls) and compared them using symptom checklists, severity ratings, and standardized measures/scores. It aimed to empirically identify discriminatory symptoms, evaluate the use of symptom severity thresholds, and test the utility of standardized tools to enhance the specificity and reliability of chronic fatigue syndrome diagnosis. Features were evaluated based on their frequency, severity, and ability to differentiate CFS from depression and healthy controls	Not applicable	No	No
Klein Meuleman, 2023	Caesarean scar disorder	Yes	New	Modified electronic Delphi study was conducted involving 31 international experts. A systematic literature review and a focus group with Dutch experts informed the development of an initial questionnaire covering nomenclature, symptoms, diagnostic criteria, and exclusion conditions. Over three Delphi rounds, experts rated statements and reviewed anonymized feedback, with consensus defined as a rate of agreement (RoA) ≥70%	No	Yes- patient focus group was used to inform initial questions, and patient representatives participated in reviewing the final consensus statements	No
Kollensperger, 2008	Multiple system atrophy-parkinsonian subtype (MSA-P)	No	Revised	European MSA Study Group (EMSA-SG) developed a standardized Red Flag Checklist (RFCL) based on literature review and expert opinion. This was applied to patients with probable MSA-P and Parkinson's disease across 15 centres. Red flags with >95% specificity were analyzed using factor analysis, and six red flag categories were identified. A diagnosis of probable MSA-P was proposed if at least one red flag was present in two or more of the six categories	No	No	No

Lamotte, 2022	Multiple system atrophy	No	Revised	The Movement Disorder Society MSA Criteria Revision Task Force performed a systematic review to identify possible diagnostic features before conducting a two-round Delphi process, a survey of society members, and a virtual consensus meeting to finalise the diagnostic criteria	No	No	Yes 1) neuropathologically established MSA, 2) clinically established MSA, 3) clinically probable MSA and 4) possible prodromal MSA
Landgraf, 2013	Fetal alcohol spectrum disorder	Yes	New	Multidisciplinary guideline group conducted systematic literature review and assessed 178 articles for evidence. Criteria were developed using Oxford Centre for Evidence-based Medicine levels, and additional recommendations were made through formal expert consensus (nominal group process with external moderation). Recommendations were based on evidence strength, clinical relevance, external validity, and ethical considerations	Yes- study quality formally assessed (specific tool/approach not specified) and strength of evidence determined using the Oxford Centre for Evidence-based Medicine's classification for diagnostic studies (2008)	Yes- representatives of patient organisations included in consensus group	No
Lebedeva, 2022	Acute headache attributed to ischaemic stroke and sentinel headache before ischemic stroke	Yes	Revised	Prospective case-control study design involving 550 patients with first-ever ischemic stroke and 192 matched controls. Structured face-to-face interviews were conducted using standardised forms to collect data on headache occurrence and characteristics before and at stroke onset. Headaches were categorized as new, altered, or unchanged. These data were used to test existing ICHD-3 criteria and derive revised or new criteria. Criteria development was based on presence in stroke patients but absence in controls, with sensitivity and specificity calculations performed	Not applicable	No	No
Lehman, 2008	Schinzel–Giedion syndrome	No	New	Review of 46 published cases performed and frequencies of craniofacial, skeletal, visceral, and neurological abnormalities calculated. Criteria formulated based on prevalence and specificity of certain clinical signs as well as author opinion though no objective thresholds used	No	No	No
Lempert, 2021	Vestibular migraine	No	Revised	Diagnostic criteria were developed jointly by the Bárány Society and the International Headache Society (IHS) with revisions based on literature updates and clinical consensus. Experts from multiple continents and specialties participated in an iterative development process	No	No	Yes 1) vestibular migraine and 2) probable vestibular migraine
Leong, 2023	Vocal cord dysfunction	No	New	Two-round modified Delphi process was conducted with 47 international experts identified through literature review and professional networks. In round 1, experts submitted qualitative statements about VCD diagnosis; these were categorised and rated in round 2 using a 5-point Likert scale with consensus defined as ≥70% agreement	No	Yes- two patient representatives reviewed the manuscript and provided feedback which was incorporated into the final version	No
Lin, 2016	Smartphone addiction	No	New	Authors developed 12 candidate symptom criteria and 4 functional impairment criteria based on prior frameworks (Internet addiction, Internet gaming disorder in DSM-5). Four psychiatrists conducted structured diagnostic interviews with 281 participants. Criteria were refined by analysing sensitivity, specificity, and diagnostic accuracy in comparison to clinical global impression (CGI)	Not applicable	No	No

