Normality Testing in Physiotherapy Research: A Full-Text Bibliometric Audit of a Common but Low-Stakes Error

Jürgen Degenfellner

Institute of Physiotherapy, Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland

degn@zhaw.ch ORCID 0000-0001-6886-4583

Abstract

Background. Normality tests such as the Shapiro–Wilk or Kolmogorov–Smirnov test are widely used to decide between parametric and non-parametric analyses. The assumption they concern applies to the residuals of the fitted model, not the raw outcome data. A recent audit reported that 70% of ecology and 90% of biology papers test the raw data instead, but no comparable audit exists for physiotherapy, a field with small samples and frequently non-normal outcomes.

Methods. A full-text bibliometric audit was performed of articles published in 16 physiotherapy and rehabilitation journals between 2015 and 2024 that reported a named normality test, retrieved through Europe PMC and the Elsevier full-text API. A deterministic rule-based pipeline coded each paper's normality-test target as residuals (correct), per group (acceptable), or pooled raw data (incorrect). The classifier was validated against the author's manual coding of 100 papers, the prevalence was corrected for the measured classification error with a design-based estimator and bootstrap interval, and an independent language-model re-coded the corpus. The power of four decision workflows was also simulated across four distributions and sample sizes from 10 to 100 per group.

Results. The search returned 1,157 articles, of which 1,155 could be definitively coded. The classifier coded 1,118 of these (96.8%) as applying the normality test to the raw data rather than the residuals; correcting for the classifier's labelling error gave a conservative prevalence of incorrect application of 89.0% (95% CI 81.4–95.2%), rising to 90.4% among determinable papers. Only 12 papers (1.0%) tested residuals. The error exceeded 85% in every journal with at least ten coded papers and showed no material change over the decade. The language-model re-coding reproduced this and matched the manual validation even more closely than the rule-based classifier. In the simulation, the power cost of the error never exceeded about three percentage points; the test eventually chosen mattered far more than the residual-versus-raw distinction.

Conclusions. Misapplication of normality tests is the norm in physiotherapy research, in line with the rates previously reported for ecology and biology. The power consequences are small, but correct practice is essentially free; a simple check is provided for authors, reviewers and educators.

Keywords: normality test; Shapiro–Wilk test; model assumptions; bibliometric review; physiotherapy; simulation study

Background

Parametric statistical tests such as the independent-samples t-test, paired t-test, and analysis of variance (ANOVA) rest on the assumption that the model errors follow a normal distribution. In practice, many researchers, particularly those outside the statistical disciplines, operationalise this assumption by applying a formal normality test (e.g., the Shapiro–Wilk or Kolmogorov–Smirnov test) before deciding whether to use a parametric or a non-parametric procedure [1, 2].

This instinct has a sound mathematical core. Because any linear combination of independent normal variables is again exactly normal [3, Cor. 4.6.10], normality of the observations guarantees normality of the very quantity a t-test depends on. For a one-sample test with $X_1, \ldots, X_n$ independent and $N(\mu, \sigma^2)$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \;\sim\; N\!\left(\mu,\; \frac{\sigma^2}{n}\right)$$

exactly; and for two independent groups of sizes m and n with $X_i \sim N(\mu_X, \sigma^2)$ and $Y_j \sim N(\mu_Y, \sigma^2)$,

$$\bar{X} - \bar{Y} \;\sim\; N\!\left(\mu_X - \mu_Y,\; \sigma^2\left(\tfrac{1}{m} + \tfrac{1}{n}\right)\right),$$

again *exactly*, at any sample size and with no appeal to the central limit theorem: it is the closure of the normal family under such linear combinations. Checking the data for normality is therefore a reasonable instinct: normal data are a *sufficient* condition for the t-test to be exact. As the next paragraphs make precise, however, the condition properly attaches to the model *residuals* rather than the raw outcome, and it is sufficient but not necessary.

The independent-samples t-test, the paired t-test, and ANOVA are not separate methods but special cases of the linear model: the two-group t-test is a linear regression on a single binary indicator, one-way ANOVA a linear regression on a categorical predictor, and the paired t-test an intercept-only linear model on the within-pair differences (equivalently, a linear model with a subject-specific intercept), so all of them share the same error assumptions [4, 5]. Writing y_i for the outcome:

$$\begin{aligned}
\text{two-group } t\text{-test:} &\quad y_i = \beta_0 + \beta_1\, g_i + \varepsilon_i, \quad g_i \in \{0,1\};\\
\text{one-way ANOVA } (k \text{ groups}): &\quad y_i = \beta_0 + \sum_{j=2}^{k} \beta_j\, d_{ij} + \varepsilon_i, \quad d_{ij} \in \{0,1\};\\
\text{paired } t\text{-test:} &\quad \delta_i = \beta_0 + \varepsilon_i, \quad \delta_i = y_{i2} - y_{i1};
\end{aligned}$$

with $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ in every case: a single shared error term. Here β_0 is the reference-group mean (group 1; in the paired model, the mean within-pair difference), while β_1 and the β_j are the group contrasts: β_1 is the fixed, unknown difference between the two group means, and each β_j the difference of group j from the reference. The normality assumption of the linear model, however, concerns the *residuals* of the fitted model, not the raw outcome variable. For a two-group design this means that normality should be evaluated either within each group separately (which, for balanced designs, is equivalent to testing residuals) or directly on the residuals after model fitting [5, 6]. Testing the *pooled* sample of all observations for normality, as is frequently reported, examines the *marginal* distribution of the outcome, whereas the linear model assumes only *conditional* normality (of the errors,

i.e. within each group; 7) and makes no assumption at all about that marginal distribution. Such a test therefore inspects a quantity the model never assumes: a difference in group means alone can make the pooled distribution non-normal even when every group is perfectly normal, so the result says nothing about the residual assumption.

Despite this clear methodological guidance, a recent bibliometric audit by Midway and White [8] found that 70% of ecology papers and 90% of biology papers (50 papers per field, published in 2022) applied normality tests to raw data rather than residuals. Using simulation, the same authors showed that the power cost of this error is generally small: when residuals genuinely violate normality, switching to the Mann–Whitney U test improves power by only about 3–4 percentage points. Hence their memorable subtitle: mistakes abound, but may not matter.

This relative unimportance of the normality assumption is itself well established. The inferential validity of the t-test, ANOVA and linear regression depends far more on the independence of observations and on homoscedasticity than on the normality of the errors; by the central limit theorem these procedures are asymptotically valid for essentially any error distribution, so that, in the words of Lumley et al. [9], their usefulness "comes from the fact that in large samples they are valid for any distribution." Monte-Carlo studies, some reviewed by Norman [10] as far back as the 1930s, repeatedly show that parametric methods retain close to nominal Type I error and near-optimal power under pronounced non-normality [10, 11], and that the transformations or preliminary normality tests applied to "repair" non-normal outcomes are frequently unnecessary and may themselves bias estimates or reduce power [2, 12]. Of the classical linear-model assumptions, normality is therefore widely regarded as the least consequential. This matters for how the results should be read: because the assumption itself matters so little, misdirecting the *test* of that assumption is unlikely to have altered many published conclusions. The error nonetheless remains common, easy to spot, and easy to correct, which is what makes it worth measuring.

To the author's knowledge, no equivalent audit has been conducted in physiotherapy or rehabilitation research. This gap matters because outcome measures in physiotherapy (range of motion, functional scores, and especially pain measured on numeric rating scales, a distribution that is typically zero-inflated and right-skewed [13]) are frequently *non-normally distributed*, which makes the choice between parametric and non-parametric procedures potentially consequential. A preliminary normality test is, moreover, a weak instrument for guiding that choice: even the Shapiro–Wilk test, the most powerful of the commonly used normality tests, has little power to detect genuine departures from normality at the sample sizes common in this field, so that a "pass" carries very little information about the underlying distribution [1]. Gating the choice of analysis on such a test is therefore doubly questionable in physiotherapy: the test is typically applied to the wrong quantity (the raw data rather than the residuals), and it has little power to answer even the narrower question it nominally addresses.

The aims of this study were therefore:

1. to estimate the prevalence of incorrect normality testing in physiotherapy and rehabilitation journals through a full-text bibliometric audit;

2. to place this estimate alongside published rates and related evidence from other empirical disciplines [8, 14, 15]; and

3. to assess, via simulation, the practical impact of the incorrect workflow on statistical power under data-generating scenarios typical of clinical physiotherapy research.

Methods

This study is a full-text bibliometric audit, a form of meta-research. No EQUATOR reporting guideline maps cleanly onto such a design, which is neither a clinical study nor a systematic review; the reporting below nonetheless follows general meta-research practice, by pre-specifying the coding criterion, describing the full retrieval and coding pipeline, validating the automated coding against human coding, and releasing the data and code for independent re-checking.

Correct versus incorrect application of normality tests

For a two-group comparison (e.g., intervention vs. control), the parametric t-test assumes that the observations within each group are drawn from a normal distribution. Equivalently, the residuals $e_i = y_i - \hat{\mu}_g$ (where $\hat{\mu}_g$ is the estimated group mean) should be normally distributed. Testing normality *per group* (i.e., running separate Shapiro–Wilk tests for each group), or testing the residuals directly, is therefore a valid way to check the assumption. In the coding scheme below, the label *correct* is nonetheless reserved for the residual check and per-group testing called *acceptable*, since the latter coincides with a residual check only for balanced designs, and even then not as an identity of p-values: running a separate Shapiro–Wilk test in each group rejects somewhat more often than a single test of the pooled residuals, through multiplicity alone. The two are equivalent only in the sense that matters here, in that they operate on the same within-group centred values and are equally immune to the between-group mean shift that inflates the pooled raw-data test. With unequal group sizes the larger group dominates the pooled residuals and dilutes a non-normal smaller group, so even this correspondence breaks down. Strictly, normality of the individual observations is *sufficient* but not *necessary*: what the test ultimately requires is that the sampling distribution of the mean difference $\bar{x} - \bar{y}$ be (approximately) normal, which by the central limit theorem holds for moderate sample sizes even when the observations themselves are not normal [3, Central Limit Theorem, Thm. 5.5.14]. Normality of the residuals guarantees this exactly, at any sample size, and is the assumption as usually stated; the point here is only about *where* that assumption is checked (residuals or within-group observations, not the pooled raw data), not that it is indispensable.

