medical sciences

MDPI

Article

Changes in Statistical Reporting and Interpretation in Clinical Neurology Journals, 2015–2025: A Meta-Research Study

Rahim R. Aliyev [1,*], Shahla N. Mehtiyeva [1], Sona E. Aliyeva [1], Ilhama M. Karimova [2], Vafa G. Alili [3], Nijat V. Gasim [4], Aida Kondybayeva [5] and Ayten K. Mammadbayli [1]

[1] Department of Neurology and Medical Genetics, Azerbaijan Medical University, Baku AZ1078, Azerbaijan; smehtiyeva@amu.edu.az (S.N.M.); saliyeva@amu.edu.az (S.E.A.); amammadbayli@amu.edu.az (A.K.M.)
[2] Department of Obstetrics and Gynecology, Azerbaijan Medical University, Baku AZ1022, Azerbaijan; ikerimova1@amu.edu.az
[3] Department of Information and Communication Technologies, Baku Slavic University, Baku AZ1014, Azerbaijan; vafa.alili.q@bsu-uni.edu.az
[4] Department of Biostatistics, Faculty of Medicine and Health Sciences, Qarabağ University, Khankendi AZ2600, Azerbaijan; nijat.gasim@qu.edu.az
[5] Scientific and Educational Center for Neurology and Applied Neuroscience, Asfendiyarov Kazakh National Medical University, Almaty 050012, Kazakhstan; kondybayeva.a@kaznmu.kz
* Correspondence: drrahimaliyev@gmail.com

Abstract

Background/Objectives: We described statistical reporting and interpretation and explored differences between two selected samples of original quantitative articles assigned to 2015 and 2025. **Methods:** We audited 200 articles from 10 clinical neurology journals (10 per journal-year) drawn from nonexhaustive pools. For 18 outcomes, we calculated uncorrected journal-stratified Mantel–Haenszel odds ratios (ORs), 95% confidence intervals and Cochran–Mantel–Haenszel p values with Benjamini–Hochberg false discovery rate (BH-FDR) adjustment. Two external assessors independently rated 20 articles. Complete full texts could not be verified for 74 articles. **Results:** No primary comparison met the BH-FDR threshold (minimum q = 0.0532). Among the evaluable articles, fully adequate reporting was uncommon in both samples (2015 vs. 2025): sample-size justification, 9.1% vs. 13.2%; assumptions, 10.8% vs. 15.4%; missing-data handling, 39.1% vs. 26.1%; and multiplicity, 15.5% vs. 23.9%. The effect-estimate reporting OR was 1.95 in the primary analysis and 1.28 in the complete-full-text subset. **Conclusions:** The audit documents specific reporting gaps in the selected articles. The comparability of the two samples remains uncertain given incomplete sampling records, source access, differences in article composition, and limited coding reproducibility.

Keywords: clinical neurology; statistical reporting; meta-research; research methodology; p values; confidence intervals; effect estimates; causal inference; reproducibility

Check for updates

Academic Editor: Chi Kwan Tsang

Received: 26 October 2026
Revised: 12 November 2026
Accepted: 18 November 2026
Published: 21 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Transparent statistical reporting helps readers evaluate analytical choices, interpret effect magnitude and precision, and reproduce or synthesize findings. SAMPL and ICMJE recommendations emphasize effect estimates, uncertainty and methodological detail [1,2]. The two selected years provide snapshots a decade apart. As historical context, 2015 precedes the 2016 ASA statement on p values, and 2025 follows subsequent statistical-reporting guidance, including the 2019 *New England Journal of Medicine* editorial [3–5]. This comparison does not evaluate the impact of those publications or of any journal policy.

Article-level audits have documented misuse of statistical methods, limited analytical transparency and mismatches between observational designs and causal claims [6–10]. An evaluation of 85 neurology and neuroscience journals found that approximately 71% gave no specific guidance on *p*-value reporting or interpretation and 46% did not require disclosure of how assumptions were assessed [11]. That evaluation concerned instructions to authors. The present audit examines what published articles reported.

Recent recommendations emphasize estimation, analytical transparency and interpretation appropriate to the study design [12–19]. Existing article-level audits have covered distinct populations, including preclinical motor-behavioral neuroscience, the *Journal of Neurophysiology*, and studies of transcranial magnetic stimulation in neurological conditions [7,9,10].