Linden, 2008	Posttraumatic embitterment disorder	Yes	New	Diagnostic criteria and interview were developed through a comparative study of 50 clinically diagnosed PTED patients and 50 control patients with other psychiatric disorders. A semi-standardised interview was conducted, and data were analysed to identify symptom patterns, refine diagnostic descriptions and develop a diagnostic algorithm. Thresholds within the final instrument were informed by empirical sensitivity and specificity analysis	No	No	No
Litvan, 2012	Mild cognitive impairment in Parkinson's disease	No	New	Movement Disorder Society Task Force developed the diagnostic criteria through a literature review and expert consensus. Experts participated in two in-person meetings and iterative manuscript reviews over several months. Decisions regarding which features to include were based on both empirical findings and practical considerations such as feasibility in clinical and research settings	No	No	Yes 1) Parkinson's disease mild cognitive impairment and 2) possible Parkinson's disease mild cognitive impairment
Lopez-Escamez, 2015	Meniere's disease	No	Revised	Structured international consensus process overseen by the Classification Committee of the Bárány Society. Collaboration involved multiple societies and was based on iterative discussion, presentation, and refinement over a four-year period. Decisions regarding inclusion of features were informed by best available evidence, expert consensus, and aimed at improving clarity and clinical utility	No	No	Yes 1) definite Meniere's disease and 2) probable Meniere's disease
Lopez, 2022	Sleep-related dissociative disorders	No	New	Authors reviewed 26 cases (8 new and 18 previously published) to identify key clinical and video-polysomnographic features of SRDD. Diagnostic criteria were developed based on features present across cases including long episode duration, timing relative to sleep, self-harm, and presence of dissociative symptoms. Criteria were finalised based on author opinion	No	No	Yes 1) definite SRDD, 2) probable SRDD and 3) possible SRDD
Markman, 2020	Chronic low back pain	No	New	A multidisciplinary working group convened through the AAPT initiative developed the criteria. Prior to meeting, group members were provided with templates and background papers from the AAPT initiative. Several group members then performed their own literature reviews. Experts in systematic review methodology and clinical epidemiology were included in the group. Evidence and group member opinions were subsequently discussed during a series of teleconferences and email exchanges	No	No	No
Matias-Guiu, 2022	Cognitive dysfunction in post-COVID condition	Yes	New	The study adapted the IC-CoDE framework, originally used in epilepsy, to develop the IC-CoDi-COVID criteria. 404 patients with post-COVID condition were evaluated with neuropsychological batteries. Cognitive impairment was defined based on performance in five cognitive domains using thresholds of 1 and 1.5 standard deviations below the normative mean. Impairment in each domain required both representative tests to fall below the threshold. Machine learning (hierarchical clustering and random forest classification) was used to validate domain impairment thresholds and phenotypic classification	Not applicable	Yes- patient groups such as LongCovid ACTS and Colectivo COVID-19 persistent Madrid were acknowledged for their collaboration. However, their role in development or review of the diagnostic criteria was not described in detail	No
McKeith, 2017	Dementia with Lewy Bodies	No	Revised	A revised consensus statement was developed through the DLB Consortium incorporating the findings of four expert working groups. These groups reviewed evidence and formulated updates through structured meetings involving multidisciplinary specialists. Patients and care partners also participated. Features were categorised as core clinical features or supportive features based on diagnostic specificity	No	Yes- patient and care partner participation was included during the consensus	Yes 1) probable DLB and 2) possible DLB

						development meeting	
Mease, 2014	Inflammatory musculoskeletal conditions	No	New	The GRAPPA group initiated a multi-step project to develop simple clinical criteria for inflammatory MSK conditions. The first step involved nominal group exercises at the 2013 GRAPPA meeting, where expert clinicians and patient researchers identified features distinguishing inflammatory from noninflammatory disease. This was followed by planned Delphi exercises, harmonisation of language, application of draft criteria to test patients, and validation across larger cohorts	No	Yes- patient advocate was included as a member of the expert panel	No
Metry, 2024	LUMBAR syndrome	No	New	A project steering team conducted a literature review of cases involving segmental infantile hemangiomas with associated anomalies. The steering team drafted preliminary diagnostic criteria, which were then evaluated by a panel of 61 international experts using a two-round modified electronic Delphi method. Features were included based on expert consensus using a threshold of ≥80% agreement. Terminology and criteria were then refined iteratively based on panel feedback	No	No	No
Michel, 2013	Piriformis muscle syndrome	No	New	Prospective study of 250 patients with suspected piriformis muscle syndrome (PMS) involving development of a 12-point clinical scoring system based on standardized physical examination findings and symptom characteristics. Criteria were included based on their specificity for PMS, with scoring informed by comparison to 60 control subjects (30 with disco-radicular conflict, 30 healthy). A score ≥8 was considered 'probable' PMS based on statistical validation showing high sensitivity and specificity	Not applicable	No	Yes 1) probable PMS, 2) unlikely PMS and 3) PMS not considered
Michelotti, 2016	Multiple temporomandibular disorders	No	Revised	Workshop held by the International RDC/TMD Consortium Network, which convened six expert-led workgroups to discuss improvements to the DC/TMD taxonomy. Decisions regarding feature inclusion were based on existing evidence, clinical challenges, and consensus within specialty-focused groups, although explicit thresholds/definitions of consensus were not used. The workgroups proposed refinements and research recommendations to improve validity, operationalization, and clinical utility of diagnostic criteria for various TMD subtypes	No	No	Yes
Miller, 2021	Apathy in neurocognitive disorders	No	Revised	The International Society for CNS Clinical Trials Methodology (ISCTM) Apathy Work Group conducted a consensus development process using a five-step modified Delphi methodology. This included a literature review, expert surveys, and multiple international meetings. Criteria were iteratively refined based on expert input, with inclusion decisions guided by evidence from literature and structured consensus surveys (≥80% agreement threshold). The process involved input from academia, industry, and regulatory bodies	No	No	No
Moore, 2014	Alzheimer disease (AD), mild cognitive impairment (MCI), and vascular cognitive impairment	No	Revised	Criteria were developed during the Fourth Canadian Consensus Conference on the Diagnosis and Treatment of Dementia (CCCDTD4), involving a multidisciplinary committee of Canadian experts. Background papers and preliminary recommendations were developed by topic-based teams and reviewed at a national consensus conference. The GRADE system was used to assess the quality of evidence and strength of recommendations. Consensus was defined as 80% or more agreement among participants. Final recommendations were voted on both online and during the in-person meeting	Yes- GRADE approach was used to evaluate the strength of recommendations and levels of evidence	Yes – a memory clinic patient was invited to participate in the conference, though did not attend	Yes 1) probable Alzheimer's disease and 2) possible Alzheimer's disease
Morris, 2014	Alzheimer's disease	No	Revised	Authors convened an international group of experts involved in the development of prior diagnostic criteria (IWG and NIA-AA) at the 2012 Key Symposium in Stockholm to develop harmonized recommendations. Through comparison of existing criteria, open discussion, and a structured consensus approach, the group proposed unified terminology and criteria modifications	No	No	No
Müller, 2021	Compulsive buying-shopping disorder	No	New	Delphi consensus approach was used to develop criteria. An international panel of experts was recruited via a literature search and invited via email. The process included two iterative rounds of surveys where experts rated the importance of 37 initial and subsequently refined criteria grouped by domains. Items reaching ≥75% consensus among all experts and master experts were included. Expert responses were also analysed for certainty levels and qualitative feedback informed refinements	No	No	No