A preliminary normality test is also an imperfect tool *irrespective* of where it is applied: underpowered in the small samples common in physiotherapy (as noted above) and, conversely, oversensitive in large ones, flagging trivial departures and prompting an unnecessary switch to a non-parametric procedure [1, 2]. Gating the choice of analysis on whether such a test crosses a fixed significance threshold is itself an instance of the dichotomous, cutoff-based reasoning that has been widely criticised [16]. Indeed, a body of work shows that such two-stage pre-testing is at best unhelpful: pre-testing the assumptions of the two-sample t "does not pay off" [17], a preliminary goodness-of-fit test does not validate the one-sample t [18], and preliminary tests of variance can degrade rather than protect the subsequent inference [19]. For this reason several authors recommend against routine preliminary testing altogether, in favour of graphical inspection of residuals or of methods that are robust by design [18, 20]; more broadly, it has been argued that the (conditional) distribution is better assessed through subject-matter reasoning, graphics and model comparison than through p-value-based tests of the assumption [7]. The present audit therefore concerns where researchers *direct* a test they have already chosen to perform;

it does not imply that performing the test was itself advisable.

In contrast, testing the *entire pooled sample* does not condition on group and so inspects the marginal outcome distribution, which a between-group mean difference can render non-normal even when each group is normal.

Throughout, this paper builds on the residuals-versus-raw-data criterion of Midway and White [8], their core distinction between testing normality on the model residuals (correct) and on the whole/raw data (incorrect), and refines it into the following five-level coding scheme; the intermediate *acceptable*, *mixed* and *unclear* categories are an elaboration introduced here:

Correct: normality tested on model residuals.

Acceptable: normality tested separately for each group (equivalent to testing residuals for balanced two-group designs).

Incorrect: normality tested on the pooled raw data or the outcome variable without conditioning on group.

Mixed: both a residual (correct) target and an unqualified pooled (incorrect) target reported in the same paper.

Unclear: no sentence describing a normality test could be located, or its target could not be determined.

For the primary outcome, *mixed* is grouped with *incorrect* (a paper that anywhere tests the pooled data is counted as committing the error). Text matched by the search that is not in fact an assumption check (e.g. "normal range" or "within normal limits") provides no determinable normality-test target and is coded *unclear*; both the classifier and the human coder use the same five levels, so this judgement is folded into *unclear* rather than given a separate label. Following the operational rule of Midway and White [8], an unqualified statement such as "the data were tested for normality using the Shapiro–Wilk test" is coded as *incorrect*, because the data, rather than the residuals or the within-group observations, are named as the object of the test.

Scope of the residuals-versus-raw criterion. The residuals-versus-raw-data distinction is sharpest for the linear model (the t-test, ANOVA, ANCOVA and linear regression), where the normality assumption attaches unambiguously to the model errors. Normality tests are occasionally reported in support of analyses outside the linear model, where the residuals-versus-raw distinction is less clean: correlation rests on bivariate normality, factor-analytic and structural-equation models on multivariate normality of the items, and grouped procedures such as MANOVA and discriminant analysis on multivariate normality within groups [21, 22]. A further group of settings matters more still: there, testing the marginal distribution of the *raw data* is the correct procedure rather than an error: estimating normative reference intervals (where the population distribution of the measurement is itself the quantity of interest) [23], constructing Bland–Altman limits of agreement (which assume the within-subject differences are normal) [24], and statistical process control [25]. Because the scheme codes the *target* of the normality test and not its appropriateness, a paper that legitimately checks raw-data normality for one of these purposes is nonetheless coded *incorrect*; this inflates the incorrect rate and partly offsets the opposite, conservative bias introduced by paired designs. Both kinds of context are,

however, a small minority of the comparative, group-difference literature audited here. To keep the primary estimate directly comparable to Midway and White [8], their convention is followed and the *target* of the normality test (residuals, per group, or raw data) is coded for every paper that reports one, regardless of the downstream analysis. The robustness of this estimate is then probed with a sensitivity analysis restricted to papers whose normality test is clearly attached to a linear-model/group-comparison analysis, with the non-linear-model contexts (chiefly correlation) reported separately. The classifier records, for each paper, which analysis is named in the vicinity of the normality sentence, which makes this restriction reproducible.

Journal selection and search strategy

A purposive sample of 16 journals was assembled to span the physiotherapy and rehabilitation literature that publishes original quantitative research: the field's core discipline-specific journals (for example the *Journal of Physiotherapy*, *Physiotherapy*, *Physical Therapy*, the *Journal of Orthopaedic & Sports Physical Therapy* and the *Journal of Physical Therapy Science*) together with the major general rehabilitation journals in which physiotherapy research regularly appears (for example *Archives of Physical Medicine and Rehabilitation*, *Clinical Rehabilitation* and *Disability and Rehabilitation*). The list was not derived from a formal bibliometric sampling frame; it is a deliberately broad but non-exhaustive selection of widely-read titles, and the consequences of this choice are considered in the Limitations. For brevity, "physiotherapy" is used as a shorthand for this physiotherapy-and-rehabilitation literature throughout.

Normality tests are almost always described in the statistical methods section rather than the abstract, so a full-text search is required (a preliminary PubMed title/abstract search of the same journals returned only 7 articles). Two complementary full-text sources were therefore searched. For the 14 journals whose full text is indexed by *Europe PMC*, a free, open repository of life-science literature that mirrors PubMed Central and exposes a programmatic search-and-retrieval interface (a REST API: a web service that returns search results and article records to a script over standard HTTP requests), that API was queried (11 June 2026). For the two remaining journals, the *Journal of Physiotherapy* and *Physiotherapy*, whose full text is available only through ScienceDirect, the Elsevier full-text API was used under the publisher's text-and-data-mining terms (16 June 2026).

In both sources the query combined (i) any of a set of normality-test terms, joined with the boolean OR so that an article qualified if it contained at least one of them (`"Shapiro-Wilk"`, `"Kolmogorov-Smirnov"`, `"Lilliefors"`, `"Anderson-Darling"`, `"D'Agostino"`, `"normality test"`, `"test for normality"`, `"tested for normality"`, `"assess normality"`, `"check normality"`); (ii) the journal; and (iii) the publication-date range 2015–2024. The most recent complete ten-year window was used: 2025 was deliberately left out because, at the search date (June 2026), that year was not yet completely or evenly indexed across the 16 sources, so including a partial, unevenly covered year would have biased both the prevalence and the temporal trend. After de-duplication, the search returned 1,173 unique articles; 16 of these carried an online-first publication year of 2025 (outside the study window) and were excluded, leaving **1,157** articles across the 16 journals for analysis (Table 1). Full text was retrieved for analysis from PubMed Central (E-utilities); for the two ScienceDirect journals from the Elsevier API; and for articles not contained in the PMC open-access subset (all of *Physiotherapy Canada* and a minority of articles in five other journals, chiefly the *Brazilian Journal of*

Physical Therapy), from the freely rendered PDF served by Europe PMC (parsed with `pdftotext`). After this recovery, full text was available for 1,151 of the 1,157 articles; for the remaining six only the abstract could be retrieved, four of which still named the test's target and were coded from the abstract, leaving just two articles that could not be coded at all. The corpus is dominated by the high-volume open-access journals (chiefly the *Journal of Physical Therapy Science* and the *Brazilian Journal of Physical Therapy*).

The full texts obtained under the Elsevier API terms and those recovered as rendered PDFs (free to read but not all openly licensed) were processed locally by the rule-based classifier only. For content that is not openly licensed, only the article identifiers and the derived codes, not the full text or extracted passages, are included in the public data and code repository (the permanent deposit described under Declarations), from which any reader with the same journal access can reproduce the analysis.

All retrieval steps use publicly documented interfaces and can be repeated by an independent researcher: the Europe PMC and PubMed Central services and the free rendered-PDF endpoint require no credentials, and the Elsevier full-text step requires only a (free) API key. The one access dependency is that downloading the full text of subscription articles (here the two ScienceDirect journals) requires the researcher's institution to hold the corresponding subscription; without it, the same query returns the open-access subset only. The retrieval scripts, with the exact queries and date stamps, are provided in the repository.

Table 1: 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.

Coding procedure and its validation

Because manual full-text coding of more than a thousand articles is labour-intensive, a transparent, deterministic, rule-based text-classification pipeline was used (implemented in Python; full source provided in the public data and code repository). For each article, the full text was parsed and every sentence mentioning a normality test was extracted. The classifier then assigned a code by inspecting these sentences: a sentence linking the test to *residuals* yielded *correct*; a sentence in which a per-group phrase (e.g. "each group", "separately") occurred within 55 characters of the normality cue yielded *acceptable*; any other normality-test sentence yielded *incorrect*; and articles for which no normality-test sentence could be located were coded *unclear*. The rules operationalise the coding scheme defined above exactly; Figure 1 summarises the decision logic. The 55-character proximity window is a deliberately loose heuristic (roughly half a clause), chosen so that the per-group phrase plausibly modifies the test rather than sitting elsewhere in a long sentence; varying it from 40 to 150 characters changes the *acceptable* count by only a handful of papers (21 to 30 of 1,155) and moves the corrected prevalence by at most 0.6 percentage points (89.0% to 89.6%), because the misclassification correction targets the human gold standard rather than the classifier's raw labels.

