BeyondMold

CLINICAL SCIENCE MONOGRAPH · III

THE CELL DANGER RESPONSE SERIES

Counting Is Not Clustering

Provenance, Derivation, and Psychometric Validity of the CIRS Symptom-Cluster Diagnostic Model

The diagnostic instrument at the center of the CIRS construct — thirteen symptom clusters at an eight-of-thirteen threshold — was never derived by any recognized procedure, and its only "validation" is a circular concordance against its own case definition.

Andrew W. Heyman, MD, MHSA

Fellowship Director, Integrative Medicine, George Washington University · Founder, Beyond Mold

BEYOND MOLD · JUNE 2026 · HAWKE'S BAY, NZ

CONTENTS

Clinical Science Monograph III

This monograph examines the diagnostic instrument at the center of the CIRS construct — a roster of multi-system symptoms partitioned into thirteen clusters and applied at an eight-of-thirteen threshold as a screening gate — and demonstrates, from its own primary sources, that it was never derived by any recognized item-generation or structure-defining procedure, and that its only quantitative "validation" is circular.

ABSTRACT

The case in brief

The diagnosis of chronic inflammatory response syndrome (CIRS) rests, at the point of clinical entry, on a symptom instrument: a roster of multi-system complaints organized into thirteen symptom clusters, with the presence of eight or more clusters used as a screening threshold that flags a patient for the laboratory and visual-contrast testing that follow. This instrument is routinely presented as though it were empirically derived from patient data and statistically validated. This paper tests that presentation against its own primary sources: the two foundational *Neurotoxicology and Teratology* time-series papers (2005, 2006) in full, together with the single study (McMahon 2017) cited as quantitative validation.

Three findings follow. **First,** neither foundational paper performs a cluster analysis or any data-driven procedure that could generate a thirteen-cluster structure; both assign symptoms to organ systems a priori by clinical judgment and analyze only the group-mean number of symptoms by paired t-tests within subjects across treatment time points. **Second,** the roster is not a fixed, derived object: it expands between the two papers from twenty-six symptoms in eight organ systems (a four-of-eight rule) to thirty-seven symptoms in ten organ systems (a five-of-ten rule), by editorial judgment rather than any selection algorithm; the thirteen-cluster, eight-threshold instrument appears in neither. **Third,** the McMahon validation applies the pre-existing instrument retrospectively and measures its "accuracy" against the construct's own case definition — an incorporation-biased, circular comparison — using elementary contingency-table statistics computed on public web calculators, in a single specialty clinic.

Measured against the established standards for symptom-roster and case-definition development — content and structural validity, reliability, criterion validity against an independent reference standard, threshold derivation, and external replication, as codified in COSMIN, STARD, and TRIPOD — the CIRS cluster model does not meet even the minimum entry criteria. The deficiency is not peripheral: a count-based instrument validated against itself will, by construction, assign high apparent prevalence to any symptomatic population and cannot distinguish its target condition from the many overlapping illnesses it resembles. The paper closes with the protocol an adequate re-derivation would require — the same prospective, biomarker-anchored design available to the Cell Danger Response program that supersedes the construct.

SECTION 1

The claim under examination

Every diagnostic system has a front door. For CIRS, the front door is a symptom instrument. Before any laboratory marker is drawn or any visual-contrast test administered, a patient is screened against a roster of multi-system symptoms that has been organized into thirteen groupings — the thirteen clusters — with a count of eight or more clusters treated as the threshold that identifies a probable case and justifies the downstream work-up. The instrument is the gate through which patients enter the construct, and it is the object whose prevalence claims have been used to assert that CIRS is, in the words of the validation study, one of the greatest public-health problems in existence [3].

Because the instrument carries this much weight, its provenance matters. A diagnostic instrument inherits its authority from the process that produced it: the steps by which its items were generated, reduced, and grouped; the analyses by which its internal structure was confirmed; the independent standard against which its accuracy was measured; the studies in which it was replicated. When those steps are present and transparent, an instrument earns the trust placed in it. When they are absent, the instrument may still be used — but it is being used on credit it has not earned.

The method here is deliberately conservative. Rather than rely on secondary descriptions — which uniformly report that some thirty-five symptoms "reached significance versus controls and were categorized into thirteen clusters" — the analysis works from the two foundational peer-reviewed papers in their entirety [1,2], and from the one study advanced as quantitative validation [3]. Section 2 first sets out the yardstick; Sections 3 and 4 establish what those papers actually did; Section 5 measures the one against the other; Sections 6 and 7 develop the consequences and the path to a defensible instrument.

One point of fairness frames everything that follows. The 2005 and 2006 papers were, on their own stated terms, within-subject time-series clinical trials of cholestyramine: their aim was to show that a group-mean symptom burden and a visual-contrast deficit move in synchrony as a binding agent is given, withdrawn, and re-given. Judged as that, they are internally coherent. The critique here is not of those trials as trials. It is of the secondary use of those papers as the derivation and validation of a thirteen-cluster diagnostic instrument — a claim the papers do not contain and cannot support.

