Supplementary material

Application of normality tests in physiotherapy and rehabilitation research and its consequences for statistical power

This file contains the supplementary tables and figure referenced in the main manuscript.

Online supplemental table S1. The 16 physiotherapy and rehabilitation journals included in the full-text audit, with the number of retrieved articles that reported a named normality test (2015–2024).

Journal	PubMed abbreviation	Articles
Retrieved via Europe PMC		
Journal of Physical Therapy Science	J Phys Ther Sci	563
Brazilian Journal of Physical Therapy	Braz J Phys Ther	216
European J. of Physical and Rehabilitation Med.	Eur J Phys Rehabil Med	67
Archives of Physical Medicine and Rehabilitation	Arch Phys Med Rehabil	40
Physiotherapy Canada	Physiother Can	31
Hong Kong Physiotherapy Journal	Hong Kong Physiother J	29
Physical Therapy	Phys Ther	30
Clinical Rehabilitation	Clin Rehabil	25
Disability and Rehabilitation	Disabil Rehabil	8
Physiotherapy Theory and Practice	Physiother Theory Pract	7
Journal of Orthopaedic & Sports Physical Therapy	J Orthop Sports Phys Ther	7
Physical Therapy in Sport	Phys Ther Sport	5
Journal of Hand Therapy	J Hand Ther	2
Musculoskeletal Science and Practice	Musculoskelet Sci Pract	1
Retrieved via the Elsevier full-text API		
Physiotherapy	Physiotherapy	110
Journal of Physiotherapy	J Physiother	16

Counts are per-journal retrieval totals (full publication years 2015–2024); the analysis uses 1,157 unique de-duplicated articles.

Online supplemental table S2. Proportion of incorrect normality tests by journal (definitively coded papers).

Journal	Papers coded	n incorrect	% incorrect
Journal of Physical Therapy Science	563	544	96.6
Brazilian Journal of Physical Therapy	216	215	99.5
Physiotherapy	110	107	97.3
European J. of Physical and Rehabilitation Med.	67	63	94.0
Archives of Physical Medicine and Rehabilitation	39	36	92.3
Physiotherapy Canada	31	31	100.0
Hong Kong Physiotherapy Journal	29	29	100.0
Physical Therapy	30	30	100.0
Clinical Rehabilitation	25	24	96.0
Journal of Physiotherapy	16	14	87.5
Disability and Rehabilitation	8	8	100.0
J. Orthopaedic & Sports Physical Therapy	7	6	85.7
Physiotherapy Theory and Practice	6	4	66.7
Physical Therapy in Sport	5	5	100.0
Journal of Hand Therapy	2	1	50.0
Musculoskeletal Science and Practice	1	1	100.0
Total	**1,155**	**1,118**	**96.8**

Counts are definitively coded papers, so two journals are one lower than their retrieval totals in S1 Table (*Archives of Physical Medicine and Rehabilitation* 40→39; *Physiotherapy Theory and Practice* 7→6): these are the two papers that could not be definitively coded.

Online supplemental table S3. Cross-disciplinary evidence on confusing the normality of the *residuals* with the normality of the *raw data*. The first three rows use the identical criterion (among papers reporting a normality test, the proportion applying it to raw data rather than residuals) and are directly comparable; the remaining rows measure related but non-identical quantities and are shown for context.

Discipline	Source	Quantity measured	Estimate
Ecology	Midway and White [1]	Normality tested on raw data, not residuals (papers reporting a normality test; 50 papers)	70%
Biology	Midway and White [1]	Normality tested on raw data, not residuals (papers reporting a normality test; 50 papers)	90%
Physiotherapy	This study	Normality tested on raw data, not residuals (definitively coded papers; 1,155 papers; corrected for classifier error)	89.0%
Clinical psychology	Ernst and Albers [2]	Regression papers unclear about which assumptions were checked (172 regression-using papers, of 893 reviewed)	92%
Clinical psychology	Ernst and Albers [2]	Explicitly tested normality of variables rather than errors (172 regression-using papers, of 893 reviewed)	4%
Experimental psychology	Hoekstra et al. [3]	Analyses with assumptions checked correctly (any technique; 30 researchers)	12%

Online supplemental table S4. Statistical power (%) for the four decision workflows at $d = 0.5$ (10,000 simulations per cell). A = incorrect (pooled); B = per group; C = always Welch; D = always Mann–Whitney.