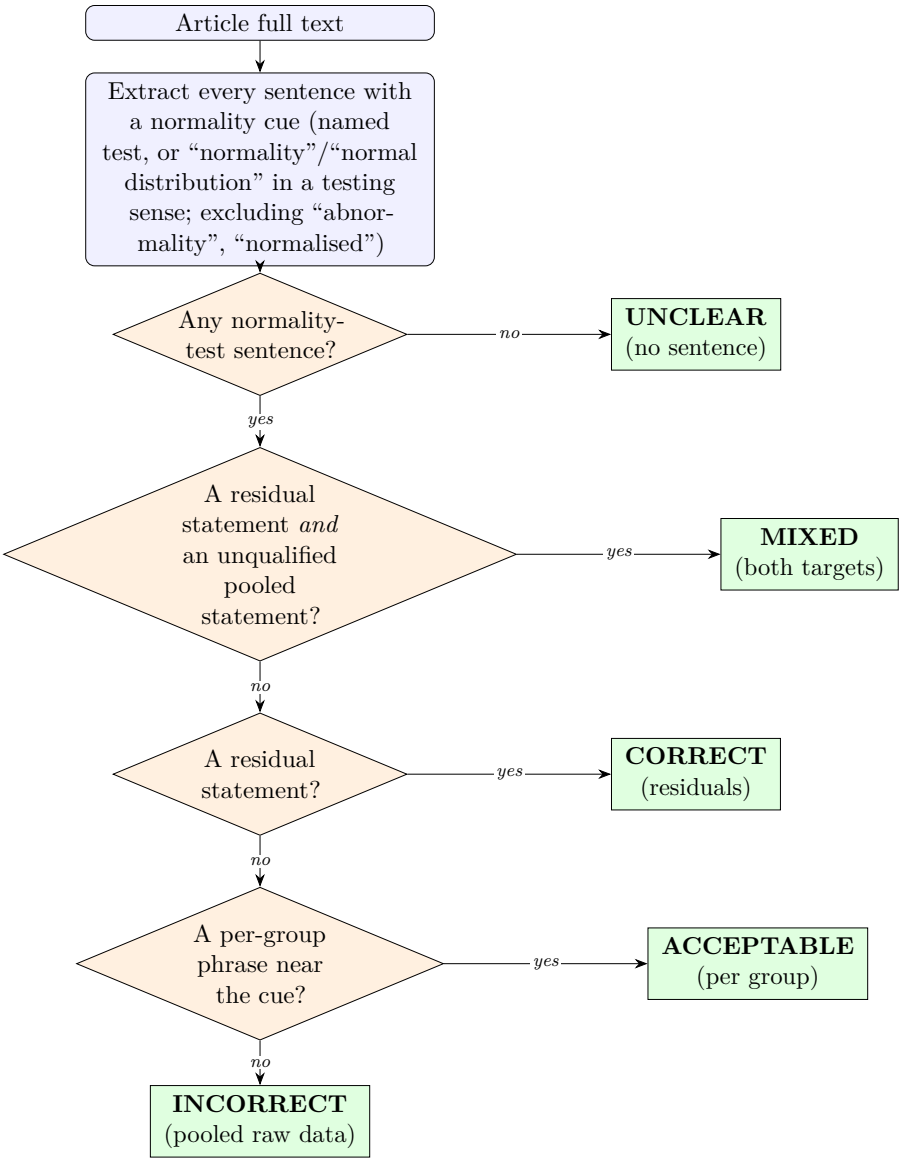

Figure 1: Decision logic of the rule-based classifier. Each article's full text is reduced to its normality-test sentences, which are then resolved to a single code in cascade order: an article with no normality-test sentence is coded *unclear*; otherwise a residual statement *together with* an unqualified pooled statement → *mixed*; otherwise a residual statement → *correct*; otherwise a per-group phrase near the cue → *acceptable*; otherwise → *incorrect*. A per-group phrase ("each group", "separately", "both groups") counts when it falls within 55 characters of the normality cue in the same sentence.

To validate the classifier and to correct the prevalence for its classification error, the author coded a stratified random sample of 100 articles by hand, blind to the classifier's output. The classifier's rules (the keyword lists and the 55-character proximity window) had been fixed before this hand-coding and were not afterwards adjusted to match it, so the agreement figures are not inflated by in-sample tuning of the rules to the validation labels. Because the papers the classifier coded as *not* incorrect (*correct* or *acceptable*) are rare in the corpus ($n = 37$), all of them were included, together with a random sample of 63 of the 1,118 papers coded *incorrect* or *mixed* (mixed being grouped with incorrect for the primary outcome, as defined above); this two-phase design estimates both directions of classifier error efficiently. The validation was capped at 100 hand-coded papers, a budget feasible for

a single blinded coder; with the 37 not-incorrect papers taken as a census, the remaining 63 fell in the dominant *incorrect* stratum. This second-phase sample is the only source of sampling uncertainty in the corrected estimate and sets the width of its bootstrap interval (about ±7 percentage points for Estimand A); a larger second-phase sample would narrow that interval, at proportional hand-coding cost. From the resulting confusion between the human and the classifier codes, two quantities were computed: (i) the corpus-weighted agreement on the binary correct-versus-incorrect distinction, summarised by Gwet's AC1 [26], a chance-corrected agreement coefficient that, by design, remains stable under the extreme class imbalance present here, where Cohen's κ is destabilised by the prevalence paradox [27], and (ii) a design-based, misclassification-corrected prevalence of incorrect application (a design-based, two-phase stratified Horvitz–Thompson estimator [28]), with a 95% confidence interval from a bootstrap that resamples the sampled *incorrect* stratum; the fully-coded not-incorrect stratum is a census and contributes no sampling uncertainty. For each paper the classifier also recorded which normality test was named and the statistical context. All passages and codes are released for independent re-checking.

Independent LLM cross-classification. The rule-based classifier and the human gold standard share a single annotator (the author) for the human side, and the classifier relies on hand-written keyword heuristics. To guard against both (an idiosyncratic rule set and a single human reader), the entire corpus was re-classified a third time with a large language model, used as an independent second coder. A locally hosted, open-weight model (Qwen2.5-32B-Instruct [29], 4-bit quantisation, run with the MLX framework [30] under greedy decoding so that the output is deterministic and exactly reproducible) received the same extracted normality passages and the same five-level coding scheme, with no access to either the rule-based code or the human code, and returned one code per paper. The system prompt is given in full in the Appendix (LLM classifier prompt). Because the language model and the rule-based classifier share no coding logic, and the model saw neither the rule-based nor the human codes, their agreement is evidence that the classification is not an artefact of one particular rule set. Two caveats temper this independence and it is not overstated here: both methods read the *same* machine-extracted passages, so the upstream sentence-extraction step is a shared point of failure (a missed residual sentence would mislead both); and because one class dominates the corpus, high *raw* agreement is partly mechanical, which is why the chance-corrected AC1 is reported rather than raw agreement for the human comparison. The model, prompt and parser are released with the other code.

Statistical analysis

The primary quantity was the proportion of papers applying normality tests incorrectly (pooled raw data), restricted to papers with a definitive code (*correct*, *acceptable*, or *incorrect*). Because the audit retrieved *all* qualifying articles in the indexed literature rather than a random sample of them, the classifier's count over the corpus is a census and carries no sampling uncertainty; the relevant source of error is instead the classifier's misclassification of individual papers. The raw classifier proportion is therefore reported as a measured quantity and, as the primary estimate, the misclassification-corrected prevalence obtained from the human validation (described above), with its bootstrap confidence interval. This corrected prevalence is reported under two estimands. *Estimand A* (the *primary*, conservative estimate) is the proportion incorrect among *all* definitively

coded papers, counting every paper a human reader judged genuinely indeterminate as not-incorrect; it is therefore conservative with respect to the indeterminate cases under these coding rules. *Estimand B* restricts the denominator to papers whose target a human reader could determine (excluding the indeterminate), and so will be slightly higher. Estimand A is taken as primary. (The boundary cases discussed later push the rate in both directions, so this conservativeness applies within the coding framework, not as an unconditional lower bound.) The change over calendar time was summarised descriptively, as the proportion of incorrect applications in each publication year. Journal-level proportions were likewise compared descriptively. As a robustness check against the corpus being dominated by a few high-volume journals, the overall incorrect rate was also recomputed weighting each journal equally (the unweighted mean of the per-journal rates) rather than each paper.

The corpus retrieval and the rule-based text-classification pipeline were implemented in Python 3 [31]. All statistical analyses and the simulation study were conducted in R version 4.6.0 [32] using the `tidyverse` [33]; the random seed for the simulation was fixed at 2025.

Simulation study

The simulation of Midway and White [8] was extended to physiotherapy-relevant scenarios. Four decision workflows were compared:

Workflow A (Incorrect): Test normality on pooled raw data; if normal ($\alpha = .05$): use t-test; else: Mann–Whitney U.

Workflow B (Per-group): Test normality per group; if both groups pass: use t-test; else: Mann–Whitney U.

Workflow C (Pragmatic): Always use Welch t-test (no preliminary test).

Workflow D (Robust): Always use Mann–Whitney U.

Data were generated for four distribution types: (1) *Normal* ($\mu = 0$, $\sigma = 1$); (2) *Skewed* (gamma, shape = 2); (3) *Pain-score-like* (a zero-inflated 0–10 numeric rating scale: 34% structural zeros plus a right-skewed remainder, calibrated to published clinical pain data: mean ≈ 3, about a third of values zero and about 13% ≥ 7 [13]); (4) *Bimodal* (mixture of two normals). Group sizes ranged from $n = 10$ to $n = 100$ per group and Cohen's d from 0 (false-positive check) to $d = 0.8$. The effect was introduced as a pure *location shift*: the second group's distribution was the first group's shifted by d standard deviations, with shape unchanged. This is deliberate. The t-test and the Mann–Whitney U test do not in general test the same hypothesis: the t-test concerns the difference in *means*, whereas the Mann–Whitney test concerns stochastic ordering, $\Pr(X < Y) \neq \frac{1}{2}$, and the two can even point in opposite directions when the shape of the distribution also changes between groups [34]. Under a location-shift alternative, however, both hypotheses coincide (mean shift, median shift and stochastic dominance move together), so comparing the power of workflows that may end up applying either test is well defined [34]. This scope condition matters: the power comparison speaks to the common situation of a location shift, not to settings where groups differ in shape, where a "significant" t-test and a "significant" Mann–Whitney test would answer materially different questions.

To probe the case in which the pooled-versus-residual distinction is most likely to bite, a *heteroscedastic* scenario was added: both groups normal, but the second with two or

three times the standard deviation of the first (SD ratio $r \in \{2, 3\}$, i.e. a variance ratio of 4 or 9), the mean shift expressed in control-group SD units. Pooling two normals of unequal variance can itself appear non-normal, so the incorrect pooled normality test is more likely to misfire there than under equal variances, and workflow A may switch to the rank-based test while the per-group workflow B correctly retains the t-test. The downstream t-test was Welch's throughout, which isolates the effect of *where* normality is tested from the separate matter of unequal variances. Each cell was replicated 10,000 times, giving a Monte-Carlo standard error of at most 0.5 percentage points (at a rejection rate of 0.5), so the small between-workflow differences reported below are resolved rather than noise. Power was estimated as the proportion of simulations in which the final test rejected at the conventional (but arbitrary) $\alpha = .05$ threshold.