SECTION 2

What the development of a validated symptom roster requires

A symptom roster used to identify a disease is a measurement instrument, and a clustered, threshold-scored roster used to gate a diagnosis is a clinical prediction instrument. Both are governed by a mature and consensual methodology — classical test theory and its modern successors, codified for health measurement in the COSMIN taxonomy [14], for diagnostic-accuracy studies in STARD [15], for prediction models in TRIPOD [16],

and in the standard scale-development literature [9–13]. It is not an unusually demanding standard; it is the ordinary one, the floor any instrument proposing to label patients with a disease is expected to clear.

2.1 Two distinctions that must be fixed first

Before any pipeline begins, two distinctions must be settled. The first is between a *case definition* (built to assemble homogeneous study groups, tolerating a sensitivity–specificity trade) and a *diagnostic instrument* (applied to an individual and calibrated to the population of actual use) [18]. An instrument derived as one cannot be silently redeployed as the other. The second is between the *index test* — the instrument under evaluation — and the *reference standard* against which it is judged. The reference standard must be independent of the index test; when the same information enters both, the resulting agreement is an artifact called *incorporation bias*, and the reported accuracy is spurious [15]. These two distinctions return as the decisive failures in Section 5.

2.2 Item generation and content validity

Development begins with a theory-anchored definition of the construct, followed by generation of an item pool that over-samples its content [10–12]. Items are drawn from three sources in combination: a systematic reading of the prior literature; structured expert input; and — indispensably — qualitative work with affected patients, whose lived symptom vocabulary frequently differs from the clinician's. The pool is then assessed for content validity by expert and patient panels, often quantified as a content-validity index, and revised. This stage answers a question no later statistic can repair: do the items, as a set, represent the disorder, and are any obviously relevant domains missing?

2.3 Item reduction and the empirical derivation of structure

An over-inclusive pool must then be reduced, and — critically — its internal structure must be *discovered, not assumed*. This is the stage at which the word "cluster" acquires a technical meaning. If an instrument asserts that its items fall into a fixed number of groupings, that grouping is an empirical claim about the covariance structure of the data, established by dimension-finding methods: exploratory factor analysis or PCA to reveal how many latent dimensions the items occupy; confirmatory factor analysis to test a hypothesized structure against fresh data; latent-class or latent-profile analysis when the structure is categorical; or formal hierarchical or k-means cluster analysis [13]. Each reports diagnostics — eigenvalues and scree, factor loadings, fit indices, class-separation statistics. A grouping produced instead by clinical intuition is an a priori organizing convenience; it is not a derived structure, and it carries none of the evidentiary weight the word "cluster" is meant to convey. The number thirteen, if it is to mean anything, must come *out* of such an analysis, not be imposed before it.

2.4 Reliability

An instrument cannot be valid if it is not reliable. Three forms are expected, each with accepted thresholds [9,14,20]. *Internal consistency* (Cronbach's α, or McDonald's ω) asks whether items within a cluster co-vary as a coherent set; a claimed cluster that does not cohere is not a cluster. *Test–retest reliability* asks whether a stable patient, re-assessed over a short interval, receives the same score (an intraclass correlation). *Inter-rater reliability* asks whether two clinicians agree — acute when the instrument is administered by interview, because the

interviewer becomes a measurement instrument whose consistency must itself be demonstrated. None is optional, and each is a number that must be reported.

2.5 Construct validity: convergent, discriminant, and known-groups

Validity is the evidence that the instrument measures what it claims. Convergent validity requires that scores correlate with established measures of the same construct; *discriminant validity* — the property most consequential here — requires that they separate the target condition from distinct conditions that produce overlapping symptoms [12,13]. For a multi-system, fatigue-and-cognition syndrome, discriminant validity is the entire game: the roster must be shown to distinguish its target from chronic fatigue syndrome, fibromyalgia, depression, anxiety, sleep disorder, hypothyroidism, and the other high-prevalence conditions whose symptom lists are nearly identical. An instrument never tested against look-alikes has no evidence that a high score means the target disease rather than simply ill health.

2.6 Criterion validity against an independent reference standard

Where an external reference standard exists, criterion validity is established by measuring sensitivity, specificity, predictive values, and likelihood ratios against it — in a representative sample, with index and reference read independently, and the reference genuinely independent of the index test [15]. When no gold standard exists — the usual situation for a contested syndrome — the honest responses are latent-class modeling, validation against hard external outcomes, or explicit acknowledgment that criterion validity cannot yet be claimed. What is *not* permitted is to validate the instrument against the construct's own case definition, which merely measures its agreement with a relabeled version of itself.

2.7 Thresholds, calibration, and the base-rate problem

A cut-point such as "eight of thirteen clusters" is not a self-evident fact; it is a parameter that must be derived and justified. Proper derivation places the threshold on a receiver-operating-characteristic analysis that displays the full sensitivity–specificity trade-off, selects the operating point against the clinical costs of misclassification, and reports calibration [16]. A threshold must then be interpreted through the base rate of the population in which it is applied: applying a low-specificity roster in a population already enriched for the condition (a specialty referral clinic) inflates the apparent positive rate. A prevalence figure generated inside a referral clinic and projected onto the public is a base-rate error, not a population estimate.

2.8 External validation, replication, and transparent reporting

Finally, an instrument is not validated by the group that built it. Credible instruments are subjected to external validation in independent samples, by independent investigators, in the settings where they will be used, and reported against published checklists — STARD for accuracy studies, TRIPOD for prediction models — that exist precisely to expose the design weaknesses that inflate performance [15,16]. An instrument whose entire evidentiary chain remains within one author group, one clinic, and venues without rigorous review has not been validated in the sense the word carries.