Our objective was to describe 13 statistical-reporting features and five interpretation or consistency problems in two selected samples of original quantitative articles assigned to 2015 and 2025 from the same ten clinical neurology journals, and to explore differences between those samples. We also assessed journal-level directions, coding reliability, sensitivity to source availability, and handling of unclear codes and the adequacy threshold [20].

2. Materials and Methods

2.1. Study Design

We conducted an article-level meta-research audit of two selected publication-year samples from ten clinical neurology journals. The unit of analysis was the journal article. The main analysis comprised 200 articles, with ten per journal-year and 100 assigned to each year.

Because the study evaluated published articles and did not involve collection of individual-level participant data, institutional ethics approval was not required.

2.2. Journal Selection

The panel was selected before article-level sampling using pragmatic criteria: broad clinical-neurology scope, continuous publication in both years and sufficiently searchable archives to assemble candidate pools. It remained unchanged throughout the study.

The fixed journal panel included *Neurology*, *JAMA Neurology*, *the Journal of Neurology, Neurosurgery & Psychiatry*, *the European Journal of Neurology*, *the Journal of Neurology*, *the Journal of the Neurological Sciences*, *Clinical Neurology and Neurosurgery*, *Neurological Sciences*, *BMC Neurology*, and *the Canadian Journal of Neurological Sciences*.

Using the same journals in both years permits a journal-stratified contrast between the selected samples. The journal panel was chosen pragmatically.

2.3. Article Eligibility

Articles were eligible if they were original quantitative studies involving human participants or human-derived clinical data and included inferential statistical analysis.

Reviews, systematic reviews, meta-analyses, letters, case reports, study protocols, and articles limited to descriptive statistics without inferential testing or modeling were excluded. Eligibility was assessed at article level from the accessible materials, including full-text assessment where available and needed. Complete full-text availability was not an eligibility requirement for the historical sample.

An article initially classified as eligible could subsequently be excluded if full-text review showed that it did not satisfy the inferential-analysis criterion. Such exclusions did not alter the predefined sampling sequence.

2.4. Candidate Pools, Calibration and Sampling

Separate, nonexhaustive pragmatic pools were assembled by journal and assigned publication year. The documented sampling pool is the pool recorded in the sampling log for each journal-year. The sampling summary and Supplementary Table S1 report those counts, cross-checked against the archived sampling-status summary. The candidate register preserves additional records. The exact search strings, screening dates, input ordering and a complete account of candidate discovery and eligibility decisions could not be reconstructed. The records do not establish that search coverage and screening were comparable across years.

The log records seed 20152025 and a within-stratum sequence for documented selections and replacements. This identifies the initial draws and recorded updates, which support the final ten-per-journal-year allocation in Supplementary Table S1. All ten logged candidates in *the Journal of the Neurological Sciences* 2015 pool entered the sample.

For the exclusions recorded after initial selection, the log documents selection of the next eligible candidate in the retained sequence, with no recorded reseeding. The documented sampling pool sizes by journal and assigned publication year are presented in Table 1, with additional details on exclusions, replacements, and final selected articles provided in Supplementary Table S1. The archived statuses distinguish records outside the logged pool from candidates ruled ineligible at later recorded steps.

Table 1. Sampling summary. Documented sampling pool sizes by journal and assigned year.

Journal	2015 Pool *n*	2025 Pool *n*
Neurology	18	18
JAMA Neurology	18	18
Journal of Neurology, Neurosurgery & Psychiatry	18	18
European Journal of Neurology	18	18
Journal of Neurology	18	18
Journal of the Neurological Sciences	10	18
BMC Neurology	20	18
Neurological Sciences	18	18
Clinical Neurology and Neurosurgery	16	18
Canadian Journal of Neurological Sciences	18	18

Notes: The final sample contains ten articles in every journal-year. The candidate register contains 21 records for *Neurology 2015* (three additional "exclude-main"), 23 for *BMC Neurology 2015* (two additional "exclude" and one "exclude-main"), and 20 for *the Canadian Journal of Neurological Sciences 2025* (two additional "exclude"). In *BMC Neurology 2015*, the logged pool of 20 includes two candidates subsequently recorded as ineligible. The other 17 strata have the same number of archived records and logged pool entries. See Supplementary Table S1 for documented replacements and final article IDs.

The ten-per-journal-year allocation was adopted for comparison feasibility. There was no formal prospective power calculation or preregistered confirmatory endpoint. The analysis pertains to the selected articles and remains vulnerable to selection effects from incomplete pool coverage and documentation.