Declaration of generative AI

The primary classification pipeline is a deterministic, rule-based algorithm. As an independent robustness check, the corpus was additionally re-classified with an open-weight large language model (Qwen2.5-32B-Instruct, run locally under greedy decoding; detailed in the Results); this use is part of the analysis, is described in full in the Methods, and is fully reproducible from the released code. During the preparation of this work the author also used Claude Code (Anthropic) for assistance with the analysis code, and the Claude Opus large language model (Anthropic) for language editing. After using these tools the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Results

Corpus and coding

After restricting to full publication years 2015–2024, the corpus comprised 1,157 unique articles from 16 journals (Table 1). The classifier coded 1094 (94.6%) as *incorrect*, 24 (2.1%) as *mixed/incorrect*, 25 (2.2%) as *acceptable*, 12 (1.0%) as *correct*, and only 2 (0.2%) as *unclear*. The Shapiro–Wilk test was the most frequently named test (762 articles, 65.9%), followed by the Kolmogorov–Smirnov/Lilliefors test (391 articles, 33.8%); the Anderson–Darling and D'Agostino tests appeared twice each. The most common statistical context was the t-test ($n = 433$), followed by Mann–Whitney/Wilcoxon ($n = 238$), ANOVA ($n = 215$), correlation ($n = 181$), regression ($n = 73$), and ANCOVA/mixed models ($n = 33$). These contexts are not mutually exclusive: context is attached by keyword matching in the extracted passages, so a paper reporting more than one analysis is counted under each (and the totals therefore exceed the number of papers), and the attribution, like the proximity rule used for the sensitivity analysis below, is approximate.

Prevalence of incorrect normality testing

Of the 1,155 papers that could be definitively coded (the 1,157 above minus the 2 *unclear*), the classifier coded **1,118 (96.8%)** as applying the normality test incorrectly to pooled raw data or the outcome variable (the 1,094 *incorrect* codes above together with the 24 *mixed* papers, which also test the raw data); 25 (2.2%) as using the acceptable approach

of testing each group separately, and 12 (1.0%) as applying the test correctly to model residuals (Figure 3).

This raw classifier figure is a census of the indexed literature, but it does not account for the classifier's own labelling error. Correcting for that error using the human validation (Table 2) lowers the estimate: the misclassification-corrected prevalence of incorrect application was **89.0% (95% CI 81.4–95.2%)** under the primary, conservative estimand (Estimand A, which counts every paper a human reader judged indeterminate as not-incorrect), rising to **90.4% (95% CI 82.9–96.7%)** among papers whose target a human reader could determine (Estimand B; Figure 2). The correction moves the estimated prevalence down from the raw 96.8% by about eight percentage points but leaves the qualitative conclusion unchanged: incorrect application remains the overwhelming norm.

Concretely, the corrected count is a two-phase stratified (Horvitz–Thompson) estimate that weights each hand-coded paper by the inverse of its probability of being hand-coded. The 37 papers the classifier called *not*-incorrect were hand-coded in full (a census, weight 1), whereas each of the 63 papers sampled from the 1,118 it called *incorrect* or *mixed* represents $1118/63 \approx 17.7$ corpus papers. On full reading the census split into 17 incorrect and 20 not, and the sample into 57 incorrect, 5 not and 1 unclear (Table 2), so the estimated number of incorrectly-analysed papers is

$$\hat{D} \;=\; \underbrace{17}_{\text{census}} \;+\; \frac{1118}{63} \times \underbrace{57}_{\text{sample}} \;\approx\; 1{,}029,$$

the same inverse-probability reweighting giving about 108 correct-or-acceptable and 18 unclear papers. This corrected count is 89.0% of all 1,155 definitive papers (Estimand A) and 90.4% of the $\approx$ 1,137 determinable papers, excluding the estimated unclear (Estimand B). The percentages are computed from the unrounded design-based estimates, so they need not equal a ratio of the rounded counts shown here.

The remainder of this section reports the per-journal and temporal patterns using the classifier codes, which give the finest-grained picture; the corrected prevalence above is the quantity carried into the cross-field comparison and discussion.

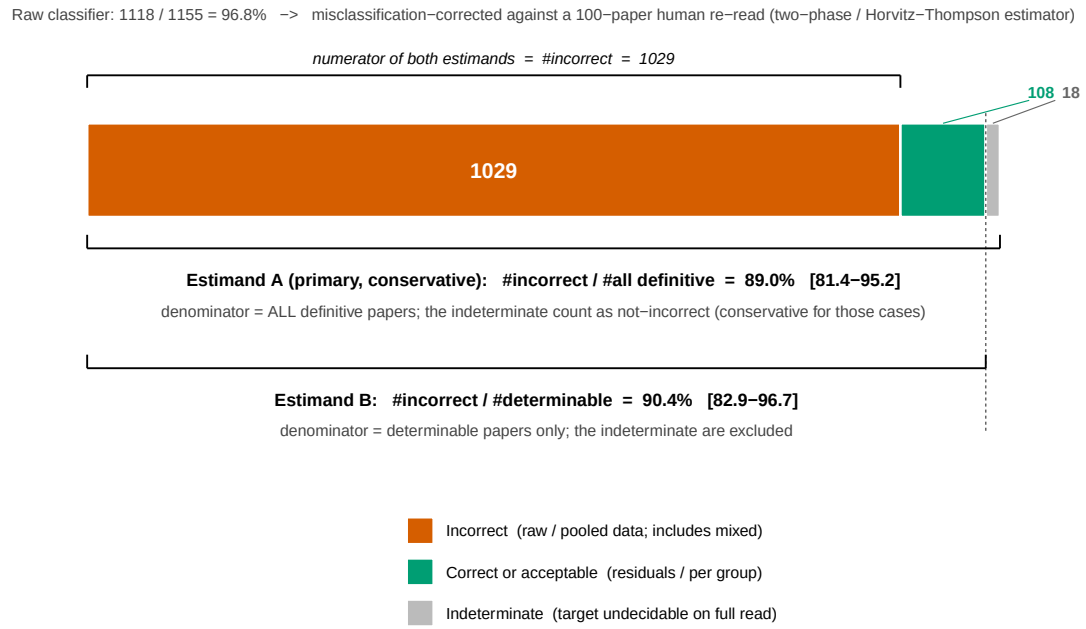

Figure 2: From the raw classifier count to the corrected prevalence. The misclassification correction (against a 100-paper human re-read; two-phase / Horvitz–Thompson estimator) maps the raw 96.8% onto the design-based composition of the 1,155 definitively-coded papers (incorrect; correct or acceptable; indeterminate). The numerator is shared (the incorrect papers); Estimands A and B differ only in whether the indeterminate papers enter the denominator. The incorrect category includes the small *mixed* stratum (papers naming both a residual and a pooled target); indeterminate papers are those whose target a human reader could not determine at all (coded *unclear*), not mixed. Counts are rounded design-based estimates and the percentages are computed from the unrounded values.

Sensitivity to analytic context. Because the residuals-versus-raw criterion is cleanest for the linear model (see Methods), the prevalence calculation was repeated on the subset of papers whose normality test sits next to a clearly identifiable linear-model or group-comparison analysis (t-test, ANOVA, ANCOVA or regression). The classifier could attach such a context to 660 of the 1,155 definitively coded papers (57%); among these the incorrect rate was 96.2% (635/660), essentially identical to the raw figure for the whole corpus (96.8%). The papers whose nearest context was instead a Pearson correlation (the most common non-linear-model case, where the appropriate assumption is bivariate normality rather than residual normality) were if anything coded incorrect slightly more often (99/101; 98.0%). The overall estimate is therefore not an artefact of pooling linear-model analyses with correlation.

Sensitivity to journal weighting. Because two open-access titles contribute most of the corpus (Table 1), the overall figure was confirmed not to be an artefact of their weight. Weighting each *journal* equally rather than each paper, the unweighted mean of the per-journal incorrect rates (classifier codes), gives 96.3% across the ten journals with at least ten coded papers (median 96.9%, range 87.5–100%), and 91.6% across all sixteen journals, several of which contribute only a handful of papers. The near-universal rate is thus a property of the indexed field, not of a single high-volume journal.

Validation of the correction. The human validation behind the misclassification correction (the 89.0%/90.4% above) is summarised in Table 2. Corpus-weighted agreement between classifier and human on the binary correct-versus-incorrect distinction was 90.7%, corresponding to a Gwet's AC1 of 0.89 [26], "almost perfect" on the conventional Landis–Koch benchmark scale (originally defined for κ) [35]. AC1 is used rather than Cohen's κ because, under the extreme class imbalance here (the incorrect category covers the great majority of papers), κ is deflated by the well-known prevalence paradox [27] and badly understates agreement that is in fact excellent (here $\kappa = 0.24$ despite 90.7% raw agreement, the paradox in plain view); AC1 is constructed to be robust to exactly this situation. The principal disagreement was the classifier's *acceptable* code: of 25 papers it labelled acceptable, a human reader judged 16 to be incorrect (Table 2), almost always because a per-group phrase near the normality cue was coincidental rather than a genuine within-group test, or because the design was paired.

Table 2: Validation: classifier codes (rows) against the author's manual codes (columns) for the 100-paper stratified sample (all 37 papers the classifier coded *correct*/*acceptable*, plus a random 63 of the 1,118 it coded *incorrect*). Counts are raw sample counts, not corpus-weighted; the corrected prevalence in the text reweights the two strata to corpus proportions.

Classifier ↓ / Human →	Correct	Accept.	Incorr.	Mixed	Unclear
Correct	11	0	1	0	0
Acceptable	1	8	16	0	0
Incorrect	1	4	56	0	1
Mixed	0	0	0	1	0

Independent LLM cross-classification

Re-coding the whole corpus with the language model, a method that shares no coding logic with the rule-based classifier (though both read the same extracted passages), gave a closely concordant picture. The LLM agreed with the rule-based classifier on 92.2% of the 1,155 definitively coded papers (binary incorrect-versus-not agreement 95.8%, Gwet's AC1 0.95). Run as a standalone classifier it independently reproduced the near-universal error: it coded 92.0% of all definitively coded papers as incorrect, or 95.2% if the papers it judged indeterminate are excluded, a rate between the raw rule-based figure (96.8%) and the misclassification-corrected estimate (89.0%).

