

From Messy Data to Meaningful Insight

Psoriasis & Wellbeing Study

Research audit | 13 October 2026 | Exploratory — not a confirmatory publication

Executive overview

A real, openly deposited 722-row psoriasis dataset looks complete until coded nonresponse is examined. Analysis needs its survey codebook, not just the CSV.

The 36-field file is semicolon-delimited with comma decimals. Without parsing it correctly and inspecting -9 response codes, a simple spreadsheet analysis could misrepresent missingness.

Dataset provenance and privacy

Uploaded file: Schuster et al_data(4).csv | Rows: 722 | Columns: 36 | Source:
<https://doi.org/10.23668/psycharchives.4451>

The uploaded participant-level CSV is deliberately withheld from all public artifacts. This report contains only aggregate, non-row-level results. Source dataset license and original codebook status must be rechecked before public publication.

Phase 1 — Audit

- 722 recorded rows, 36 fields; 0.00% literal missing cells.
- 0 whole-row duplicates; structure: One row per survey record; no independently verified participant identifier or repeat index.

Null cells: 0 (0.000% of all cells); blank-string cells: 0. Whole-row duplicate count: 0. Numeric columns: 36; other: 0.

Missingness and coding patterns

- AL01: 0.00% literal nulls
- AL02: 0.00% literal nulls
- AL02_01: 0.00% literal nulls
- AL02_02: 0.00% literal nulls
- AL02_03: 0.00% literal nulls
- AL02_04: 0.00% literal nulls

Observed negative codes: -9→50. These are NOT automatically classified as missing. Consult source coding manual.

Observed integrity flags

- minus9_cells: 50
- minus9_rows: 20
- ha03_minus9_cells: 37

- ha03_minus9_rows: 11
- age_min: 12
- age_max: 85

Phase 2 — Analyses and reproducibility

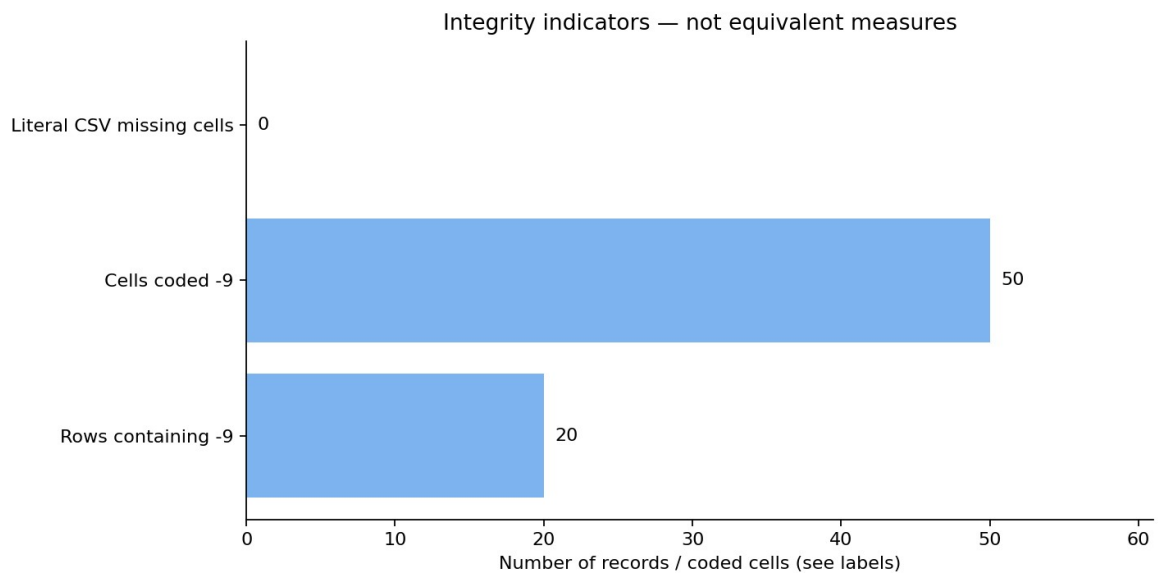
Exploratory associations use Spearman rank correlation with pairwise available cases. Group comparisons use Welch t-tests and pooled-SD Cohen d, but group meanings must be verified. Reliability uses complete cases only and is diagnostic in the absence of confirmed scale keying.