The pipeline in one view. Table 1 collects these stages, the question each answers, the accepted methods, and the reporting standard. It is the template against which the CIRS instrument is read in Section 5.

Stage	Question it answers	Accepted methods / evidence	Reporting standard
Item generation	Do the items represent the construct?	Literature review + expert panel + patient qualitative work; content-validity index	COSMIN content validity
Structure derivation	How many groupings exist, and which items belong?	EFA/PCA → CFA; latent-class/profile analysis; formal cluster analysis with reported diagnostics	COSMIN structural validity
Reliability	Are results consistent?	Internal consistency (α/ω) per subscale; test–retest ICC; inter-rater agreement	COSMIN reliability
Construct validity	Does it measure the target and not look-alikes?	Convergent, discriminant, and known-groups testing against overlapping conditions	COSMIN hypotheses testing
Criterion validity	How accurate against truth?	Sensitivity/specificity/LRs vs an *independent* reference standard; latent-class methods if none exists	STARD
Threshold & calibration	Where is the cut-point and is it honest?	ROC-derived threshold; calibration; base-rate–aware predictive values	TRIPOD
External validation	Does it hold elsewhere, in other hands?	Independent samples, independent investigators, multiple settings; replication	STARD / TRIPOD; peer review

Table 1. The accepted development pipeline for a validated symptom roster or diagnostic case definition. Each stage is a prerequisite, not an option; the right-hand column names the consensus reporting framework that governs it.

SECTION 3

What the foundational papers actually did

Against the template of Section 2, the two foundational papers can be read for what they contain. The reading is unambiguous: both are within-subject time-series trials whose symptom methodology is a priori organ-system assignment followed by analysis of a symptom *count*. Neither contains any element of structure derivation, reliability testing, construct validation, or criterion validation. Neither contains a cluster analysis.

3.1 The 2005 study: twenty-six symptoms, eight organ systems, four-of-eight

The 2005 paper enrolled twenty-one volunteers from five water-damaged buildings, nineteen of whom completed all five time points [1]. Its case definition required, as its second criterion, symptoms in "at least four of the eight organ systems" in its Table 2 — a four-of-eight organ-system rule, not an eight-of-thirteen cluster rule. The symptom list comprised twenty-six items; completers "averaged 14.9 symptoms out of 26" at baseline. The grouping of those items into eight organ systems is presented as a fixed organizing scheme, assigned by clinical category. No procedure is described, or implied, by which that grouping was derived from the data. The methods confirm the point: the analysis section states that "the MANOVA analyses reduced to paired t-tests" comparing group-mean values across time points, with Bonferroni correction [1]. There is no test of any

individual symptom, and no analysis of how symptoms co-vary. The only comparison group is historical — eighty-seven earlier controls whose data "are shown for comparison only and were not statistically analyzed."

3.2 The 2006 study: thirty-seven symptoms, ten organ systems, five-of-ten

The 2006 companion enrolled twenty-eight participants, twenty-six completing all five time points, and added a double-blinded cholestyramine–placebo arm [2]. Its case definition required symptoms in "at least five of the ten organ systems" — a five-of-ten rule — and its roster had grown to thirty-seven items. The methods are explicit: "symptoms were categorized into organ systems," with headache and skin sensitivity deliberately set aside as "unique, multifactorial symptoms." That is clinical categorization performed before analysis — the opposite of a derived grouping. The analysis is the same as 2005: "the MANOVA analyses reduced to paired t-tests" across time points [2]. Baseline burden is reported as a count — "an average of 23 out of 37 symptoms" — that falls on therapy, rises on re-exposure, and falls again on re-treatment. No symptom is tested individually, no covariance structure is examined, and no cluster analysis is performed. The eight-of-thirteen cluster instrument is absent here as well: the operative rule is five-of-ten organ systems.

3.3 The roster grew by judgment; the clusters are nowhere derived

Read side by side, the two papers expose the central fact. The instrument is not a fixed, derived object that the papers validate; it is a clinically curated list that changed between the two studies. The symptom count rose from twenty-six to thirty-seven, the organ-system count from eight to ten, and the threshold from four-of-eight to five-of-ten — all by editorial revision, none by any selection procedure. An instrument whose very dimensionality shifts between its two founding papers, without a methodological step to account for the shift, is by definition a work of judgment rather than derivation.

Feature	2005 paper [1]	2006 paper [2]	13-cluster instrument (in use)
Symptom items	26	37	~37 (variable)
Grouping unit	8 organ systems	10 organ systems	13 clusters
How grouping was set	A priori clinical category	A priori clinical category	Not derived in either paper
Inclusion threshold	4 of 8 systems	5 of 10 systems	8 of 13 clusters
Symptom analysis	Count; paired t-tests	Count; paired t-tests	None reported
Cluster analysis?	None	None	None
Control comparison	Historical; not analyzed	Within-case across time	—
Sample (completers)	19	26	—

Table 2. Evolution of the symptom instrument across the two foundational papers, with the thirteen-cluster instrument actually deployed. The grouping unit, threshold, and item count all change by judgment; no cluster analysis appears in either source, and the deployed structure matches neither paper.