Most usefully, on the 100-paper validation sample the LLM acted as an independent second coder. Weighted to corpus proportions on the same basis as the rule-based validation, its agreement with the human gold standard was 96.0% (Gwet's AC1 0.96), somewhat *higher* than the rule-based classifier's own 90.7% (AC1 0.89; Table 3). Where the LLM and the rule-based classifier disagreed on the binary distinction (17 of the 100 papers), the LLM matched the human reading in 14. It also reproduced, unprompted, the rule-based classifier's chief weakness: several papers the proximity rule had labelled *acceptable* on a coincidental per-group phrase were read by both the human and the LLM as *incorrect*. That a third method, sharing no coding logic and with no sight of the human codes, both corroborates the manual coding and recovers the corpus-wide prevalence is evidence that the result is not an artefact of one particular coding procedure; it does not by itself establish the codes' correctness, since the two automated methods share the

upstream passage extraction, but it does supply a measure of the second reading that a sole-author audit would otherwise lack.

Table 3: The two automated classifiers against the human gold standard (100-paper validation sample, corpus-weighted), and each classifier's standalone incorrect rate over the full corpus. The LLM (Qwen2.5-32B-Instruct, run locally under greedy decoding) saw the same extracted passages but neither the rule-based nor the human codes.

Classifier	Agreement with human (weighted)	Gwet's AC1	Standalone incorrect rate (definitive)
Rule-based	90.7%	0.89	96.8%
LLM (Qwen2.5-32B)	96.0%	0.96	92.0%

The pattern was uniform: across journals with at least 10 coded papers the incorrect rate ranged from 87.5% to 100% (Table 4). The single largest contributor, the *Journal of Physical Therapy Science*, had an incorrect rate of 96.6% (544/563); several journals (including *Physiotherapy Canada* and the *Hong Kong Physiotherapy Journal*) reached 100%. Among the journals with at least ten coded papers, the lowest rate was in the field's flagship journal, the *Journal of Physiotherapy* (14/16; 87.5%), the only journal in the sample that formally advocates estimation over significance testing. The three journals with nominally lower rates (the *Journal of Hand Therapy*, *Physiotherapy Theory and Practice* and the *Journal of Orthopaedic & Sports Physical Therapy*) each contributed fewer than ten papers, so their percentages rest on very small denominators.

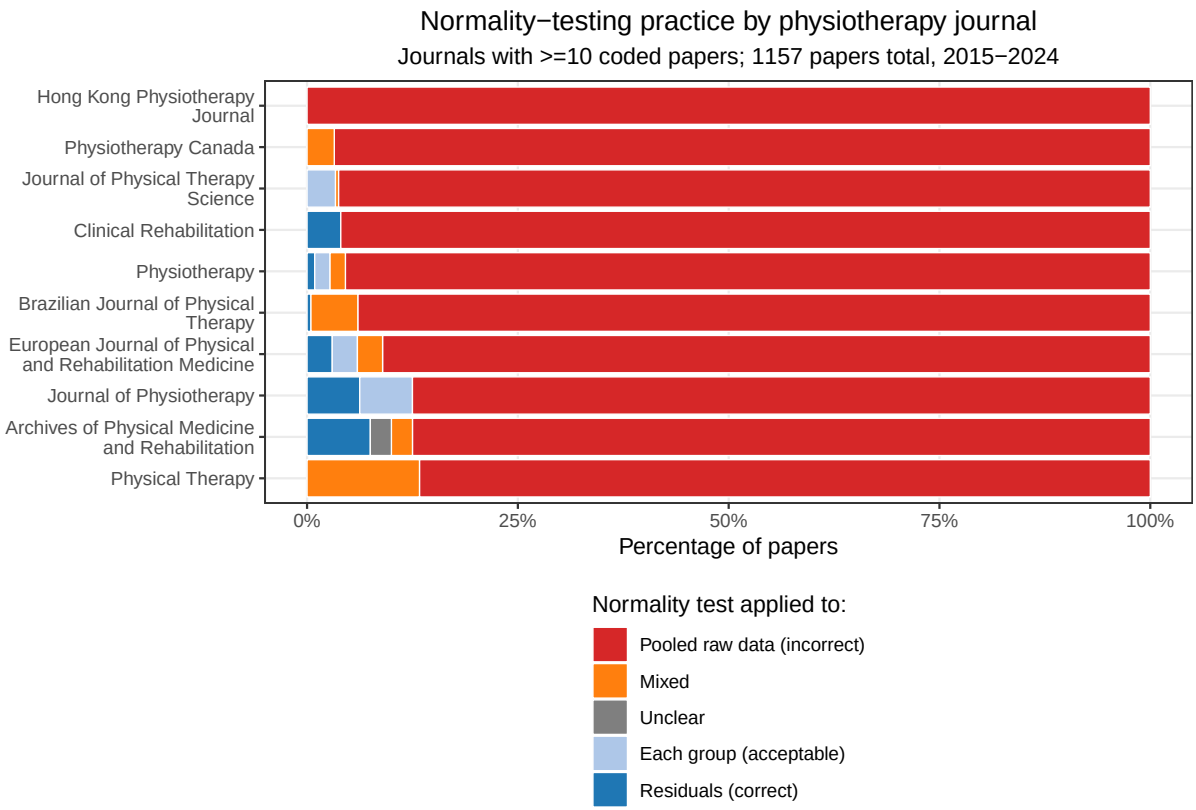

Figure 3: Distribution of normality-testing codes across physiotherapy journals with ≥ 10 coded papers, 2015–2024. Journals are ordered by decreasing proportion of incorrect applications. *Correct* = residuals; *Acceptable* = each group; *Incorrect* = pooled raw data; *Unclear* = no codeable normality-test sentence located.

Table 4: 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 Table 1 (*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.

Temporal trend

Figure 4 shows the proportion of incorrect normality tests by publication year. The rate was uniformly high, ranging from 95% to 99% across years, with no visible upward or downward trend over the decade: every year sits near the ceiling, and the largest annual proportions occur as readily early as late in the period. The yearly proportions, each weighted by the number of coded papers behind it (shown by point size), make this plain without any fitted model.

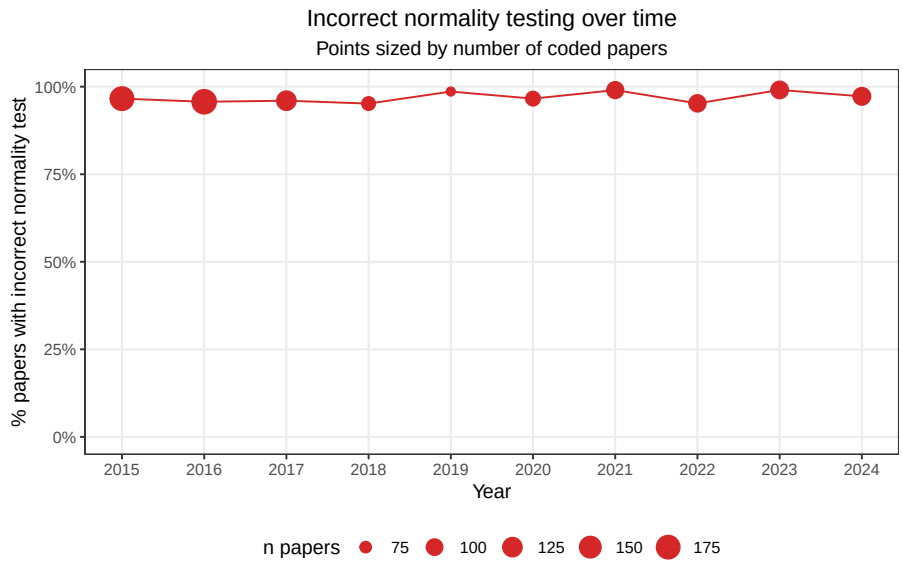

Figure 4: Proportion of papers with incorrect normality testing per year, 2015–2024. Point size is proportional to the number of coded papers.

Comparison across disciplines

The most direct comparators are the two fields audited by Midway and White [8] with the same residuals-versus-raw-data criterion. Figure 5 places the estimate alongside theirs: physiotherapy's conservative misclassification-corrected 89.0% (95% CI 81.4–95.2%), biology's 90%, and ecology's 70%. The two comparators each rest on only 50 hand-coded papers, so their values carry wide binomial intervals (ecology 95% Wilson CI 56–81%, biology 79–96%; Figure 5). Because whether two fields genuinely differ is a question about the distribution of their difference rather than about whether these separate intervals overlap, and the comparators are small, no between-field difference is claimed; restricting to papers whose target could be determined (90.4%) leaves the picture unchanged. Read descriptively, the error is common in all three fields, and physiotherapy's rate is close to the higher of the two published figures.

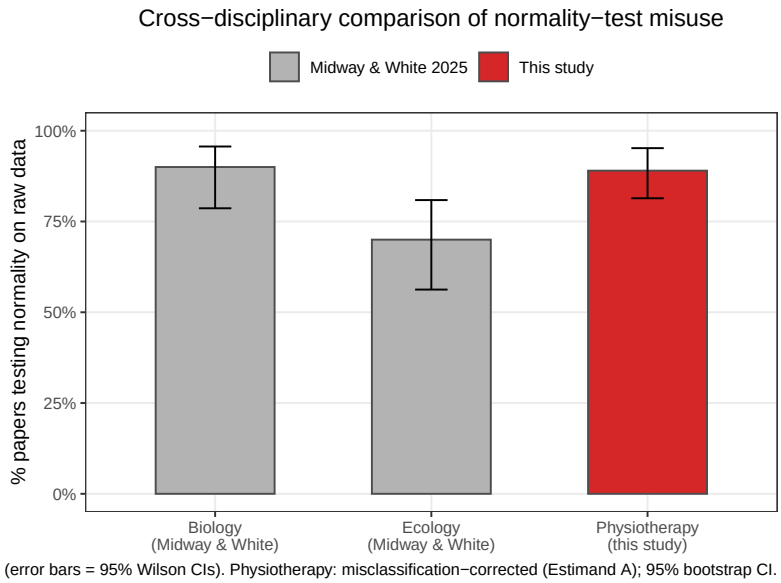

Figure 5: Comparison of the three fields audited with the identical "normality tested on raw data rather than residuals" criterion, among papers reporting a normality test. Ecology and biology values from Midway and White [8] (50 papers per field; error bars are 95% Wilson binomial intervals). The physiotherapy value is the misclassification-corrected prevalence (Estimand A, the primary, conservative estimand); its error bar shows the 95% bootstrap confidence interval.