Mysliwiec, 2018	Trauma associated sleep disorder	Yes	New	Narrative review of literature regarding trauma survivors with parasomnia-like symptoms. Criteria proposed based on author interpretation of the literature (eg polysomnographic data and patterns of symptoms onset)	No	No	No
Neul, 2010	Rett syndrome	No	Revised	An expert consensus group (RettSearch Consortium) reviewed clinical data and evaluated the existing criteria against a large clinical cohort. Features were retained, modified, or excluded based on diagnostic clarity and consistency as deemed appropriate by the authors	No	No	Yes 1) atypical Rett syndrome, 2) probably atypical Rett syndrome
Ohayon, 2016	Restless legs syndrome	No	Revised	Large, cross-sectional, population-based survey was conducted across 14 US states using a telephone-based method with the Sleep-EVAL system to assess RLS symptoms, frequency, duration, and associated impairments in 19,136 adults. The system applied DSM-5 and IRLSSG criteria and excluded mimics using structured questions. Decisions regarding inclusion of features focused on how frequently and for how long symptoms occurred, and whether they caused clinically meaningful distress or impairment. Data analyses were weighted and included regression models to assess prevalence and comorbidities	Not applicable	No	No
Olin, 2002	Depression of Alzheimer's disease	Yes	New	Workshop involving 21 experts in dementia and late-life depression who were provided access to relevant review articles prior to meeting. Subset of members drafted criteria modeled on DSM-IV Major Depressive Episode criteria (modified to suit the Alzheimer's disease population). Draft criteria were then subject to eight rounds of feedback via emails and calls	No	No	No
Parker, 2020	Bipolar disorder	No	Revised	international Task Force was formed, with members invited based on prior publications in the field. Task force members met to discuss potential criteria before recruiting 178 patients to complete a structured questionnaire on manic/hypomanic symptoms and impairment. Patient-reported data were used to empirically derive revised definitions and symptom constructs. Symptom prevalence was analyzed and compared to a unipolar depression group to determine specificity, and factor analysis was used to identify symptom domains. Final criteria were devised based on questionnaire data and task force/author interpretation and opinion though no objective thresholds were used	Not applicable	No	No
Parker, 2022	Bipolar disorder	No	Revised	AREDOC taskforce recruited international experts in bipolar disorder to review and revise DSM-5 criteria. This included eliciting views on current DSM-5 criteria, developing provisional revisions, and conducting empirical analyses involving bipolar I, bipolar II, and unipolar depressed patients. Patients rated symptom checklists, and expert opinion plus empirical data informed item selection and threshold setting. Machine learning (prediction rule ensembles), factor analyses, and latent class analysis were used to refine symptom criteria and determine classification accuracy, sensitivity, and specificity for proposed diagnostic thresholds	No	Yes- patients were asked to review definitional statements	No
Parvaneh, 2016	Benign joint hypermobility	No	New	New criteria were developed based on a narrative review examining limitations of the previous criteria and pediatric range of motion norms by age and gender as well as the clinical experience of the authors. A case–control study including 200 children (100 cases and 100 matched controls) aged 3–16 was performed. Each child was evaluated using the standard Beighton criteria and the newly proposed criteria. Diagnostic accuracy of the new criteria was assessed using sensitivity, specificity, and ROC analysis	No	No	No
Patel, 2014	Mucous recirculation syndrome	No	New	Authors conducted literature review to identify published cases of mucous recirculation syndrome and supplemented these with four new patients from their own institutional database. Total of 12 patients were included who met defined inclusion criteria. Diagnostic criteria were formulated based on the prevalence/consistency of features across these patients and prior anatomical/surgical studies	No	No	No
Peralta, 2001	Catatonia	No	New	Study used a four-step method inspired by Pfohl and Andreasen's recommendations for psychiatric classification systems: (1) patient selection and symptom variable identification using the Modified Rogers Scale, (2) cluster analysis to separate patients into catatonic and noncatatonic groups, (3) development of diagnostic criteria through discriminant and ROC analyses, and (4) validation through assessment of psychometric and construct validity	Not applicable	No	No