Distribution	n/group	A: Incorrect	B: Per-group	C: Welch	D: Mann–Whit.
Normal	10	18.6	18.8	18.5	16.8
Normal	20	33.2	33.4	33.3	32.0
Normal	30	48.8	48.9	48.9	46.2
Normal	50	69.3	69.4	69.5	67.1
Skewed	10	21.7	23.1	19.6	22.3
Skewed	20	42.1	43.2	34.3	43.1
Skewed	30	61.4	61.6	49.4	61.7
Skewed	50	83.3	83.3	70.0	83.3
Pain score	10	14.0	14.3	10.6	14.1
Pain score	20	26.0	26.0	17.8	25.9
Pain score	30	36.7	36.7	24.0	36.7
Pain score	50	54.9	54.9	37.0	54.9
Bimodal	10	17.7	17.9	17.1	17.7
Bimodal	20	35.9	36.2	32.9	36.5
Bimodal	30	52.5	52.3	47.7	52.4
Bimodal	50	73.8	73.6	69.5	73.6

Online supplemental table S5. Heteroscedastic scenario: statistical power (%) at $d = 0.5$ when the two normal groups have unequal variances (group B's SD is 2× or 3× group A's; 10,000 simulations per cell). A = incorrect (pooled); B = per group; C = always Welch; D = always Mann–Whitney. The final column is the absolute power gap between the incorrect and per-group workflows.

SD ratio	n/group	A: Incorrect	B: Per-group	C: Welch	D: Mann–Whit.	|A−B|
2×	10	9.5	10.1	9.7	9.5	0.6
2×	20	15.7	16.7	16.6	16.8	1.0
2×	30	21.0	23.0	22.7	22.5	1.9
2×	50	32.9	35.2	35.0	34.2	2.4
2×	100	57.1	60.2	60.5	58.0	3.1
3×	10	7.5	8.0	7.6	8.3	0.5
3×	20	10.3	10.9	10.3	12.2	0.7
3×	30	14.1	14.5	14.0	15.5	0.4
3×	50	19.5	19.9	19.4	20.4	0.4
3×	100	35.1	35.1	34.8	35.1	0.0

Mean Type I error (at $d = 0$, averaged over n) stayed near nominal for the Welch-keeping workflows B and C ($\leq 5.6\%$); it drifted upward for the workflows that fall back on the rank-based test under unequal variances, reaching 5.8% for the incorrect pooled workflow A (at the 3× ratio) and 5.7%/6.6% for always-Mann–Whitney D at SD ratios of 2/3.

Online supplemental figure S1. Power difference between the incorrect (A) and per-group (B) workflows across distributions, sample sizes, and effect sizes. Positive values mean the incorrect workflow had higher power. Most differences lie within ±1.5 percentage points; the few exceptions, all at the largest effect ($d = 0.8$) on skewed and bimodal data, reach at most about 2.2 points.

Online supplemental text S1. Classifier and coding rules

The classifier locates, in each article's full text, every sentence containing a normality cue (a named test (Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, Anderson–Darling, D'Agostino) or the words "normality"/"normal distribution" used in a testing sense, excluding strings such as "abnormality" and "normalised"). It then assigns a single code in the following cascade order (the first matching rule wins; cf. Figure **??**):

UNCLEAR: no normality-test sentence could be located.

MIXED: a residual statement *and* an unqualified pooled statement both occur.

CORRECT: a residual statement occurs (and, by the order above, no unqualified pooled statement).

ACCEPTABLE: a per-group phrase ("each group", "per group", "separately", "both groups", etc.) occurs within 55 characters of a normality cue in the same sentence (this proximity window, and its insensitivity, are discussed in the Methods).

INCORRECT: a normality-test sentence exists but names the data, the variables, the outcome, or the distribution as the object of the test, without a residual or per-group qualifier.

Validation and misclassification correction

The classifier was validated against the author's manual full-text coding of a stratified random sample of 100 articles, coded blind to the classifier's output. Because the papers the classifier coded as *not* incorrect (*correct* or *acceptable*) are rare ($n = 37$ in the whole corpus), all 37 were included as a census, together with a random sample of 63 of the 1,118 papers coded *incorrect* or *mixed*. The resulting confusion between human and classifier codes is shown in Table **??**. Corpus-weighted agreement on the binary correct-versus-incorrect distinction was 90.7% (Gwet's AC1 = 0.89, "almost perfect"; AC1 is used rather than Cohen's κ, which is destabilised by the extreme class imbalance).