Restricting the comparison to a single, perfectly matched metric would, however, understate how broad the phenomenon is, because few other disciplines have been audited with exactly this criterion. A wider (if necessarily more heterogeneous) body of evidence points the same way (Table 5). In clinical psychology, Ernst and Albers [14] found that, among the 172 of 893 reviewed articles that actually used linear regression, 92% did not even make clear which assumptions they had checked, and that where the practice could be identified the normality assumption was typically attributed to the variables rather than to the model errors. In an experimental study, Hoekstra et al. [15] observed that when researchers re-analysing data checked normality at all, they examined the *scores* rather than the residuals, and for regression did so "never correctly"; overall only 12% (95% CI 8–18%) of analyses had their assumptions checked correctly. Across every empirical field examined to date (ecology, biology, clinical psychology, experimental psychology, and now physiotherapy), the same conceptual error recurs.

Table 5: 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 [8]	Normality tested on raw data, not residuals (papers reporting a normality test; 50 papers)	70%
Biology	Midway and White [8]	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 [14]	Regression papers unclear about which assumptions were checked (172 regression-using papers, of 893 reviewed)	92%
Clinical psychology	Ernst and Albers [14]	Explicitly tested normality of variables rather than errors (172 regression-using papers, of 893 reviewed)	4%
Experimental psychology	Hoekstra et al. [15]	Analyses with assumptions checked correctly (any technique; 30 researchers)	12%

Simulation study

When no true difference existed ($d = 0$), the rejection rate of all four workflows was very close to their nominal $\alpha = .05$ (A: .049, B: .050, C: .049, D: .049), confirming that the simulated workflows are well calibrated and that the comparisons below reflect genuine differences in power rather than differences in false-positive behaviour.

Figure 6 shows statistical power as a function of group size for a medium effect ($d = 0.5$) across distributions. Under normal data, all four workflows performed virtually identically. The incorrect workflow A and the per-group workflow B were nearly indistinguishable under *every* distribution: at a medium effect ($d = 0.5$, Table 6) the power difference between them stayed within about one and a half percentage points, and across *all* simulated conditions it never exceeded about 2.2 percentage points, the largest difference occurring at a large effect ($d = 0.8$) on skewed data at the smallest sample size. For the realistic pain-score distribution the two workflows differed by at most 0.3 percentage points. The largest gap between the incorrect workflow and the *best* available strategy was likewise only about 2.2 percentage points, again for a large effect on skewed data at $n = 10$ per group. What mattered more than the residual-versus-raw distinction was the eventual choice of test: for skewed and pain-score data, always using the Welch t-test (workflow C) lost up to about 20 percentage points of power relative to the rank-based workflows, whereas the

Mann–Whitney test (workflow D) was at least as powerful as the t-test in every non-normal scenario. This dwarfs the $\leq$2.2-point cost of the raw-versus-residuals error.

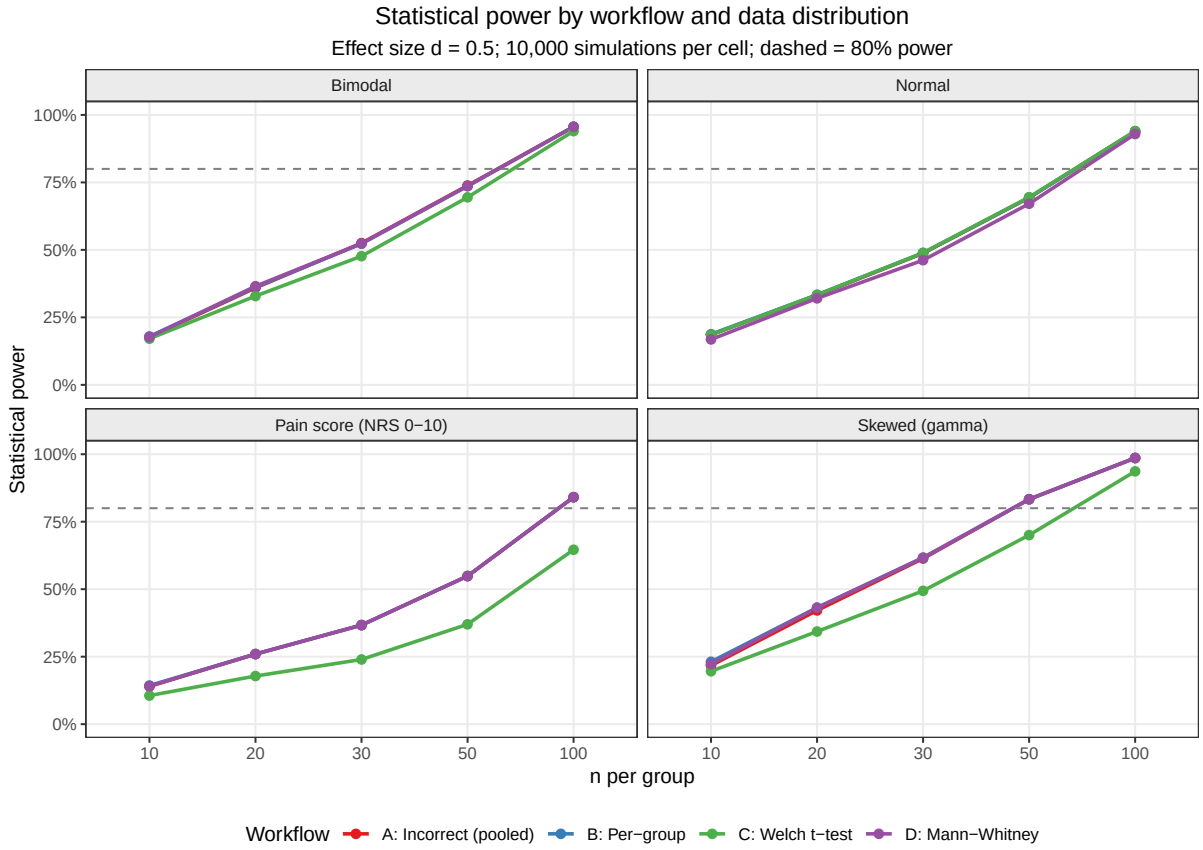

Figure 6: Statistical power by workflow, sample size, and data distribution ($d = 0.5$, 10,000 simulations per cell). The dashed horizontal line marks 80% power, the conventional target corresponding to a 20% type-II error rate; this $\beta = 0.20$ convention is widely used but arbitrary, and is shown here only as a familiar reference, not as an endorsed threshold. Workflows A (incorrect, pooled) and B (per-group) overlap almost exactly.

Table 6: 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

The full set of power differences between the incorrect (pooled) and per-group workflows is shown in Figure 7; the differences cluster tightly around zero (within about 2.2 percentage points throughout, and below one and a half points except at the largest effect ($d = 0.8$) on skewed and bimodal data, and at most 0.3 points for the realistic pain-score distribution), confirming that, for the purpose of choosing between a t-test and a Mann–Whitney test, it makes little practical difference whether the normality test was applied to raw data or to residuals.

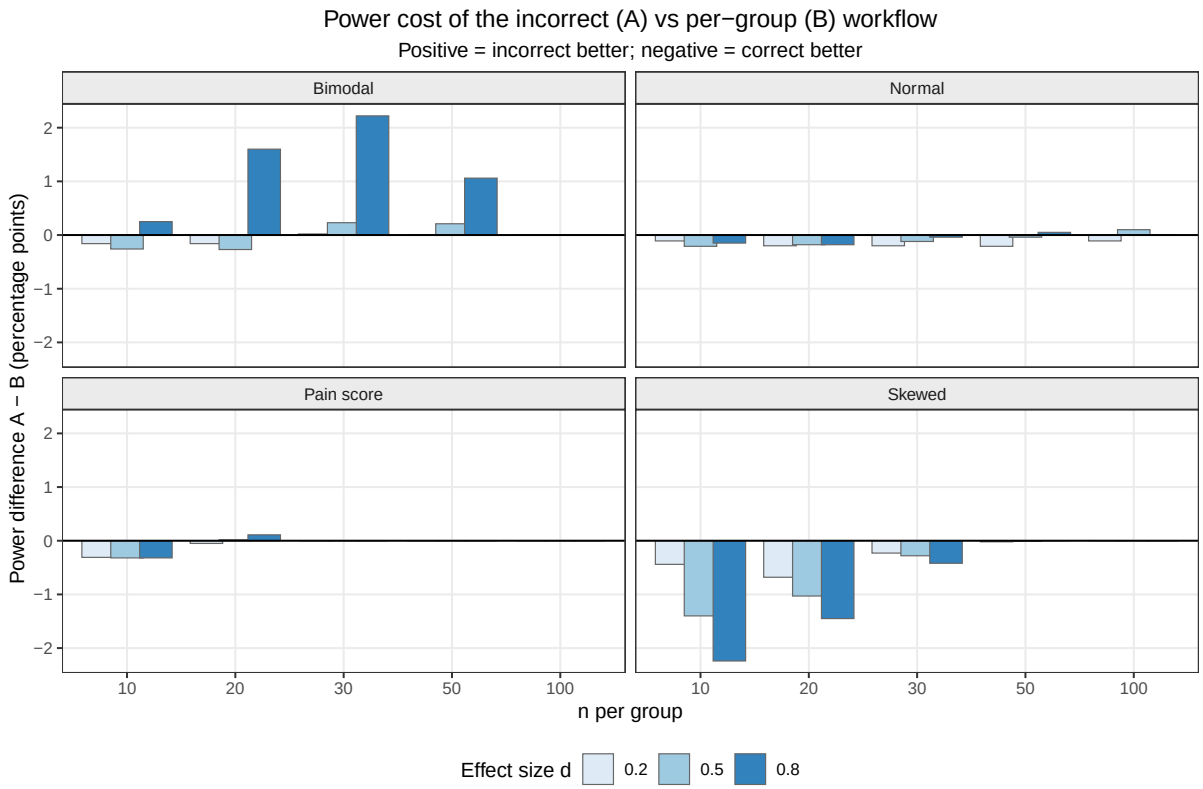

Figure 7: Power difference between the incorrect (A) and per-group (B) workflows across distributions, sample sizes, and effect sizes. Positive values mean the incorrect workflow had higher power. Most differences lie within ±1.5 percentage points; the few exceptions, all at the largest effect ($d = 0.8$) on skewed and bimodal data, reach at most about 2.2 points.