Petzold, 2022	Optic neuritis	No	New	Panel of 101 international experts in neurology, ophthalmology, neuro-ophthalmology and neuroradiology developed criteria using an electronic Delphi process conducted over three years. Diagnostic features were selected based on literature review, expert consensus, and analysis of paraclinical findings. Clinical features were categorised and aligned with specific paraclinical thresholds, with a tiered diagnostic approach (ie definite vs possible optic neuritis)	No	No	Yes 1) definite optic neuritis and 2) possible optic neuritis
Picchietti, 2013	Restless legs syndrome	No	Revised	Panel of seven pediatric restless legs syndrome experts were appointed by the IRLSSG. Panel developed 15 consensus questions and reviewed 71 selected articles from a structured literature search. Issues were discussed in person and via correspondence, and consensus was achieved through iterative discussion. The updated criteria were endorsed by the IRLSSG Executive Committee and distributed for feedback among IRLSSG members	No	No	Yes 1) probable restless legs syndrome and 2) possible restless legs syndrome
Poole, 2019	Hand-arm vibration syndrome	Yes	Revised	Modified Delphi process was used involving 7 international experts in HAVS. Experts participated anonymously in multiple rounds to review the Stockholm Workshop Scale and suggest improvements. Consensus was defined as agreement from ≥72% (5/7) of experts. Decisions on diagnostic criteria were based on both existing literature and expert experience	No	No	No
Poretti, 2012	Oral-facial-digital syndrome type IV	No	New	16 patients with previously diagnosed OFD VI were included and their clinical data characterised. The authors proposed diagnostic criteria based on the consistent presence of the molar tooth sign combined with one or more other features, though no objective thresholds were used	No	No	No
Postuma, 2015	Parkinson's disease	No	Revised	Criteria developed by the MDS Task Force through a structured process involving two brainstorming teleconferences, an in-person meeting, and six months of iterative revisions. After an internal draft was finalized, "cognitive pretesting" was performed by neurologists not involved in the criteria development, who tested the criteria in real patient assessments and provided feedback. Final criteria were ratified by the task force	No	No	Yes 1) clinically established Parkinson's disease and 2) clinically probably Parkinson's disease
Pringsheim, 2023	Functional tic-like behaviours	No	New	A formal multi-round web-based Delphi consensus process was used following an in-person discussion at the 2022 ESSTS meeting. Experts in Tourette syndrome and FTLBs were invited to participate based on their level of clinical experience. Criteria were developed from clinical observation and literature review, and panelists rated items on a 9-point Likert scale. Items with a median score ≥7 were retained, and revisions were made based on comments to refine the criteria.	No	No	Yes 1) clinically definite and 2) clinically probable
Prost-Squarcioni, 2018	Epidermolysis bullosa acquisita	No	Revised	International Bullous Diseases Group (IBDG) convened three times in 2015 to discuss and vote on diagnostic criteria for EBA. 50 proposals were put forward for discussion and tables summarising the literature were constructed. A structured consensus methodology was used involving 22 experts from 14 countries who voted on 50 proposals related to diagnostic features. Consensus was defined as > 66% agreement. Discussions occurred when this threshold was not met, and re-voting was conducted. Decisions on criteria inclusion were based on both clinical relevance and supporting evidence, however the specific role of evidence in informing decision-making was not described in detail	No	No	Yes 1) definite EBA, 2) highly probable EBA and 3) possible EBA
Ravina, 2007	Parkinson's disease psychosis	No	New	Criteria developed by NIH-sponsored work group comprising clinical and research experts in Parkinson's disease and psychosis. Three subgroups conducted structured literature reviews on clinical features, epidemiology/pathophysiology, and differential diagnosis. Based on this evidence, the group drafted and ratified diagnostic criteria during a workshop, using consensus discussion rather than formal Delphi or voting procedures. Criteria were modeled after DSM-IV-TR standards and are intended to be provisional pending validation.	No	No	No
Read, 2001	Vogt-Koyanagi-Harada disease	No	Revised	Criteria were developed by a consensus committee convened at the First International Workshop on VKH Disease. Discussions among experts led to the formation of a draft, which was circulated among committee members for feedback and finalised based on agreement	No	No	Yes 1) complete VKH disease, 2) incomplete VKH disease and 3)