A separate 12-article internal calibration set was used to refine and standardize application of the operational coding framework before completion of the main audit. These articles and their supporting evidence were retained separately and did not contribute to the 200-article analytical sample or any temporal comparisons.

2.5. Human Coding, Adjudication and Reliability

Two independent external assessors (Assessors 1 and 2) completed a 20-article Round-1 calibration exercise; its ratings were used to identify ambiguities in the application of the operational rules before the revised codebook was locked. These were not included in the final reliability estimates. A separate 12-article internal calibration resource was also archived and excluded from the 200-article analysis. The two calibration resources should not be conflated with the 20-article Round-2 reliability sample.

A distinct 20-article Round-2 sample (one record per journal-year stratum; seed 26092027) was independently rated by the same assessors using the revised rules before adjudication. The publication-year column was removed from the Excel assessment forms provided to the external assessors. Both assessors completed their evaluations independently, without access to each other's ratings. However, publication-year information remained visible in the source article PDFs; therefore, complete blinding to publication year was not ensured. The archived Round-2 templates and completed workbooks contain a publication-year field. The distributed copies have not been independently authenticated. Binary raw agreement, Cohen's kappa (κ), and Gwet's AC1 were calculated using the primary outcome recordings after excluding pairs with codes 8 or 9. Raw-agreement intervals used Wilson's method. Coefficient intervals used 4000 percentile-bootstrap resamples of evaluable pairs (seed 20261004), retaining finite replicates. Gwet's AC1 complements κ when rating marginals are imbalanced [21,22]. Supplementary Table S2 gives the domain-specific denominators and binary and original-code reliability results. The later targeted adjudication is described separately below.

Following historical source review, a separate pair of independent assessors (Assessors A and B), who had not participated in the Round-1 or Round-2 exercises, reassessed 33 targeted article-domain items, comprising 32 historical coding discrepancies and one additional verification involving P004. Their initial independent ratings were concordant for 23 items. Seven disagreements were resolved through documented discussion, while the remaining three items required determination by an independent third adjudicator. Comparison of the final human-adjudicated decisions with the frozen pre-adjudication dataset identified exactly 11 article-domain coding changes, which were incorporated without modifying any other article-level fields. This targeted, disagreement-enriched adjudication exercise was distinct from the Round-2 reliability assessment. It was not a random reliability sample, and no separate inter-rater reliability coefficient was calculated from these 33 items. No additional independent reliability round or comprehensive independent reassessment of the 200-article dataset was conducted. Twenty-three additional historical discrepancies across six articles remained unresolved because sufficient original full-text evidence was unavailable for independent reassessment. Their original codes were preserved. The individual human-adjudication records and unresolved evidence register are retained in the confidential internal audit archive. The public Supplementary Materials provide the operational codebook, inter-rater reliability results, and reproducible article-level analytical data.

2.6. Audit Framework and Outcomes

A structured audit framework was developed to assess statistical reporting, analytical transparency, and interpretation of quantitative findings. Detailed operational definitions and decision rules for all coded domains are provided in Supplementary Table S3.

The primary coding of the 200-article study sample was performed by Rahim R. Aliyev (R.R.A.) using the accessible publication materials and article-specific evidence notes. IBM SPSS Statistics, version 27, was used for selected checks; final inferential calculations were reproduced in Python. The subsequent external assessments covered selected items and the

separate Round-2 subsample, not complete duplicate coding of the 200 articles. Although 33 targeted judgments were adjudicated, this does not establish reliability of all final article-domain classifications.

Coding distinguished assessable absence (0), partial evidence (1), adequate reporting (2), not applicable (8), and not verifiable (9) according to the locked domain-specific rules. Item-specific evidence and context, not lack of a searchable phrase alone, governed the distinction between absent reporting and unavailable evidence.

Complete full text could not be verified for 74/200 articles (37 per year). For these articles, accessible abstracts, publisher or PubMed descriptions, or excerpts supported only fields verifiable from that material. Seventy-one articles had at least one unclear outcome code (9). Twenty-three historical article-domain discrepancies across six source-limited articles remained unresolved, and their original classifications were retained. Source restriction and unclear-code scenarios assess sensitivity to the available evidence and associated coding assumptions.

Statistical complexity was an ancillary structural descriptor: basic unadjusted analysis (low), conventional adjusted modeling (moderate), and advanced or multiple structures including mixed effects, weighting, recurrent events, high-dimensional imaging or machine learning (high). Application of the revised rule yielded 94 changes (92 high to moderate, two moderate to high). For example, P161's standard Cox subgroup/interaction modeling and P004's logistic and conventional comparisons were not by themselves high-complexity structures under the locked rule. This is post-audit harmonization, not separately replicated coding.