A pre-emptive clarification is warranted, because the 2005 paper does contain the phrase "cluster analysis." It occurs once, in the literature review, describing a separate study by Linz and colleagues in which "a statistical cluster analysis defined problematic areas in which potentially etiologic microbes were identified" [8] — a spatial clustering of building locations by another group, with nothing to do with symptoms or with the thirteen-cluster roster. A second observation, recorded for the companion HLA analysis: the 2005 Discussion is also where the genetic-susceptibility claim first enters the peer-reviewed literature, supported by a single citation to a 2002 conference abstract [7]. Two of the construct's load-bearing pillars trace to the same paper, and neither is actually built there.

SECTION 4

The validation that was not: McMahon 2017

If the clusters are not derived in the foundational papers, the weight of empirical justification shifts entirely onto the one study presented as their validation [3]. That study does not bear it. It is a retrospective review of 1,061 consecutive patients at a single CIRS specialty clinic, of whom 371 met the existing case definition; the thirteen-cluster instrument, with panels of laboratory and screening tests, was then applied to those 371 patients across four age bands. The verb is decisive: the study *applies* the cluster instrument; it does not *derive* it. Nothing in the study generates the thirteen-cluster structure or the eight-cluster threshold; both are taken as given.

The deeper defect is the reference standard. The clusters' reported "sensitivity" is measured against the construct's own case definition — a definition that itself rests on the same symptom roster and requires improvement on therapy to confirm [3]. Index test and reference standard are therefore not independent; they are two expressions of one construct, applied to one population. This is incorporation bias in its textbook form (§2.1), and the high concordance it produces is an artifact of the circularity, not evidence of accuracy. The study measures how well a relabeled version of the instrument agrees with itself.

The statistical machinery matches the modesty of the design. The reported "error rates," quoted as ranging from 1.24×10^{-3} to 1.10×10^{-6}, are the outputs of contingency-table tests computed on two public web calculators — a chi-square page and a two-by-two page — named directly in the reference list [3]. These are p-values from elementary association tests, not cross-validated estimates of diagnostic accuracy; there is no ROC analysis, no calibration, no multivariable modeling, and no out-of-sample testing. A small p-value establishes that two columns are associated; it says nothing about how an instrument will perform as a diagnostic test in a new patient.

Finally, the prevalence claim that gives the study its rhetorical force — a pediatric prevalence of 7.01 percent, advanced as evidence that CIRS is a public-health emergency — is generated from this same single specialty clinic together with one partner pediatric practice [3]. A rate computed inside a referral population enriched for

the condition, then projected onto the general public, is precisely the base-rate error of §2.7. The instrument's limited specificity, never independently established, guarantees that such a projection will overstate true prevalence. The cluster instrument and its prevalence claim were subsequently codified in a consensus statement issued by the same construct community [4], which formalizes the instrument but adds no independent derivation or validation to it.

SECTION 5

Five structural failures against the standard

The findings of Sections 3 and 4 can now be laid directly against the pipeline of Section 2. The instrument fails at distinct stages, each independently disqualifying for a tool used to assign a disease label. Table 3 states the crosswalk; the subsections elaborate.

Required property	What the standard requires	What the construct provides	Verdict
Content & structure derivation	Item pool from literature/experts/patients; structure found by factor or cluster analysis	Symptoms grouped a priori; roster changed between papers; no derivation	Absent
Reliability	Internal consistency, test–retest, inter-rater — each reported	None reported for the cluster structure	Absent
Discriminant validity	Separation from CFS, fibromyalgia, depression, etc.	Never tested against look-alike conditions	Absent
Independent criterion validity	Accuracy vs a reference standard independent of the index test	Validated against the construct's own case definition	Circular
Threshold & calibration	ROC-derived cut-point; calibration; base-rate–aware	8-of-13 cutoff has no published derivation; web-calculator p-values	Absent
External replication & reporting	Independent samples/investigators; STARD/TRIPOD reporting	Single author group, single clinic; no STARD/TRIPOD report	Absent

Table 3. The CIRS cluster instrument measured against the minimum development standard. Each row is independently disqualifying for a diagnostic instrument; the construct fails all six.

5.1 No derivation of content or structure

The instrument never completes the first two stages of the pipeline. There is no documented item-generation process — no qualitative patient work, no content-validity assessment — and, decisively, no structure-derivation step. The thirteen-cluster partition is asserted, never derived; no factor analysis, latent-class analysis, or cluster analysis of the symptom data has ever been published, and the only grouping the foundational papers contain is the a priori organ-system scheme that itself changed from eight categories to ten [1,2]. In COSMIN

terms, both content validity and structural validity are simply missing. A claimed structure that has never been subjected to a structure-finding analysis is not a finding; it is a layout.

5.2 No reliability evidence

No internal-consistency coefficient has been reported for any cluster, so there is no evidence that the items within a cluster cohere as a measurable dimension. No test–retest statistic establishes that a stable patient scores consistently. And because the roster is administered by clinician interview, the absence of any inter-rater analysis is especially damaging: the interviewer is part of the instrument, and the consistency of that interviewer has never been demonstrated [9,20]. An instrument of unknown reliability cannot be an instrument of established validity, because reliability bounds validity.