Unequal variances. The scenarios above hold the two groups' variances equal. Because that is exactly the condition under which the pooled and the per-group normality tests tend to agree, the comparison was repeated with heteroscedastic groups, the case most adverse to the pooled test (Table 7). Even when the second group's standard deviation was two or three times the first's, the incorrect (pooled) and per-group workflows differed in power by at most 3.1 percentage points (mean 1.1), the same order of magnitude as under equal variances. The per-group (B) and always-Welch (C) workflows, which keep the Welch t-test, held closest to the nominal 5% false-positive rate under heteroscedasticity (mean Type I $\leq 5.6\%$). The two workflows that drifted upward both fall back on the rank-based test under unequal variances: always-Mann–Whitney (workflow D), whose mean Type I reached 5.7% and 6.6% at SD ratios of 2 and 3, and, more mildly, the incorrect pooled workflow (A), which switches to the rank test when the pooled data look non-normal (mean 5.8% at the 3× ratio). Even so, the audited contrast in *where* normality is tested (A versus B) stayed small in this adversarial setting, in both power (at most 3.1 points) and Type I error (about 0.3 points).

Table 7: 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.

Discussion

Main finding

Almost every physiotherapy paper that reports a normality test applies it incorrectly to the raw or pooled outcome data rather than to model residuals or within-group distributions. The classifier coded 96.8% of papers this way; after correcting for the classifier's labelling error against a hand-coded validation sample, the primary, conservative estimate, which treats every indeterminate paper as not-incorrect, is 89.0% (95% CI 81.4–95.2%), rising to 90.4% (95% CI 82.9–96.7%) among the papers whose target could be determined. All three figures fall in the same broad range as the rates reported by Midway and White [8] for ecology (70%) and biology (90%). The results are consistent with the broader cross-disciplinary evidence summarised in Table 5, including independent findings from clinical [14] and experimental [15] psychology that the normality assumption is routinely attributed to the raw data rather than the residuals. On this evidence, normality-test misapplication is not a quirk of one discipline; it is essentially the default practice in the physiotherapy literature.

Several factors plausibly explain why the mistake is so common. First, popular statistical textbooks and software defaults present descriptive statistics and histograms for the entire dataset, making it natural to test the full distribution. Second, statistical teaching in clinical disciplines tends to emphasise rules of thumb ("test whether your data are normal") rather than the formal definition involving residuals [36]. Third, statistical reviewers may not consistently scrutinise the *target* of normality tests during peer review. The near-uniformity of the error across journals, countries, and a full decade suggests a deeply embedded convention rather than isolated lapses.

The simulation, like that of Midway and White [8], shows that the power consequences

of the mistake are small. The incorrect workflow (testing pooled raw data) and the per-group workflow led to power differences of at most about 2.2 percentage points under every distribution examined (and below one and a half points except for a large effect on skewed data at the smallest sample size), including the skewed, bounded and zero-inflated pain-score distribution typical of physiotherapy (where the two differed by at most 0.3 points), and including the small samples ($n \leq 20$ per group) common in the field. For a two-group comparison with a modest effect, the pooled distribution and the within-group distributions usually lead to the same parametric-versus-non-parametric decision.

This finding sits squarely within the broader literature on the robustness of linear models to assumption violations. Because the t-test and linear regression are, by the central limit theorem, asymptotically valid regardless of the error distribution [9], and because parametric procedures retain near-nominal error rates and power even under marked non-normality [11], the question of whether normality was assessed on the raw data or on the residuals is, for the purpose of inference, largely moot. The point is especially relevant in physiotherapy, where outcomes are frequently bounded, discrete, or Likert-type: parametric analyses of exactly such data have repeatedly been shown to perform well despite the familiar ordinal-measurement objection [10]. In other words, the near-universal error documented here is real, but it is an error about a quantity (the normality of the errors) that the analysis is largely robust to in the first place.

One scope condition deserves emphasis, because the "low-stakes" half of the message is prominent enough to appear in the title. The simulation models a two-group, location-shift comparison, but the t-test is only one of the contexts in which the audited papers sit: the corpus also contains ANOVA ($n = 215$), correlation ($n = 181$), regression ($n = 73$) and ANCOVA or mixed models ($n = 33$), none of which was simulated. The power conclusion is therefore demonstrated *directly* only for the two-group case and *extended* to these other contexts by the robustness argument above rather than by direct simulation. As all of them are linear models (or, for correlation, rest on the same large-sample logic), the residuals-versus-raw distinction is expected to be just as inconsequential there; but readers should treat that step as a reasoned extrapolation, not a simulated result.

The more consequential choice is which test is ultimately used. For skewed data, defaulting unconditionally to the t-test (workflow C) sacrificed up to several percentage points of power relative to the rank-based approaches, whereas the Mann–Whitney test never lost meaningful power even under normality. This suggests that the energy currently spent debating where to apply a normality test might be better directed at the prior question of whether a preliminary normality test is needed at all [2].

Recommendations

1. **For researchers:** If a normality check is performed, apply it to model residuals or within each group, not to the pooled outcome variable. For small samples ($n < 30$ per group), normality tests themselves have low power [1]; for very small samples ($n < 10$) formal testing is uninformative and visual inspection of residual plots is preferable [36]. Defaulting to a robust or rank-based method is a defensible alternative that avoids the issue entirely.

2. **For reviewers and editors:** When a paper reports a normality test, check explicitly whether the authors specify the target (raw data vs. residuals/per group). Requests for clarification should be routine.

3. **For educators:** Courses for health and rehabilitation professionals should emphasise the distinction between the distribution of the outcome variable and the distribution of residuals, ideally with interactive demonstrations.

Classifier performance and sources of disagreement

The human validation (Table 2) clarifies both why the corrected prevalence is a few points below the raw classifier figure and where an automated target-identification rule reaches its limits. Three patterns account for almost all of the disagreement.

First, the classifier's *acceptable* code was the least reliable: of 25 papers it labelled acceptable, the author judged 16 to be incorrect on reading the full passage. The classifier assigns *acceptable* when a per-group phrase ("each group", "both groups", "separately") sits close to a normality cue, but such phrases are often coincidental, describing, for example, how the sample was *described* per group while normality was in fact tested on the pooled data. A human reading the surrounding sentences recovers the actual target; the proximity rule cannot. This is the dominant reason the small not-incorrect stratum shrank under manual review, and it implies that the raw classifier figure, if anything, *understates* the incorrect rate within that stratum.

Second, and pulling in the opposite (and numerically larger) direction, a small fraction of the papers the classifier coded *incorrect* were, on close reading, genuinely indeterminate or not an instance of the error: of 63 sampled, the author coded 5 as acceptable or correct and 1 as unclear. Because this *incorrect* stratum contains the overwhelming majority of the corpus, even this small fraction, reweighted to corpus proportions, accounts for most of the gap between the raw 96.8% and the corrected 89.0%. These were typically papers whose methods text named the normality test without an unambiguous object, so that a careful reader would not commit to "raw data" with confidence.

Third, several disagreements arose in analytic contexts where the residuals-versus-raw-data dichotomy does not map cleanly onto a single "error". For a paired or pre–post design analysed with a paired t-test, the assumption concerns the *differences* [5], so testing the two time points separately is not the acceptable per-group practice it would be in a between-groups design but is itself incorrect; conversely, the classifier's uniform *acceptable* coding of per-group testing over-credits such designs, meaning the corrected estimate is conservative with respect to them. Designs identifiable as paired or repeated-measures from the reported analyses were, however, a small minority of the corpus (about 6%), which bounds the size of this effect. A few analysis types fit the raw-versus-residuals criterion less cleanly. Correlation has no designated outcome and therefore no residual to test: the relevant assumption is bivariate normality of the pair, which the normality of each variable's marginal distribution does not by itself establish [3], so neither side of the raw/residual distinction applies. In factor-analytic and structural-equation models, maximum-likelihood estimation assumes multivariate normality of the observed variables themselves [21, 22], so testing those variables is appropriate rather than the error audited here. The author coded the few such papers as unclear rather than incorrect. Reliability and intraclass-correlation analyses, by contrast, do rest on a model (a random-effects structure with normally distributed errors), so the raw-versus-residuals criterion applies to them as it does to the linear model, and they were treated under the same rule rather than as exceptions. And in a handful of settings testing the raw distribution is positively *correct* (normative reference intervals, Bland–Altman limits of agreement, or process control), so a paper doing so for those purposes is over-counted as incorrect by a target-based rule.

These boundary cases pull in both directions and are a minority; the sensitivity of the main conclusion to them is captured by the gap between the two estimands; but they mark the edge of what a single raw-versus-residuals criterion can adjudicate, and they are flagged explicitly rather than forced into a code.