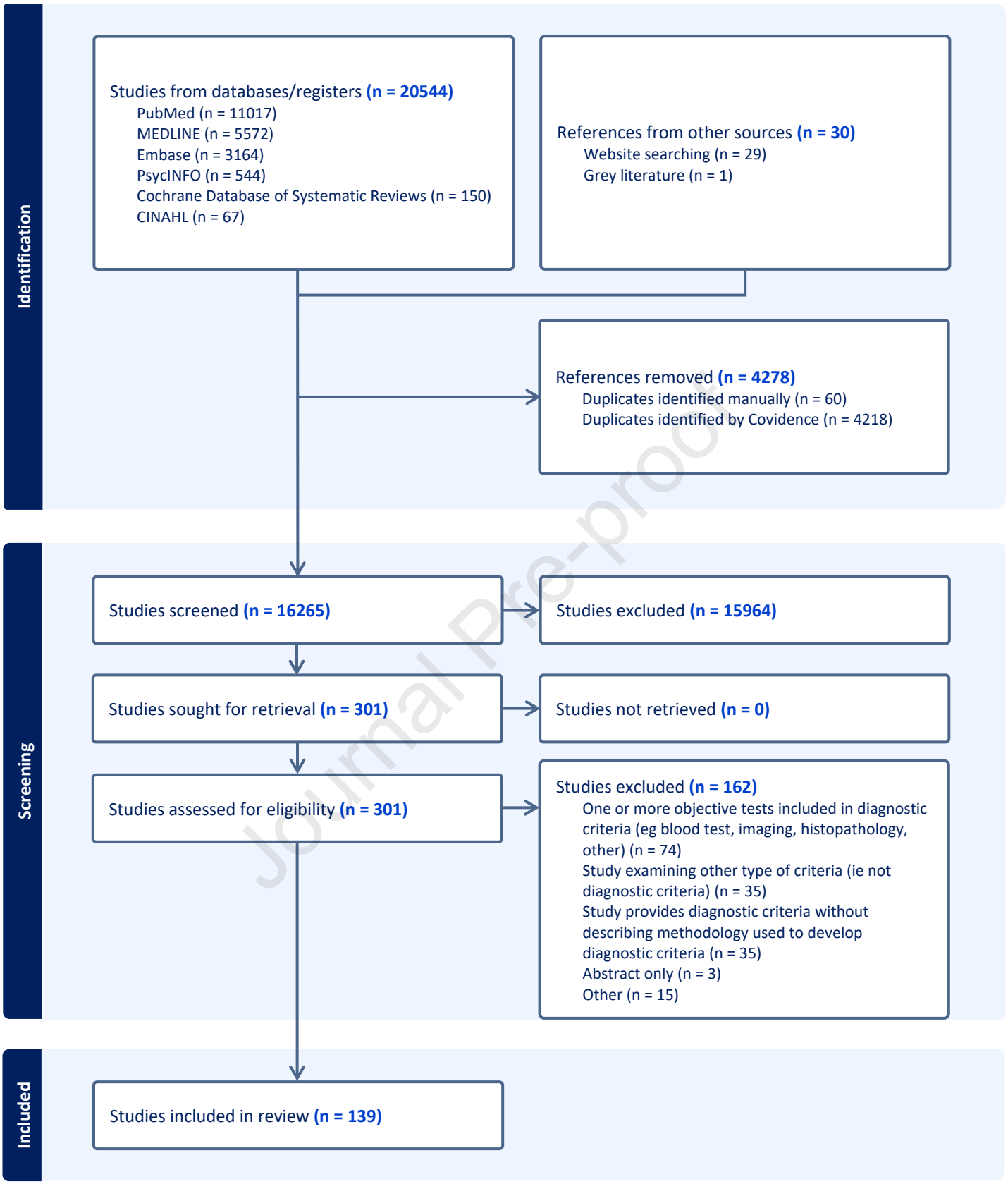
							probable VKH disease
Reid, 2024	Fetal alcohol spectrum disorder	Yes	Revised	The study adapted the GRADE approach to update Australia's diagnostic criteria for FASD, focusing on integrating systematic evidence into decision-making. Candidate diagnostic features were identified through systematic reviews and meta-analyses evaluating their association with prenatal alcohol exposure. Evidence-to-decision frameworks (EtDFs) were developed and modified to support decision-making about inclusion/exclusion of features. Features were assessed based on strength of association, diagnostic utility, certainty of evidence, feasibility, and other GRADE-informed domains	Yes- GRADE approach for prognostic factors was adapted	Yes – Consumer-advocacy organisations were part of the national consortium leading the update. Additionally, "lived experience statements" were developed based on a systematic review of qualitative evidence examining individuals' experiences with the diagnostic process	No
Reilmann, 2014	Huntington's disease	No	New	Development of diagnostic categories and criteria was informed by natural history studies (notably PREDICT-HD and TRACK-HD) and longitudinal clinical observation. Criteria were developed based on the authors' interpretation of the literature as well as their own clinical experience	No	No	No
Rinella, 2024	Metabolic dysfunction-associated steatotic liver disease	Yes	Revised	Multi-interest holder Delphi process was conducted under the leadership of major hepatology societies, involving experts across disciplines and sectors. The process spanned three years and included four Delphi surveys and two large in-person meetings. Consensus was defined a priori as ≥67% agreement, with features included in the criteria based on conceptual clarity, inclusivity, and clinical practicality, while aiming to avoid stigma and ensure continuity with historical NAFLD populations.	No	Yes- patient advocates involved in Delphi panel contributed to discussions on terminology and potential stigma	No
Rivers, 2015	Myofascial pain syndrome	No	New	International survey was conducted among clinician members of the International Association for the Study of Pain and the American Academy of Pain Medicine. Respondents evaluated a list of palpatory findings, signs, symptoms, and treatment responses, classifying each as "essential," "associated," "irrelevant," or "exclusionary." These data informed the development of a preliminary set of diagnostic criteria. Decisions regarding feature inclusion were based on thresholds of endorsement (eg >50% identifying a feature as essential or >80% as essential/associated)	No	No	No
Robert, 2009	Apathy in Alzheimer's disease	No	Revised	Task force composed of experts from the Association Française de Psychiatrie Biologique, the European Psychiatric Association, and the European Alzheimer's Disease Consortium was convened to revise previously proposed apathy criteria. Experts were selected based on neuropsychiatric expertise and engaged in iterative reviews of proposed drafts, culminating in a consensus meeting held during the 2008 EPA conference. Decisions regarding inclusion of diagnostic features focused on their observability, clinical relevance, and ability to distinguish apathy from similar conditions	No	No	No
Rosowsky, 2019	Late-onset personality disorder	No	New	Three-round Delphi consensus process involving 21 international experts in geriatric psychiatry, psychology, and personality disorders. Experts were asked structured questions about the validity of applying DSM-V personality disorder criteria to late-onset presentations and whether such cases could be diagnosed de novo in old age. Consensus was defined as >67%	No	No	No
Rothman, 2014	Circumferential skin folds syndrome	No	New	Literature review including 42 cases with characteristic phenotype. Reported associated anomalies were compiled, categorized and sorted by frequency. Author selected major and minor diagnostic criteria based on frequency across reported cases, though no objective thresholds/cutoffs were used	No	No	Yes 1) circumferential skin fold syndrome and 2)