5.3 No discriminant validity — the decisive clinical gap

This is the failure that matters most at the bedside. The CIRS roster consists of fatigue, cognitive difficulty, pain, headache, mood change, and other non-specific multi-system complaints — the shared vocabulary of chronic fatigue syndrome, fibromyalgia, depression, anxiety, obstructive sleep apnea, hypothyroidism, and many other common conditions. An instrument built from such items can only mean "this disease" rather than "some illness" if it has been shown to discriminate the target from those alternatives [12,13]. That test has never been performed. Without discriminant validity, a high cluster count is uninterpretable: it is fully consistent with the patient having any of a dozen other conditions, or several at once.

5.4 Circular criterion validity and an underived threshold

The single study offered as validation measures the instrument against the construct's own case definition, which is incorporation bias rather than criterion validity [3] (§4). Independently of that circularity, the operative threshold — eight of thirteen clusters — has no published derivation at all: no ROC analysis locating it on a sensitivity–specificity curve, no calibration, and no statement of the clinical loss function that would justify that cut-point. The statistics that do appear are p-values from public contingency-table calculators applied to tiny, pre-selected samples (nineteen and twenty-six completers) [1,2]. Statistical significance of a within-subject symptom-count change under treatment is a real finding about cholestyramine; it is not, and cannot be converted into, the diagnostic accuracy of a thirteen-cluster classifier.

5.5 No external validation, replication, or transparent reporting

The entire evidentiary chain remains inside one author group and, effectively, one clinic. There is no external validation by independent investigators, no multi-site testing, and no report conforming to STARD or TRIPOD — the checklists that exist to expose exactly the design features (incorporation bias, selected samples, missing calibration) that characterize this instrument [15,16]. The key codifying documents appear in venues without rigorous methodological peer review [3,4], and the literature advances by a citation cascade in which each paper cites the others, producing the appearance of accumulating evidence without the substance of independent confirmation.

SECTION 6

Why the defect is clinically consequential

These are not bookkeeping objections. A count-based instrument that has never been tested for discriminant validity and is validated against itself has a predictable and harmful behavior: it labels a large fraction of any symptomatic population as cases. Because its items are the common currency of chronic illness, and because no step ever established that a high count is specific to one disease, the instrument cannot help but capture patients whose fatigue and cognitive complaints arise from depression, sleep disorder, autoimmune disease, or simple multimorbidity. The high prevalence the construct reports is therefore not a discovery about the world; it is a property of the instrument — the arithmetic consequence of a sensitive, non-specific, self-validated screen applied in an enriched population [3].

The clinical cost is twofold. Patients who in fact have a different, possibly treatable condition may be routed into a single explanatory channel and away from the correct diagnosis — the precise risk that discriminant-validity testing exists to prevent. And the credibility of genuine environmental-illness research is burdened by a prevalence claim the instrument cannot support. None of this implies that patients are not ill, or that water-damaged buildings cause no disease; it implies only that this instrument cannot establish who has *this* disease, because it was never built to.

SECTION 7

What an adequate re-derivation would require

The remedy is not rhetorical but methodological, and it is entirely achievable. A defensible instrument would be rebuilt through the full pipeline of Section 2. Item generation would begin from the literature and from structured qualitative work with affected patients, producing a content-validated pool rated by clinical and methodological experts. Structure would be *discovered, not assumed*: exploratory factor or latent-class analysis in a derivation cohort would determine how many symptom dimensions actually exist — a number that may or may not be thirteen — with the solution then confirmed by confirmatory factor analysis in a separate cohort [13].

Reliability would be quantified — internal consistency per derived dimension, test–retest in stable patients, inter-rater agreement for the interview administration. Discriminant validity, the indispensable step, would test the instrument against prospectively enrolled comparison groups with chronic fatigue syndrome, fibromyalgia, depression, sleep disorder, and thyroid disease [12]. Criterion validity would be assessed against a reference standard genuinely independent of the symptom roster — and where no gold standard exists, latent-class methods or hard external outcomes would replace the circular comparison. The threshold would be derived from an ROC analysis with explicit attention to the clinical costs of misclassification, and prevalence estimated only in appropriately sampled populations [15,16]. Finally, the instrument would be externally validated by independent groups and reported against STARD and TRIPOD.

There is an opportunity in this for the program that supersedes the construct. The same prospective, multi-site, biomarker-anchored design required to rebuild a symptom instrument correctly is the design the Cell Danger Response framework needs in any case to validate its phenotypes and its readiness instrument. An objectively grounded staging — mitokine markers such as GDF15 and FGF21 read alongside a properly derived symptom structure — would let the symptom roster be anchored to measurable biology rather than to itself [19]. The path out of the circularity is to derive the structure, validate it against something other than itself, and let independent hands try to break it.

SECTION 8

The wider literature: symptom rosters in water-damaged-building research

The CIRS instrument does not exist in a vacuum. Symptom rosters have been used in water-damaged-building (WDB) and sick-building-syndrome (SBS) research for four decades, and reading the CIRS construct against that wider literature is clarifying: the field already contains a properly validated symptom instrument, and the part of the field most sympathetic to a broad, multi-system illness is candid about lacking one. The construct examined here is the outlier on both counts.