Thirteen positive statistical-reporting domains were evaluated:

1. Sample-size justification;
2. Appropriate descriptive statistics;
3. Reporting of an interpretable effect estimate;
4. Reporting of confidence intervals or other appropriate statistical uncertainty;
5. Exact *p*-value reporting;
6. Clear identification of statistical methods;
7. Assessment or reporting of relevant statistical assumptions;
8. Adequate specification of the principal statistical model;
9. Reporting and handling of missing data;
10. Handling of multiplicity;
11. Sensitivity or robustness analysis;
12. Identification of statistical software;
13. Identification of the software version.

Most reporting variables were coded on a three-level scale:

- 0 = not reported or inadequate;
- 1 = partially reported or partially adequate;
- 2 = adequately reported.

Where applicable, additional codes were used:

- 8 = not applicable;
- 9 = unclear or not verifiable from the available material.

Five additional outcomes assessed potentially problematic statistical interpretation or reporting:

- Interpretation of a nonsignificant result as evidence of no effect or no difference;
- Use of "trend" or "tendency toward significance" terminology;
- Unsupported causal or mechanistic interpretation not warranted by the study design or analysis;

- Numerical inconsistency;
- Presence of at least one major reporting, interpretation, or internal-consistency problem.

The first three and the major-problem indicator were coded as absent, present, or unclear. Numerical consistency was classified as no discrepancy, minor discrepancy, potentially major discrepancy, or not verifiable.

2.7. Outcome Coding and Dichotomization

The primary operational target for the 13 reporting domains was full adequacy (code 2), with absent or inadequate reporting (0) and partial reporting (1) grouped together. The post hoc alternative threshold grouped partial and adequate reporting (1/2) and compared them with absent or inadequate reporting (0). Both used the same exclusions for codes 8 and 9. The second definition targets the presence of at least partial reporting. The full ordinal distributions and alternative estimates are available in the supplementary results and Supplementary Table S4.

Codes 8 and 9 were excluded from the relevant outcome-specific evaluable denominator. Software identification used the 0/1/2/9 rule and was evaluable in 134 articles; the locked software version rule instead assigned 0 when no version was found in the reviewed material, including where software identity was itself unverifiable, so all 200 articles had a software version code. These distinct denominators reflect operational conventions, not equally complete verification; especially for source-limited records, version code 0 must not be interpreted as verified absence from the unavailable full text.

For adverse interpretation outcomes, the event was defined as presence of the problematic reporting practice. Articles coded as unclear were excluded from the corresponding evaluable denominator.

For numerical consistency, minor or potentially major discrepancies were combined into an "any discrepancy" category and compared with articles in which no discrepancy was identified.

2.8. Statistical Analysis and Multiplicity

The primary comparison contrasted articles assigned to the 2025 and 2015 study-year strata.

The primary estimator was a standard uncorrected journal-stratified Mantel–Haenszel (MH) common OR comparing 2025 with 2015, with Robins–Breslow–Greenland log-scale 95% confidence intervals and the uncorrected Cochran–Mantel–Haenszel (CMH) score-test p. No 0.5 pseudo-counts were added to zero cells, and no Yates/continuity subtraction was applied to the CMH statistic. Only journal strata informative for the relevant outcome contribute to the estimate; all evaluable records contribute to descriptive counts. Where the pooled OR lies on a boundary, the log-scale CI is marked not estimable (NE), and asymptotic p values warrant particular caution. A historical alternative that adds 0.5 to all four cells of any zero-containing journal stratum was applied to the final 200-article analytical dataset as a post hoc sparse-data sensitivity; this alternative was not selected based on inferential significance.

For positive reporting outcomes, an OR greater than 1 indicated higher odds of adequate reporting in 2025. For adverse interpretation or reporting outcomes, an OR greater than 1 indicated a higher frequency of the problematic practice in 2025. Absolute differences in proportions were also reported in percentage points.

Benjamini–Hochberg false-discovery-rate adjustment [23] was applied only to the defined family of 18 primary uncorrected CMH p values (13 reporting and five adverse outcomes). The statistical-analysis plan was finalized after descriptive inspection and was not prospectively registered. Both nominal p values and FDR-adjusted q values were reported, with effect estimates and 95% confidence intervals emphasized in interpretation.