Limitations

First, full-text search favours journals whose full text is indexed, and the corpus is heavily dominated by a few high-volume open-access journals: almost half of all articles come from a single open-access title (the *Journal of Physical Therapy Science*) and roughly two-thirds from just two (with the *Brazilian Journal of Physical Therapy*). That full text could be retrieved for almost the entire corpus is largely a consequence of this open-access dominance, not of comprehensive coverage of the field: open-access articles are freely available to anyone, the two subscription journals not full-text indexed by Europe PMC (the *Journal of Physiotherapy* and *Physiotherapy*) were obtained through institutional access to the Elsevier text-and-data-mining API, and the remaining non–open-access articles through the rendered PDFs served by Europe PMC. The 16 journals were moreover a purposive rather than a formally sampled selection, so they need not be representative of the field as a whole; smaller or non-indexed journals may be under-represented, and the estimate should be read as characterising the indexed physiotherapy literature in widely-read journals. Second, classification was automated, and the classifier is imperfect: corpus-weighted agreement with manual coding was 90.7% on the binary distinction, and the *acceptable* code in particular proved unreliable (as detailed above). This was addressed directly by correcting the prevalence for the measured classification error rather than reporting only the raw classifier count, and all passages and codes are released (where licensing permits) for independent verification. Third, the human gold standard against which the classifier was corrected was produced by the sole author, who was not blinded to the study's expected result. Several factors limit the scope for bias: the coder *was* blinded to the classifier's output (the coding sheet omitted it); the coding task is mechanical, requiring identification of an explicitly named target in a methods sentence rather than a judgement of quality; and the full validation sample, passages, and codes are released so that any reader can re-code them. The single-coder concern was further mitigated with the independent LLM cross-classification reported above: a third method that saw neither the rule-based nor the human codes agreed with the human gold standard at least as well as the rule-based classifier did (corpus-weighted 96.0% vs 90.7%; AC1 0.96 vs 0.89) and reproduced the corpus-wide prevalence, providing a partial second reading a sole-author audit would otherwise lack (partial because both automated methods share the upstream passage-extraction step). A blinded human second coder would strengthen this further still. Fourth, coding reflects what authors *reported*, not necessarily what they did, and, as discussed above, for paired designs and for correlation and factor-analytic contexts the raw-versus-residuals criterion is an imperfect fit, so individual codes at those boundaries carry more uncertainty than the aggregate figure. Fifth, the simulation assumed a simple two-group design; more complex physiotherapy designs (repeated measures, hierarchical data) may behave differently, although the broad insensitivity of power to the residual-versus-raw distinction is likely to persist. Sixth, the corpus is by construction restricted to papers that report a *named* normality test (Shapiro–Wilk, Kolmogorov–Smirnov and the related terms in the query): papers that assess normality only graphically (for example with a normal quantile–quantile plot of the residuals, which is correct practice) carry no

matching search term and are not captured, as are any tests named with terms outside the list (e.g. Shapiro–Francia or Jarque–Bera). The prevalence should therefore be read as conditional on reporting a formal normality test, the same population of papers the comparator audits of normality testing address; it does not speak to the (unknown) share of authors who check residuals graphically and correctly, and it cannot, on its own, distinguish a field that tests normality badly from one that mostly avoids formal testing altogether.

Comparison with related work

The conceptual error audited here, conflating the distribution of the outcome with the distribution of the residuals, has long been flagged in the methodological literature [14, 37], yet the results show that this guidance has had little effect on practice. Statistical practice in physiotherapy has itself been examined before, including the methods used across the field's journals [38] and the reporting of statistical significance and clinical relevance in its trials [39]; to the author's knowledge, however, the specific question of *where* the normality assumption is checked has not previously been audited in this field. The contribution of this study is to isolate a single, precisely defined, and easily checkable error, to quantify it at scale in physiotherapy for the first time, and to show through simulation that its practical cost is small.

It is worth being explicit about why a near-universal but low-stakes error is worth measuring, since the two findings might seem to cancel out. They do not. That the error is near-universal makes it a reliable marker of a shared conceptual gap (many authors evidently treat "test the data for normality" as the assumption itself, rather than as a proxy for a condition on the residuals), and a gap this widespread is an educational target, not a series of isolated slips. That its inferential cost is nonetheless small carries a reassuring message that is itself useful: the existing physiotherapy literature is not invalidated by it, and authors need not anxiously re-run published analyses. Quantifying the smallness is what licenses that reassurance instead of mere hope. Finally, the same simulations redirect attention to the choice that does matter (which test is ultimately used, and whether a preliminary normality test is needed at all [2]), where the stakes (here up to about 20 percentage points of power) dwarf the residuals-versus-raw distinction. An error that is free to fix, reflects a common misunderstanding, and points to a more consequential question is an unusually tractable target for correction.

Conclusion

Normality tests are misapplied in the large majority of physiotherapy papers that report them: after correcting for classifier error, a conservatively estimated 89.0% (95% CI 81.4–95.2%) test the raw outcome data rather than model residuals or within-group distributions, a rate in line with those previously reported for biology and ecology. The consequences for statistical power are nonetheless small, echoing Midway and White [8]. Because correct practice costs nothing and occasionally helps, and because the error is so easily spotted, there is an opportunity for a simple, high-yield improvement in statistical reporting: targeted education for physiotherapy researchers, clearer reporting standards, and a one-line check during peer review.

Declarations

Ethics approval and consent to participate

Not applicable. This study analysed only previously published, aggregate bibliographic data obtained from public databases; it did not involve human participants, their identifiable personal data, or animals.

Consent for publication

Not applicable.

Availability of data and materials

The data and code supporting the conclusions of this article (the corpus metadata, the extracted normality passages, the per-paper classifier codes, the 100-paper validation sample, and all retrieval, classification, validation, simulation and analysis scripts) are available in a public repository at https://github.com/jdegenfellner/normality-testing-physiotherapy-audit_public; a permanent, versioned archive of this repository, with a citable DOI, will be deposited on Zenodo upon acceptance. The underlying literature records were obtained from the public Europe PMC (https://europepmc.org) and PubMed Central (https://www.ncbi.nlm.nih.gov/pmc) application programming interfaces. For content that is not openly licensed, namely the two journals retrieved through the Elsevier full-text API (*Journal of Physiotherapy* and *Physiotherapy*) and articles recovered as rendered PDFs from non–open-access sources, the licence terms preclude redistribution of the full text; for these the repository therefore contains only the article identifiers (DOIs) and the derived per-paper codes, from which the analysis can be reproduced using the provided scripts.

Competing interests

The author declares that he has no competing interests.

Funding

No funding was received for this study.

Authors' contributions

J.D. is the sole author and conducted all aspects of the study. Use of a large language model is documented in the Methods.

Acknowledgements

Not applicable.

References

[1] Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[2] Justine Rochon, Matthias Gondan, and Meinhard Kieser. To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12:81, 2012. doi: 10.1186/1471-2288-12-81.

[3] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition, 2002.

[4] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and Other Stories*. Cambridge University Press, Cambridge, 2021.

[5] Julian J. Faraway. *Linear Models with R*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition, 2014.

[6] John H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, MD, 3rd edition, 2014. URL http://www.biostathandbook.com.

[7] Peter H. Westfall and Andrea L. Arias. *Understanding Regression Analysis: A Conditional Distribution Approach*. Chapman and Hall/CRC, Boca Raton, FL, 2020.

[8] 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.

[9] Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23:151–169, 2002. doi: 10.1146/annurev.publhealth.23.100901.140546.

[10] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5):625–632, 2010. doi: 10.1007/s10459-010-9222-y.

[11] Ulrich Knief and Wolfgang Forstmeier. Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53(6):2576–2590, 2021. doi: 10.3758/s13428-021-01587-5.

[12] Amand F. Schmidt and Chris Finan. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98:146–151, 2018. doi: 10.1016/j.jclinepi.2017.12.006.

[13] Joseph L. Goulet, Eugenia Buta, Harini Bathulapalli, Ralitza Gueorguieva, and Cynthia A. Brandt. Statistical models for the analysis of zero-inflated pain intensity numeric rating scale data. *The Journal of Pain*, 18(3):340–348, 2017. doi: 10.1016/j.jpain.2016.11.008.

[14] 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.

[15] 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.

[16] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19, 2019. doi: 10.1080/00031305.2019.1583913.

[17] Dieter Rasch, Klaus D. Kubinger, and Karl Moder. The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers*, 52(1):219–231, 2011. doi: 10.1007/s00362-009-0224-x.

[18] William R. Schucany and H. K. T. Ng. Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. *Communications in Statistics — Theory and Methods*, 35(12):2275–2286, 2006. doi: 10.1080/03610920600853308.

[19] Donald W. Zimmerman. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181, 2004. doi: 10.1348/000711004849222.

[20] Craig S. Wells and John M. Hintze. Dealing with assumptions underlying statistical tests. *Psychology in the Schools*, 44(5):495–502, 2007. doi: 10.1002/pits.20241.

[21] Rex B. Kline. *Principles and Practice of Structural Equation Modeling*. Guilford Press, New York, NY, 4th edition, 2016.

[22] Barbara G. Tabachnick and Linda S. Fidell. *Using Multivariate Statistics*. Pearson, Boston, MA, 6th edition, 2013.

[23] Graham R. D. Jones and Andrew Barker. Reference intervals. *The Clinical Biochemist Reviews*, 29(Suppl 1):S93–S97, 2008.

[24] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.

[25] Douglas C. Montgomery. *Introduction to Statistical Quality Control*. Wiley, Hoboken, NJ, 8th edition, 2020.

[26] Kilem L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600.

[27] Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990. doi: 10.1016/0895-4356(90)90158-L.

[28] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

[29] Qwen Team. Qwen2.5 technical report, 2024.

[30] Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on Apple silicon. https://github.com/ml-explore/mlx, 2023.

[31] Python Software Foundation. *Python Language Reference, version 3*. Python Software Foundation, 2024. URL https://www.python.org.

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org.

[33] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milborrow Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the `tidyverse`. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

[34] Michael P. Fay and Michael A. Proschan. Wilcoxon–Mann–Whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010. doi: 10.1214/09-SS051.

[35] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

[36] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2): 486–489, 2012. doi: 10.5812/ijem.3505.

[37] Douglas Curran-Everett. Explorations in statistics: the assumption of normality. *Advances in Physiology Education*, 41(3):449–453, 2017. doi: 10.1152/advan.00064.2017.

[38] James Roush, James W. Farris, Lori M. Bordenave, Samantha Sesso, Ali M. Benson, and Carissa Millikan. Commonly used statistical methods in the journals associated with physical therapy and physiotherapy. *Journal of Physical Therapy Education*, 29 (1):5–9, 2015. doi: 10.1097/00001416-201529010-00003.

[39] Arianne Verhagen, Peter William Stubbs, Poonam Mehta, David Kennedy, Anthony M. Nasser, Camila Quel de Oliveira, Joshua W. Pate, Ian W. Skinner, and Alana B. McCambridge. Comparison between 2000 and 2018 on the reporting of statistical significance and clinical relevance in physiotherapy clinical trials in six major physiotherapy journals: a meta-research design. *BMJ Open*, 12(1):e054875, 2022. doi: 10.1136/bmjopen-2021-054875.

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 1):

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 2. 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.

LLM classifier prompt

The independent language-model cross-classification (Independent LLM 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.

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.