8.1 The validated tradition: the MM / Örebro questionnaire

The reference instrument of the field is the "MM" questionnaire, developed at Örebro University Hospital from the early 1980s and refined across dozens of pilot studies. Unlike the CIRS roster, it was explicitly "tested for usefulness, reliability and validity" before deployment [21], and has since been used, translated, and re-tested in hundreds of studies worldwide. Its content is deliberately narrow — non-specific symptoms of the eyes, upper airways, and skin, with headache and fatigue — and its purpose is equally specific: to estimate the prevalence of building-related symptoms, not to diagnose an individual with a discrete disease. That modesty is a feature; it is what allows the instrument to be valid for what it claims. The same tradition shows the field is fully capable of the multivariate methods the CIRS construct never used — for example, separating high- and low-prevalence buildings by principal-components analysis [22]. Their absence from the CIRS papers is a choice, not a limitation of the era.

8.2 The epidemiologic syntheses: what is actually established

The weight of the field rests on large, careful evidence reviews — the WHO indoor-air guidelines on dampness and mould [24], the Institute of Medicine's *Damp Indoor Spaces and Health* [25], and the comprehensive epidemiologic review by Mendell and colleagues [23], with quantitative meta-analyses for specific outcomes such as asthma development [26]. Their convergent conclusion is precise and worth stating exactly: indoor dampness or mould is consistently associated with respiratory and allergic outcomes — asthma, wheeze, cough, upper-respiratory symptoms — while causal links remain unproven and, critically, current evidence does not support measuring specific indoor microbiologic factors to guide health decisions [23]. The robustly supported health

effects are thus respiratory and allergic; the broad multi-system, neurocognitive syndrome that the CIRS and DMHS constructs describe is the *least* established part of the literature, not the most.

8.3 The DMHS strand: broad in scope, candid about its limits

A more recent European literature — led by Hyvönen, Tuuminen, Lohi, and Valtonen — proposes a broader entity, "dampness and mold hypersensitivity syndrome" (DMHS), that, like CIRS, encompasses multi-system and neurological complaints [27,28]. What distinguishes this strand from the CIRS construct is its candor. Valtonen's review, proposing a five-part clinical DMHS criterion set, states plainly that the literature contains no previously proposed diagnostic clinical criteria, that no unanimously accepted laboratory test exists, and that diagnosis therefore remains clinical and provisional [29]; related work frames SBS as a possible "starting point" for an as-yet-uncharacterized chronic process [30]. This strand makes the same broad multi-system claim the CIRS construct makes, but it does not pretend the supporting instrument has been validated. That honesty is exactly what is missing from the cluster model.

Three strands, three postures. Table 4 places the validated SBS tradition, the DMHS strand, and the CIRS construct side by side.

Strand	Instrument	Validation status	Posture toward its own limits
MM / Örebro (SBS epidemiology) [21]	Standardized, pilot-tested questionnaire; narrow symptom set	Tested for reliability and validity; used worldwide	Honest — measures association/prevalence, not individual diagnosis
DMHS strand (Hyvönen, Valtonen, Tuuminen) [27–30]	Proposed clinical criteria; broad multi-system scope	Not validated; no accepted laboratory test	Honest — explicitly states criteria are unvalidated and provisional
CIRS cluster model [1–4]	13-cluster / 8-threshold roster; broad multi-system scope	Presented as validated; is not	Not candid — claims a derivation and validation it does not have

Table 4. Three strands of symptom-roster use in water-damaged-building research. The validated instrument is deliberately narrow; the broad DMHS proposal is openly unvalidated; the CIRS construct is the only strand that combines broad scope with an unearned claim of validation.

SECTION 9

The gap in the literature

Read as a whole, the field reveals a genuine and consequential gap — one that the CIRS construct does not fill but obscures. The gap has four components.

- **Association, not diagnosis.** The validated instruments were built to relate building conditions to symptom prevalence in populations; they were never designed, and do not claim, to assign a discrete disease to an

individual [21,23]. There is, at present, no symptom instrument in this domain validated as a diagnostic classifier for a distinct multi-system mold illness.

- **The best-validated outcomes are the narrowest.** The respiratory and allergic effects of dampness are the ones the evidence supports; the multi-system neurocognitive syndrome is the part of the field with the weakest validation [23–25]. The CIRS and DMHS constructs operate precisely in that weakest region.
- **Discriminant validity is missing field-wide.** The non-specific symptoms at issue are shared across SBS, CFS, fibromyalgia, MCS, depression, and other persistent-symptom conditions, all of which display the same discrepancy between subjective burden and objective findings. No instrument in the field has been shown to separate a distinct mold illness from these neighbors. The DMHS literature concedes this directly [29]; the CIRS construct simply does not test it.
- **Honesty is unevenly distributed.** The gap is openly acknowledged by the validated tradition and by the DMHS strand. The distinctive failure of the CIRS construct is not that it confronts an unsolved problem — the whole field does — but that it presents the problem as solved: a broad, unvalidated instrument marketed as a derived and validated diagnostic test, with a referral-clinic prevalence projected onto the public as fact [3].

SECTION 10

A roadmap to validate a symptom set for water-damaged-building illness

Section 7 set out the generic requirements for any validated instrument; this section grounds them in the water-damaged-building literature specifically, so that the path is concrete. The encouraging fact is that the field already supplies most of the building blocks — a validated item base, an established quasi-experimental design, candidate objective markers, and a clear catalogue of the conditions an instrument must be distinguished from. What has never been done is to assemble them into a single, transparently reported validation program.