The FDR significance criterion was $q < 0.05$; comparisons with nominal $p < 0.05$ but $q \geq 0.05$ were interpreted as exploratory rather than statistically confirmed temporal differences.

Journal-stratified conditional logistic regression was used as a post hoc secondary estimator to examine within-journal associations under sparse-data conditions. Conditioning on journal-specific outcome totals removes journal intercepts. The conditional likelihood was evaluated from within-journal year/outcome margins; model-based Wald CIs and p values were reported. These estimates describe within-journal year associations and do not constitute an exact conditional hypothesis test.

For secondary models including categorical statistical complexity and journal indicators, Firth bias-reduced logistic regression used a Jeffreys information penalty [24,25]. Penalized likelihood-ratio p values and profile penalized-likelihood 95% CIs were calculated. These secondary estimators were added after observing sparse-data instability and did not replace the uncorrected-MH primary estimates or their 18-outcome BH family.

Complexity-adjusted models assessed whether descriptive year associations changed under an alternative estimator and adjustment specification. They do not estimate a causal year effect or supersede the primary MH estimates and multiplicity adjustment.

Secondary-model validation included optimizer score checks and profile-interval bracketing. The descriptive-statistics year contrast remained at the boundary for unpenalized conditional estimation; its Firth estimate is secondary and penalty-dependent.

For each outcome, changes in reporting proportions from 2015 to 2025 were examined separately within each journal. Journals were classified as showing improvement, worsening, or no change according to the direction of the within-journal difference.

Homogeneity was explored using the unadjusted Breslow–Day test in informative strata where the pooled OR was finite and positive. Sparse or boundary tables can make this test unstable or non-estimable; no zero-cell correction was added for this diagnostic.

2.9. Sensitivity Analysis Specifications

Two source-accessibility checks were specified in the archived analysis plan. The first retained the primary coding definitions and restricted the population to the 126 articles with verified complete text (63 per year). The second kept the original sample frame and, for the 13 reporting domains, classified unverifiable observations (code 9) as not adequate while continuing to exclude not-applicable observations (code 8). This assumption changes the outcome-specific evaluable denominator and classification of unclear records. It does not verify the missing content.

Post hoc code-9 scenarios assigned unclear records to favorable or adverse outcomes in both years, leaving code 8 excluded. These scenarios illustrate the effect of specific classification assumptions. Other post hoc checks used a 0.5 sparse-cell correction and a lower adequacy threshold for the 13 reporting domains (codes 1/2 versus 0), which changes the reporting estimand while retaining the primary exclusions (Supplementary Tables S4 and S5). The conditional and Firth models and the stratified exact boundary check were also exploratory. The primary BH-FDR family comprised the same 18 uncorrected comparisons throughout.

Statistical complexity was summarized descriptively before and after harmonization (Table 2 and accompanying text). The reproducibility package contains the final harmonized complexity codes used as a categorical covariate only in secondary sensitivity models. It was not one of the 18 primary reporting or interpretation outcomes and therefore was not included in their BH-FDR adjustment family.

Table 2. Characteristics of the audited article sample.

Characteristic	2015 ($n = 100$)	2025 ($n = 100$)
Articles, *n*	100	100
Full text available, *n* (%)	63 (63.0%)	63 (63.0%)
Full text unavailable for complete verification, *n* (%)	37 (37.0%)	37 (37.0%)
Low statistical complexity, *n* (%)	3 (3.0%)	3 (3.0%)
Moderate statistical complexity, *n* (%)	78 (78.0%)	62 (62.0%)
High statistical complexity, *n* (%)	19 (19.0%)	35 (35.0%)
Journals represented, *n*	10	10
Articles per journal-year stratum, *n*	10	10

Notes: Values are article counts and percentages within each 100-article year sample. Full-text availability denotes complete verification of the article. Complexity uses the final harmonized operational definition and is a descriptive characteristic, not a primary reporting-quality outcome; pre-harmonization distributions are reported in the text.

2.10. Analysis Plan and Reproducibility

The retained analysis plan was finalized after descriptive review but before inferential modeling and was not prospectively preregistered. The uncorrected MH/CMH estimator was retained as the primary analytical specification. The source-accessibility checks and post hoc analyses described in Sections 2.8 and 2.9 did not replace this specification or alter the fixed family of 18 comparisons subjected to BH-FDR adjustment.