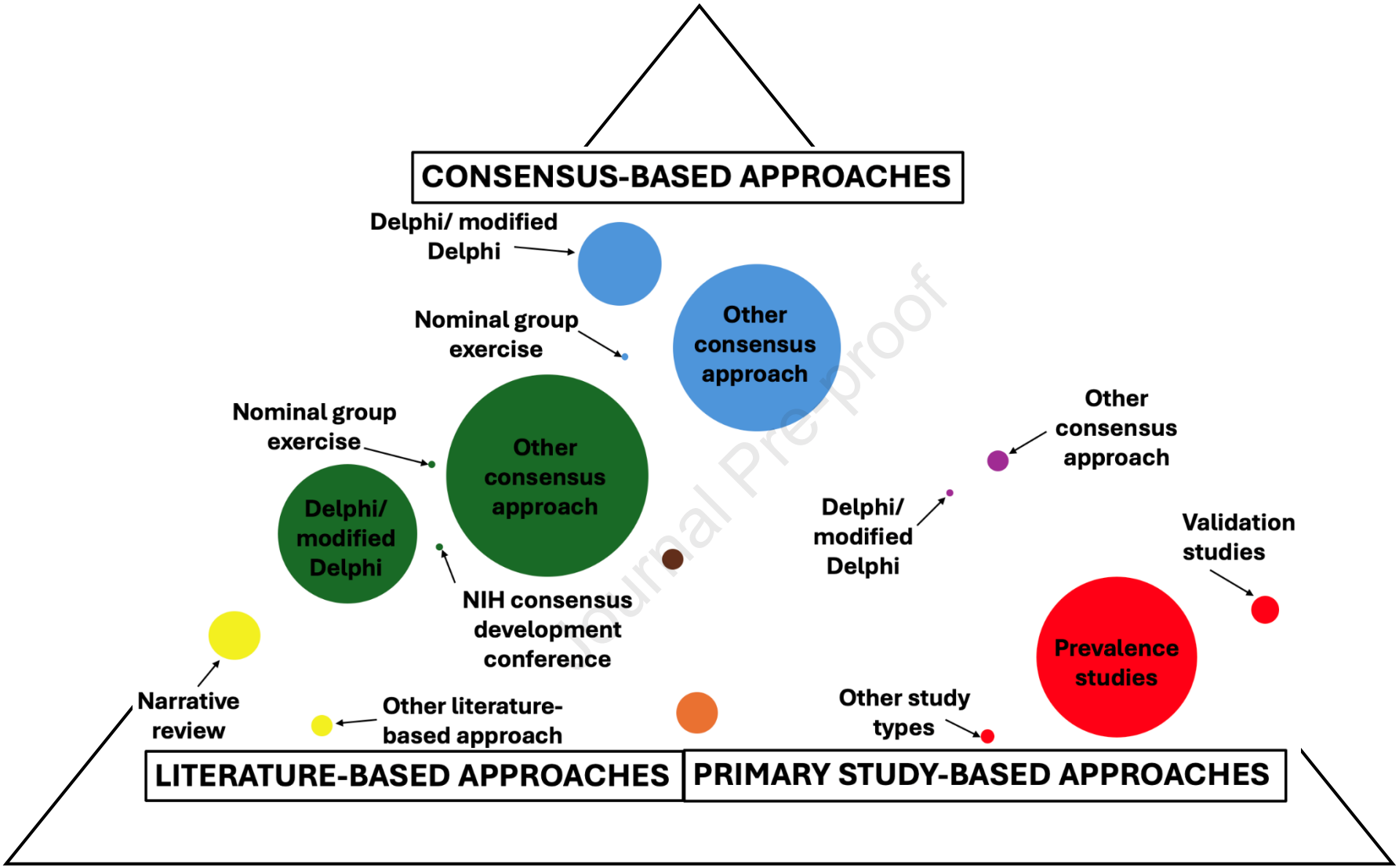
							possible circumferential skin fold syndrome
Rudnicka, 2015	Satoyoshi syndrome	No	New	Authors conducted a systematic review of 49 previously published cases of Satoyoshi syndrome, along with 3 new cases they reported. Clinical features, treatment responses and coexisting conditions were analysed. Diagnostic criteria were proposed based on the findings of this review	No	No	No
Sachdev, 2014	Vascular cognitive disorders	No	New	A symposium of the International Society for Vascular Behavioral and Cognitive Disorders (VASCOG) in 2009 initiated the development of the criteria. The draft was iteratively revised through expert feedback and further discussed at subsequent VASCOG meetings	No	No	Yes 1) probable VCD and 2) possible VCD
Sheldon, 2006	Vasovagal syncope	No	New	A structured 118-item questionnaire was administered to 418 patients with syncope and no structural heart disease. Diagnostic classification was based on comparison between patients with positive tilt tests (vasovagal syncope) and patients with known other causes of syncope. Logistic regression was used to identify predictive features and a point score system was developed. Thresholds for the score were validated using receiver operating characteristic analysis and adjusted for optimism using bootstrapping	Not applicable	No	No
Smith, 2021	Acute otitis externa	No	New	Three-stage process was used: 1) interviews with nine former patients to gather candidate outcomes, 2) systematic review of the literature to identify previously used definitions, diagnostic criteria and outcomes, and 3) a three-round modified Delphi process involving experts and former patients (patients rated only outcomes, not criteria). Items rated on a 1–9 scale were included if ≥70% scored them 7–9 and <15% scored them 1–3. The steering committee reviewed borderline items and grouped included features to form the final diagnostic criteria	No	Yes- nine former patients were interviewed to generate patient-centred outcomes. They also participated in the Delphi process (outcomes only), helping identify important domains like quality of life and impact on work, which were underrepresented in prior research	No
Staab, 2017	Persistent postural-perceptual dizziness	No	New	Criteria were developed by expert consensus through the Bárány Society's Committee for the Classification of Vestibular Disorders. The process included iterative draft revisions and feedback from international experts and scientific societies across otorhinolaryngology, neurology, psychiatry, and psychosomatic medicine. The committee reviewed 30 years of research on related syndromes and synthesized common features. Consensus was achieved through structured meetings and extensive consultation, with refinements based on feedback from the society and broader medical community	No	No	No
Strupp, 2016	Vestibular paroxysmia	No	Revised	Classification Committee of the Bárány Society conducted a four-year multidisciplinary internationally represented iterative consensus process incorporating presentations and discussions. Appraisal of evidence was performed; however minimal detail was provided with respect to how this informed decision-making	No	No	Yes 1) vestibular paroxysmia and 2) probable vestibular paroxysmia
Sumitani, 2010	Complex regional pain syndrome	Yes	New	Multi-centre between-subjects study comparing CRPS and non-CRPS patients in Japan. A standardised sign/symptom checklist (adapted from the CRPS database) was used to collect data from 195 CRPS and 146 non-CRPS patients. Factor analysis identified five symptom/sign clusters, which were then used to develop and test 18 decision rules via discriminant function analysis. The selection of features for inclusion in diagnostic criteria was based on statistical groupings and diagnostic performance metrics such as sensitivity and specificity	Not applicable	No	No