Several steps deserve emphasis because they exploit something specific to this field. Item generation need not start from nothing: the validated MM/Örebro item base [21] is the natural scaffold, extended with the multi-system items the DMHS literature describes [27–29] and refined through structured qualitative work with affected patients. Structure would then be discovered by exploratory and confirmatory factor analysis or latent-class modeling [22]. Discriminant validity is the decisive and field-defining step: prospectively enrolled comparison groups against which the instrument must demonstrate separation. A genuine strength available here is the quasi-experimental removal-and-remediation design — the within-subject logic the foundational papers already used — which, paired with blinded re-assessment and a properly developed instrument, can establish criterion validity against an exposure manipulation rather than against the construct's own definition. Finally, the symptom instrument should be anchored to objective biology: validated respiratory and allergic markers where the outcome is respiratory [23], and candidate mitochondrial mitokines such as GDF15 and FGF21 as objective correlates of the multi-system component [19].

Step	What to do in the WDB context	What it guards against
Anchor items	Start from the validated MM/Örebro base [21]; add multi-system items via patient qualitative work [27–29]	Reinventing a weaker instrument; omitting real symptoms
Define the construct	Pre-register: distinct entity vs severity gradient; specify intended population and use	Post-hoc category drift; case/diagnostic confusion
Measure exposure	Documented or measured dampness, not self-report alone (heeding the caution against specific microbiologic measures) [23]	Recall bias; confounding by expectation
Discover structure	EFA → CFA or latent-class analysis; let the data set the number of clusters (cf. PCA in SBS [22])	Imposing a fixed, underived cluster count
Establish reliability	Internal consistency per dimension; test–retest; inter-rater for interview administration	Inconsistent, unrepeatable measurement
Test discriminant validity	Prospective comparison vs CFS, fibromyalgia, MCS, depression, anxiety, sleep disorder, thyroid disease	Labeling general ill-health as one specific disease
Use the removal design	Exploit remediation/removal and re-exposure with blinded re-assessment	Placebo and expectation effects; circularity
Anchor to biology	Validated respiratory/allergic markers; candidate CDR mitokines GDF15/FGF21 [19]	Symptom-only self-referential validation
Derive threshold & prevalence	ROC-based cut-point with calibration; estimate prevalence only in representative samples	Arbitrary cutoffs; base-rate inflation
Validate externally	Independent investigators, multiple sites; STARD/TRIPOD reporting; replication	In-house, non-replicated, non-transparent claims

Table 5. A field-grounded roadmap for validating a water-damaged-building symptom set. Each step uses a resource the field already possesses; the program's novelty is in assembling them into a single, transparently reported validation rather than any one technique.

> **A minimum bar.** Pending such a program, a defensible position is simply stated: no symptom instrument should be used to assign a multi-system mold-illness diagnosis to an individual until it has, at minimum, demonstrated a data-derived internal structure, established reliability, and shown discriminant validity against the major look-alike conditions in an independent sample. That is the ordinary threshold every diagnostic instrument is expected to clear — and the one the CIRS cluster model has not approached.

SECTION 11

Conclusion

The thirteen-cluster CIRS instrument is presented with the authority of a derived and validated diagnostic tool. Its own primary sources show that it is neither. The two foundational peer-reviewed papers contain no cluster analysis and no derivation of any kind; they assign symptoms to organ systems by clinical judgment and analyze a within-subject symptom count, and the roster they describe changes — in item number, grouping, and threshold — between the two studies [1,2]. The single study advanced as validation applies the instrument rather than deriving it, measures it against the construct's own case definition, computes elementary contingency statistics on public web calculators, and generates a prevalence figure by projecting a referral-clinic rate onto the public [3]. Measured against the ordinary, consensual standards for building a symptom roster, the instrument fails at every stage of the pipeline.

Placed against the wider water-damaged-building literature, the verdict sharpens rather than softens. That literature already contains a properly validated symptom instrument — narrow, honest, and used worldwide [21] — and a parallel strand that makes the same broad multi-system claim as CIRS while openly conceding that its criteria are unvalidated [29]. The CIRS construct is the single place in this landscape where a broad, unvalidated, multi-system roster is presented as a derived and validated diagnostic test. The gap it claims to fill is real, but it fills that gap with the appearance of evidence rather than the substance.

The pattern is the same one documented for the genetic-susceptibility claim: an instrument presented as statistically derived whose foundational papers contain no derivation, validated in-house against a non-independent standard, and propagated through a closed citation cascade. The two defects even share a birthplace in the 2005 Discussion. The remedy is achievable with resources the field already holds (Section 10): derive the structure with the methods designed for the purpose, validate it against something other than itself, report it transparently, and submit it to independent replication.

Until that is done, the cluster count should be read for what it is — a clinically curated tally of common symptoms — not for what it is presented to be: a validated diagnostic classifier.

REFERENCES

References

1. Shoemaker RC, House DE. A time-series study of sick building syndrome: chronic, biotoxin-associated illness from exposure to water-damaged buildings. *Neurotoxicol Teratol.* 2005;27(1):29–46.

2. Shoemaker RC, House DE. Sick building syndrome (SBS) and exposure to water-damaged buildings: time series study, clinical trial and mechanisms. *Neurotoxicol Teratol.* 2006;28(5):573–588.