The misclassification-corrected prevalence is a design-based, two-phase stratified estimator (a Horvitz–Thompson reweighting of the hand-coded sample to corpus proportions). Writing $N_{\text{neg}} = 37$ and $N_{\text{pos}} = 1{,}118$ for the two strata, the corrected count of incorrect papers is the (exact, census) number of human-incorrect papers in the not-incorrect stratum plus ($N_{\text{pos}}/63$) times the number of human-incorrect papers among the 63 sampled, with analogous reweighting for the other human categories. The primary, conservative estimand (Estimand A) divides the corrected count of incorrect papers by all definitive papers, counting every indeterminate paper as not-incorrect; the secondary estimand (Estimand B) divides instead by the corrected count of *determinable* papers (excluding those a reader judged unclear). Confidence intervals come from a bootstrap that resamples the 63 sampled *incorrect* papers (20,000 replicates); the fully-coded not-incorrect stratum is a census and contributes no sampling uncertainty. This yields 89.0% (95% CI 81.4–95.2%) for Estimand A and 90.4% (95% CI 82.9–96.7%) for Estimand B.

The full validation sample, the author's codes, and all extracted passages are released (where licensing permits) for independent re-checking. Articles whose full text is not openly licensed (the two Elsevier journals and articles recovered as rendered PDFs from non–open-access sources) were classified with the same classifier but, to respect their licence terms, the repository provides only their identifiers and derived codes.

Online supplemental text S2. Language-model classifier prompt

The independent language-model cross-classification (**??**) used the following system prompt, followed by the extracted normality passages as the user message in the form `EXCERPT: """`

... """. Decoding was greedy (temperature 0) for determinism. The exact string, including whitespace, is in `llm_classify.py`.

> You are a meticulous methodologist coding how a study applied a NORMALITY TEST (e.g. Shapiro-Wilk, Kolmogorov-Smirnov). You are given the methods sentences that mention normality. Decide the TARGET the normality test was applied to, and return exactly one code.
>
> **CODES:**
>
> - `CORRECT`: normality assessed on the model RESIDUALS (or 'errors'), e.g. 'Q-Q plot of residuals', 'residuals were normally distributed'.
> - `ACCEPTABLE`: normality assessed SEPARATELY WITHIN EACH GROUP, e.g. 'for each group', 'in both groups separately', 'within groups'. (Not residuals, but per-group.)
> - `INCORRECT`: normality assessed on the POOLED / RAW outcome data or variables, not conditioned on group and not on residuals. Following Midway & White, an UNQUALIFIED statement such as 'the data were tested for normality' or 'normality of the variables was checked' counts as INCORRECT, because the data/variables (not the residuals or within-group observations) are named as the object of the test.
> - `MIXED`: BOTH a residual target (correct) AND an unqualified pooled/raw target (incorrect) are reported.
> - `UNCLEAR`: no normality TEST of an assumption is actually described, or its target cannot be determined. Examples: 'within normal limits', 'normal range', text not about checking a distributional assumption.
>
> **RULES:**
>
> - Use ONLY the excerpt. Do not infer beyond what is written.
> - If residuals are named as the target → CORRECT, even if the word 'data' appears elsewhere.
> - Per-group testing → ACCEPTABLE (do not upgrade to CORRECT).
> - An unqualified 'tested X for normality' with no residuals and no per-group wording → INCORRECT.
> - Reply with STRICT JSON on a single line:
> `{"code": "CORRECT|ACCEPTABLE|INCORRECT|MIXED|UNCLEAR", "reason": "<=20 words"}`

For each paper the model thus returned a `code` and a short free-text `reason` (a justification of at most 20 words). Only the `code` entered the analysis; the `reason` field was retained solely for auditing and is released with the codes.

Online supplemental text S3. Reproducibility

The pipeline comprises: `fetch_corpus.py` (Europe PMC search), `fetch_fulltext.py` (PubMed Central full-text retrieval and passage extraction), `fetch_elsevier.py` (Elsevier full-text API search and retrieval for the two ScienceDirect journals), `fetch_render_pdf.py` (rendered-PDF retrieval for articles outside the PMC open-access subset), `classify.py` (rule-based coding), `validation_analyze.py` (agreement and Gwet's AC1), `run_simulation.R` (power simulation, seed 2025), and `analysis.R` (figures and tables). R version 4.6.0 was used for all statistical analyses.

References

[1] Stephen R. Midway and J. Wilson White. Testing for normality in regression models: mistakes abound (but may not matter). *Royal Society Open Science*, 12(4):241904, 2025. doi: 10.1098/rsos.241904.

[2] Anja F. Ernst and Casper J. Albers. Regression assumptions in clinical psychology research practice — a systematic review of common misconceptions. *PeerJ*, 5:e3323, 2017. doi: 10.7717/peerj.3323.

[3] Rink Hoekstra, Henk A. L. Kiers, and Addie Johnson. Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3:137, 2012. doi: 10.3389/fpsyg.2012.00137.