Tao, 2010	Internet addiction disorder	No	New	Study was conducted in three phases—development, validation, and clinical evaluation. Initially, eight symptoms were identified based on clinical experience and prior instruments and were then evaluated in 110 patients. A revised version excluding the least endorsed symptom was tested in 408 additional patients. A final validation phase involved 405 students and 150 clinic patients where the criteria were tested for inter-rater reliability and diagnostic accuracy. Criteria were finalised using a "2 + 1" rule (preoccupation and withdrawal symptoms plus one additional symptom), supported by functional impairment, duration, and exclusion criteria	Not applicable	No	No
Taylor, 2004	Psoriatic arthritis	No	Revised	Authors reviewed existing diagnostic criteria for psoriatic arthritis and highlighted the methodological limitations in their development such as lack of patient-derived data and inappropriate control groups. The study informed the design of the CASPAR (ClASsification criteria for Psoriatic ARthritis) project, which used prospectively recruited cases and matched controls from rheumatology clinics across multiple countries. Potential diagnostic features were selected using a Delphi approach and logistic regression and classification tree analysis were employed to derive data-driven criteria	No	No	No
Tinuper, 2016	Sleep-related hypermotor epilepsy	No	New	Formal consensus conference approach inspired by the NIH Consensus Development Program and adapted from the Italian National Guideline System. Experts from various disciplines participated in structured phases including systematic review, evidence mapping, and multi-group deliberations. Features were included in the diagnostic criteria based on available evidence, discussions among workgroups, and structured consensus procedures, culminating in the proposal of three levels of diagnostic certainty.	Yes- quality of evidence assessed using the Classification of Evidence Schemes of the Clinical Practice Guideline Process Manual of the American Academy of Neurology (2011)	No	Yes 1) confirmed SHE, 2) clinical SHE and 3) possible SHE
Vaivre-Douret, 2011	Developmental coordination disorder	No	New	Prospective, descriptive and analytical study of 43 school-aged children selected using DSM-IV-R criteria. Detailed standardized assessments were conducted in neuropsychological, neuropsychomotor, and neurovisual domains. Cluster analysis (hierarchical and k-means) was applied to clinical variables to identify subgroups with similar impairment profiles, leading to the classification of three subtypes	Not applicable	No	No
van de Berg, 2021	Multiple vestibular disorders	No	New	Structured consensus development process overseen by the Classification Committee of the Bárány Society in collaboration with the International Headache Society. Initial draft criteria were based on the best available scientific evidence and refined through discussions among international experts in pediatric neurology, otolaryngology, and related fields. Criteria were revised after feedback and re-evaluation until a final consensus was reached, with an emphasis on broad clinical practicality and clarity in pediatric settings.	No	No	Yes 1) vestibular migraine of childhood and 2) probable vestibular migraine of childhood
van de Merwe, 2008	Bladder pain syndrome	No	Revised	European Society for the Study of Interstitial Cystitis (ESSIC) conducted four meetings comprising experts from multiple disciplines. Experts reviewed prior definitions and proposed standardized methods for evaluating symptoms and exclusion of confusable diseases, which were finalised through consensus	No	No	No
van der Meche, 2001	Guillain-Barré syndrome	No	Revised	Dutch Neuromuscular Research Support Centre formed a panel of experts from all neuromuscular centers in the Netherlands. A consensus protocol was used involving plenary meetings and written feedback. Criteria were developed through review of recent literature, incorporation of previous classification systems, and consensus	No	No	Yes 1) definite GBS, 2) probable GBS and 3) possible GBS
van Santen, 2023	Hypothalamic syndrome	No	New	Authors developed a new set of diagnostic criteria based on existing literature and their own expert understanding of hypothalamic dysfunction. Whilst reference is made to the literature, the specific role of evidence in informing decision-making is not described in detail	No	No	No
van Veen, 2022	Irremediable psychiatric suffering in the context of medical	No	New	Two-round Delphi process was conducted among Dutch and Belgian psychiatrists with at least five years' experience and involvement in medical assistance in dying assessments. Criteria were developed from a systematic review and qualitative interviews. Experts rated 20 initial criteria across diagnostic, treatment, and treatment refusal domains using Likert scales, and consensus was defined as ≥70% agreement. Criteria not reaching consensus in round 1 were modified and reassessed in round 2, with	No	No	No