3. McMahon SW. An evaluation of alternate means to diagnose chronic inflammatory response syndrome and determine prevalence. *Med Res Arch.* 2017;5(3).

4. Shoemaker RC, et al. A diagnostic process for chronic inflammatory response syndrome (CIRS): a consensus statement. *Intern Med Rev.* 2018;4(5).

5. Shoemaker RC, Hudnell HK. Possible estuary-associated syndrome: symptoms, vision, and treatment. *Environ Health Perspect.* 2001;109(5):539–545.

6. Grattan LM, Oldach D, Perl TM, et al. Learning and memory difficulties after environmental exposure to waterways containing toxin-producing *Pfiesteria* or *Pfiesteria*-like dinoflagellates. *Lancet.* 1998;352(9127):532–539.

7. Shoemaker RC. Differential association of HLA-DR genotypes with susceptibility to chronic, neurotoxin-mediated illness. Poster, 51st Annual Meeting, American Society of Tropical Medicine and Hygiene; Nov 2002; Denver, CO. *Am J Trop Med Hyg.* 2002;67(Suppl):160 (abstract).

8. Linz DH, Pinney SM, Keller JD, White M, Buncher CR. Cluster analysis applied to building-related illness. (Cited in ref 1, reference 104.)

9. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use.* 5th ed. Oxford University Press; 2015.

10. DeVellis RF. *Scale Development: Theory and Applications.* 4th ed. SAGE Publications; 2017.

11. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front Public Health.* 2018;6:149.

12. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. *Psychol Assess.* 1995;7(3):309–319.

13. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess.* 1995;7(3):286–299.

14. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63(7):737–745.

15. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ.* 2015;351:h5527.

16. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162(1):55–63.

17. Fukuda K, Straus SE, Hickie I, et al. The chronic fatigue syndrome: a comprehensive approach to its definition and study. *Ann Intern Med.* 1994;121(12):953–959.

18. Aggarwal R, Ringold S, Khanna D, et al. Distinctions between diagnostic and classification criteria? *Arthritis Care Res (Hoboken).* 2015;67(7):891–897.

19. Naviaux RK. Metabolic features of the cell danger response. *Mitochondrion.* 2014;16:7–17.

20. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60(1):34–42.

21. Andersson K. Epidemiological approach to indoor air problems. *Indoor Air.* 1998;8(Suppl 4):32–39. (Development and validation of the Örebro/MM questionnaire.)

22. Pommer L, Fick J, Sundell J, Nilsson C, Sjöström M, Stenberg B, Andersson B. Class separation of buildings with high and low prevalence of SBS by principal component analysis. *Indoor Air.* 2004;14(1):16–23.

23. Mendell MJ, Mirer AG, Cheung K, Tong M, Douwes J. Respiratory and allergic health effects of dampness, mold, and dampness-related agents: a review of the epidemiologic evidence. *Environ Health Perspect.* 2011;119(6):748–756.

24. World Health Organization Regional Office for Europe. *WHO Guidelines for Indoor Air Quality: Dampness and Mould.* Copenhagen: WHO; 2009.

25. Institute of Medicine. *Damp Indoor Spaces and Health.* Washington, DC: National Academies Press; 2004.

26. Quansah R, Jaakkola MS, Hugg TT, Heikkinen SAM, Jaakkola JJK. Residential dampness and molds and the risk of developing asthma: a systematic review and meta-analysis. *PLoS One.* 2012;7(11):e47526.

27. Hyvönen S, Lohi J, Tuuminen T. Moist and mold exposure is associated with high prevalence of neurological symptoms and MCS in a Finnish hospital workers cohort. *Saf Health Work.* 2020;11(2):173–177.

28. Hyvönen S, Tuuminen T, Lohi J. Occupants in moisture-damaged buildings may be at risk for various symptoms and inflammatory reactions: a case series report and literature review. *Arch Clin Med Case Rep.* 2019;3:692–701.

29. Valtonen V. Clinical diagnosis of the dampness and mold hypersensitivity syndrome: review of the literature and suggested diagnostic criteria. *Front Immunol.* 2017;8:951.

30. Tuuminen T. The roles of autoimmunity and biotoxicosis in sick building syndrome as a “starting point” for irreversible dampness and mold hypersensitivity syndrome. *Antibodies (Basel).* 2020;9(2):26.

Note on sources and evidence grading. The forensic claims about the foundational instrument (refs 1–2) and the validation study (ref 3) are drawn from the full texts of those papers; the quoted thresholds (4-of-8, 5-of-10), symptom counts (26, 37), statistical methods (paired t-tests on symptom counts), and the absence of any cluster analysis are verbatim from their Methods and Results sections. The measurement-science references (refs 9–18, 20) are standard, widely cited methodological sources. The wider water-damaged-building literature (refs 21–30) — the validated Örebro/MM tradition, the Mendell, WHO, and IOM evidence syntheses, and the Hyvönen/Valtonen/Tuuminen DMHS strand — was verified against current bibliographic databases. A small number of issue, page, or article-number details (notably refs 4, 7, 8, 21, and 26) should be confirmed against the originals before formal submission. This monograph is a methodological critique of a diagnostic instrument and a statement of the standards it should meet; it is not a claim that affected patients are not ill or that water-damaged-building exposure is without health consequence.