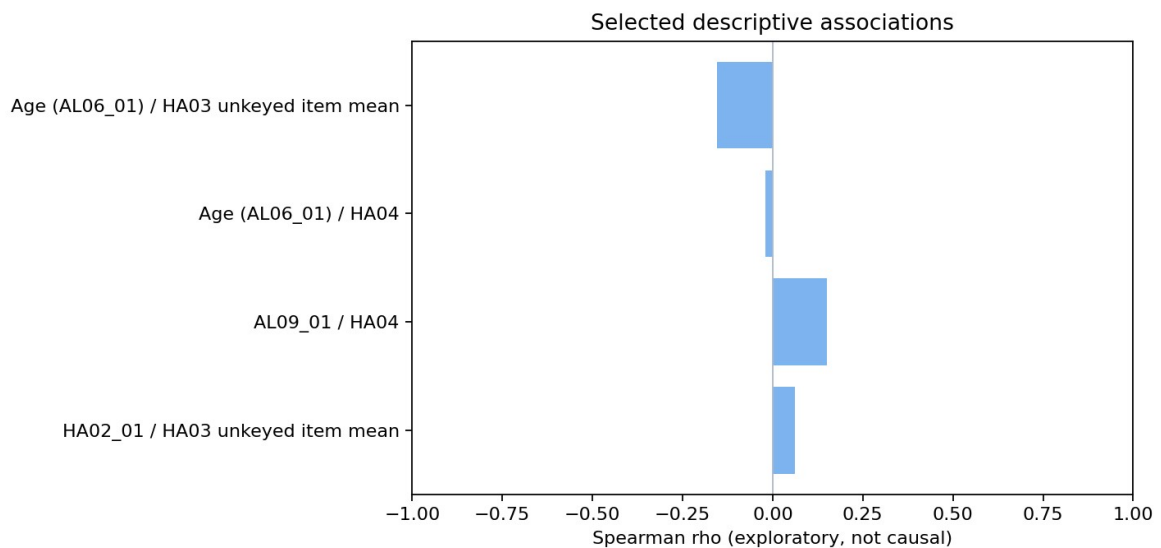
- ha03_raw_item_alpha_diagnostic: {"alpha": 0.2481, "n": 711, "items": 12}
- ha03_raw_item_omega_diagnostic: {"omega_one_factor_approx": 0.1812, "n": 711, "note": "Approximate one-factor omega; factor dimensionality and reverse scoring not validated"}
- age_group_raw_HA03: {"group1": "Age <40", "group2": "Age >=40", "n1": 268, "n2": 454, "mean1": 2.968, "mean2": 2.848, "cohen_d": 0.334, "p": 1.400592316880116e-05, "exploratory": true}

Sensitivity assumption ONLY: -9 is unavailable, and HA03 item values 1–5 are included. The unkeyed mean and reliability are not valid clinical outcomes; source scoring has not been verified.

Selected Spearman associations

- Age (AL06_01) / HA03 unkeyed item mean: rho=-0.1554, n=722, p=2.74e-05 (exploratory)
- Age (AL06_01) / HA04: rho=-0.0199, n=720, p=0.594 (exploratory)
- AL09_01 / HA04: rho=+0.1502, n=720, p=5.21e-05 (exploratory)
- HA02_01 / HA03 unkeyed item mean: rho=+0.0625, n=721, p=0.0933 (exploratory)





From findings to action

The file format matters

What we found — 722 records and 36 variables parse correctly only with semicolon delimiters; comma decimals require a separate option.

Why it matters / What to do — Import using the source-specific delimiter and decimal configuration.

Encoded missingness changes the audit

What we found — The file contains zero literal nulls but 50 -9 cells spread across 20 records.

Why it matters / What to do — Preserve -9 and verify its meaning against the published codebook before recoding.

Internal consistency is not validation

What we found — The raw-direction 12-item HA03 coefficient alpha is 0.248 for 711 complete cases under a -9-unavailable sensitivity assumption.

Why it matters / What to do — Obtain scale keys and investigate multidimensionality before reporting a wellbeing score.

Age-related patterns are exploratory

What we found — Age spans 12 to 85 and a raw HA03 item mean differs modestly across an explicitly chosen age-40 split.

Why it matters / What to do — Avoid clinical or causal conclusions until variable definitions and scoring are verified.

Data limitations — retained, not silently fixed

- The study dataset is published with a separate codebook, but the codebook was not part of the uploaded files and the original variable labels are not embedded in CSV.
- The -9 code appears in multiple questions; the sensitivity calculation assumes it denotes an unavailable response, but the original CSV remains unchanged.
- HA03 reliability is computed on raw directions and complete cases; its score semantics and reverse coding are not established. No clinical diagnosis or outcome score is inferred.
- The survey has no participant identifier or repeated-measures index; one row per respondent is assumed, not externally verified.
- Observational cross-sectional results cannot support causal claims.

Published as Schuster et al. (2021), PsychArchives CC BY 4.0, DOI 10.23668/psycharchives.4451. Only aggregate derivatives are included in this local review ZIP.

Reproducibility and citations

This bundle contains aggregate-results.csv, variable-dictionary.csv, two figures in PNG and SVG, a Python script, and references.bib. Original participant-level files are withheld. See REPRODUCIBILITY.md for exact command lines. This report is exploratory. Original source licenses are not established for the first two uploads.

Core methods: Spearman (1904), Cronbach (1951), McDonald (1999), Welch (1947), Cohen (1988), statsmodels MixedLM, scikit-learn FactorAnalysis. DOI/URL in references.bib.