	assistance in dying			qualitative comments analysed thematically to inform revisions. Decisions on inclusion of features into the final criteria were driven by consensus thresholds and stability of responses			
Visser, 2008	Predementia Alzheimer's Disease	No	New	Prospective, multi-centre cohort study design involving 881 individuals without known dementia with cognitive complaints. Data were collected on a wide range of potential predictive variables. Diagnostic algorithms were developed using logistic regression and ROC curve analysis, with variables entered stepwise based on ease of collection. Candidate variables for inclusion were selected based on univariate tests and final models aimed to maximize predictive accuracy	No	No	Yes 1) probable Alzheimer's disease and 2) possible Alzheimer's disease
Vogels, 2022	Chronic exertional compartment syndrome	No	New	A three-round Delphi method was used to collect and refine expert opinion regarding diagnosis and management of CECS in civilian and military populations. Statements for the panel to consider were derived from an initial literature search identifying publications that described one or more of (1) history-taking questions that aided in the differential diagnosis of exercise-related leg pain, (2) physical examination and special tests, (3) indications for and types of diagnostic investigations and (4) treatment interventions. Consensus was defined as >70% agreement and results of previous rounds were anonymously disclosed	No	No	No
von Brevern, 2017	Benign paroxysmal positional vertigo	No	New	Criteria were developed by the Committee for Classification of Vestibular Disorders of the Bárány Society through an international consensus process. This included multiple rounds of feedback from five professional neuro-otology societies and the Bárány Society membership, with drafts discussed at two meetings (2010 and 2012). Final consensus was achieved via an online review system. Objective thresholds such as duration and nystagmus direction were applied to delineate subtypes	No	No	Yes 1) definitive BPPV, 2) probable BPPV and 3) possible BPPV
Watkins, 2012	Fetal alcohol spectrum disorder	Yes	Revised	Modified Delphi process was used involving 139 invited health professionals (130 Australian, 9 international) with experience in FASD diagnosis or screening. Participants completed a two-round online survey with 36 Likert-scale statements regarding components of various published diagnostic criteria. Consensus was defined as ≥70% agreement with a statement. Criteria for round inclusion/reassessment/rejection were pre-specified based on levels of agreement	No	No	No
Wenning, 2022	Multiple system atrophy	No	Revised	Movement Disorder Society (MDS) established a Criteria Revision Task Force. Experts were organized into thematic working groups to review literature and generate evidence-based questions. Two Delphi rounds, an international MDS member survey, and a virtual consensus conference were used to refine and finalize the criteria. Consensus was defined as ≥80% agreement, and inputs were incorporated through a transparent, iterative process involving expert discussions, survey feedback, and structured methodological oversight	Yes	No	No
Wetselaar, 2023	Tooth wear	No	New	Criteria were developed using a two-step process. This study reports the first step: a preliminary beta version based on the authors' clinical experience and a narrative literature review. A systematic literature search was conducted, identifying 383 relevant publications, which were categorised under four topics—nomenclature/taxonomies, self-report tools, clinical assessment tools, and clinical decision-making. Inclusion of features in the criteria was based on consistency with the literature and the authors' shared expertise	No	No	No
Widerström-Noga, 2017	Central neuropathic pain	No	New	Diagnostic criteria were developed by a working group with clinical and research expertise in central pain, convened through the ACTTION–APS Pain Taxonomy (AAPT) initiative. Group members conducted literature reviews and participated in in-person meetings, email discussions, and teleconferences. A consensus approach was used to develop criteria across five AAPT dimensions	No	No	No
Wilson, 2001	Major depressive disorder in chronic pain	Yes	Revised	The study compared three diagnostic approaches for MDD in patients with chronic pain: inclusive (counting all DSM-IV symptoms), aetiologic (excluding symptoms attributed solely to pain), and substitutive (replacing somatic symptoms with psychological alternatives). Semi-structured interviews with 129 patients were used, and symptoms were rated on standard severity thresholds. Decisions to include or exclude symptoms were based on patient-reported attributions of symptom causality	Not applicable	No	No
Yamani, 2019	Migraine with brainstem aura	No	Revised	Authors performed literature review and identified 79 cases, which were combined with clinical cases from the Danish Headache Center (DHC) and a large cohort of telephone-interviewed patients. Using	No	No	No

				these sources, the authors identified the 20 most clinically 'convincing' cases and compared them to broader diagnostic cohorts. New stricter criteria were proposed based on the prevalence of features in the most 'convincing' cases although no objective threshold was used. Consensus was based on clinical judgement by two neurologists			
Yang, 2022	Fuchs' uveitis syndrome	No	Revised	Clinical data from 970 patients with FUS and 1128 patients with non-FUS uveitis were retrospectively analysed using multivariate two-step cluster analysis, logistic regression, and decision tree algorithms. Variables were selected based on statistical significance and clinical relevance, with diffuse iris depigmentation, mild anterior chamber inflammation, and absence of posterior synechiae identified as essential. These variables were integrated into the final diagnostic criteria through expert clinical judgement and data-driven decision modeling	Not applicable	No	No
Yesavage, 2003	Sleep disturbance in Alzheimer's disease	No	New	A taskforce was formed an initial workshop sponsored by the NIH performed. Criteria developed based on expert opinion and with reference to existing literature, however the specific role of evidence in informing decision-making is not described in detail	No	No	No
Zhou, 2018	Irritable bowel syndrome	No	Revised	Consensus meeting of 13 multidisciplinary North American experts was held to initiate development of diagnostic criteria for abdominal, pelvic, and urogenital pain conditions using the AAPT framework. A literature review identified IBS as a high-priority condition, and the Rome IV criteria were retained with stylistic modifications to align with the AAPT taxonomy. Features were included based on their established relevance in the Rome IV criteria and to maintain consistency across chronic pain disorders, with the final criteria refined through expert discussion in working group meetings	No	No	No
Zuraw, 2012	Hereditary angioedema with normal C1 inhibitor function (HAE-nC1INH)	No	New	A 1-day international expert symposium was convened by the US HAE Association. Experts undertook a systematic review of the available literature relating to what is known about the disease, gaps in understanding and developing concrete criteria for the diagnosis of the disease before the group proposed an agenda to close these gaps. The final criteria were determined through expert discussion and group agreement	No	No	No





WHAT IS NEW?

Key findings

- Methodological approaches to developing diagnostic criteria for conditions without objective tests, biomarkers or reference standards were variable and inconsistently reported, acknowledging not all included studies were explicitly designed to develop diagnostic criteria
- Whilst consensus methods were employed by 71% of included studies to develop diagnostic criteria, few studies adopted strategies to ensure diverse or representative panel composition
- Within the limitations of the suboptimal reporting encountered, evidence synthesis was variably integrated into the diagnostic criteria development process and only 4% of groups used formal tools to assess the strength of evidence

What this adds to what is known?

- This scoping review introduces a conceptual map of current methodological approaches to developing diagnostic criteria, including consensus-based, literature-based and primary study-based approaches
- We identify several critical gaps relating to the integration of evidence synthesis and appraisal into diagnostic criteria development as well as methodological reporting practices
- Adequate diversity amongst panels developing diagnostic criteria using consensus methodology and patient/advocate consultation are infrequently considered, presenting barriers to health equity

What is the implication/what should change now?

- Future diagnostic criteria development initiatives should explicitly describe how evidence was used to inform each stage of development, for example by integrating systematic reviews, undertaking structured evidence appraisal and using evidence-to-decision frameworks. Specific methodological guidance is needed to ascertain the types of evidence that should be used
- Additional priorities should include proactive promotion of panel diversity when using consensus methods and inclusion of patients and advocates to ensure criteria are comprehensive and clinically relevant

None of the authors have any competing interests to disclose.