Public reproducibility files contain coded article-level data for all 200 sampled publications, including an original-code analytical comparator needed by the reproduction script, the 19-rule operational codebook, two Python source files, dependency requirements, and machine-readable primary and sensitivity outputs. The separate frozen historical workbook, individual auditor worksheets, human-adjudication records, detailed discrepancy registers and publisher PDFs are retained outside the publication supplement. The analyses were run using Python 3.12.14, NumPy 2.3.5, pandas 2.2.3 and SciPy 1.17.0. Reproducible code does not establish source completeness or post-harmonization reliability.

2.11. Generative Artificial Intelligence Assistance

OpenAI ChatGPT-6.1 Sol assisted with language editing, manuscript organization, statistical code development and troubleshooting, and consistency checking. The primary author reviewed the accessible source material and performed the original article-level coding. Two external assessors independently evaluated the Round-2 reliability sample, while a separate pair of assessors subsequently reassessed selected coding discrepancies, with three unresolved items referred to an independent third adjudicator, as described in Section 2.5. The authors reviewed the outputs and are responsible for the coding decisions, analyses, interpretation and final text.

3. Results

3.1. Study Sample and Statistical Complexity

The 200 audited articles were evenly distributed across the ten journals and two assigned study-year strata (100 articles per year; Supplementary Table S6). Complete full text was independently verified for 63/100 articles in each year; verification remained incomplete for 37/100 articles in each year (Table 2). Twenty-three historical article-domain discrepancies involving six source-limited articles remained unresolved.

High statistical complexity under the locked rule was 19/100 (2015) and 35/100 (2025). The historical pre-harmonization counts were 66/100 and 78/100; 92 articles moved from high to moderate and two from moderate to high. These reclassifications, including conventional Cox/logistic analyses formerly deemed high, materially affect complexity-adjusted post hoc models. They do not reflect independently confirmed improvement or worsening in study design.

3.2. Inter-Rater Reliability and Targeted Adjudication

Round-2 agreement varied by domain. Table 3 pairs the six nominal outcomes with the corresponding binary reliability estimates. For exact *p*-value reporting, 12/17 pairs agreed (70.6%, Wilson 95% CI 46.9–86.7%), κ was −0.10 (95% CI −0.25 to 0.00) and AC1 was 0.61 (95% CI 0.13–0.87). For causal overstatement, all 16 evaluable pairs agreed (Wilson 95% CI 80.6–100%), with one concordant positive pair. Both bootstrap coefficient intervals collapsed to 1.00 in this small sample. Supplementary Table S2 reports results across all domains. The subsequent, separate 33-item adjudication yielded 23 concordant initial ratings, seven discussion-based resolutions and three third-adjudicator decisions. Eleven codes were changed. This exercise did not estimate reliability of the final 200-article classifications.

Table 3. Journal-stratified comparisons of reporting and interpretation in the 2025 and 2015 samples.

Outcome	2015 *n*/N (%)	2025 *n*/N (%)	Δ pp	MH OR (95% CI)	CMH *p*	BH q	Journals Better/Worse/Tie
Sample-size justification	9/99 (9.1%)	12/91 (13.2%)	+4.1	1.48 (0.56–3.90)	0.430	0.517	3/3/4
Descriptive statistics	100/100 (100.0%)	95/100 (95.0%)	−5.0	0.00 (CI NE)	0.024	0.107	0/4/6
Effect estimate	63/100 (63.0%)	74/98 (75.5%)	+12.5	1.95 (1.03–3.72)	0.036	0.113	5/2/3
CI/uncertainty	48/98 (49.0%)	58/92 (63.0%)	+14.1	1.94 (1.03–3.63)	0.038	0.113	7/2/1
Exact *p*-value reporting	63/92 (68.5%)	81/95 (85.3%)	+16.8	2.55 (1.23–5.28)	0.009	0.053	7/2/1
Methods identified	87/98 (88.8%)	88/99 (88.9%)	+0.1	1.01 (0.40–2.54)	0.989	0.989	4/3/3
Assumptions reported	7/65 (10.8%)	10/65 (15.4%)	+4.6	1.72 (0.57–5.15)	0.332	0.493	6/2/2
Model specification	59/81 (72.8%)	59/84 (70.2%)	−2.6	0.84 (0.42–1.70)	0.608	0.643	3/6/1
Missing-data handling	25/64 (39.1%)	18/69 (26.1%)	−13.0	0.55 (0.25–1.18)	0.124	0.248	3/6/1
Multiplicity handling	13/84 (15.5%)	16/67 (23.9%)	+8.4	1.78 (0.72–4.39)	0.202	0.364	7/1/2
Sensitivity/robustness	16/73 (21.9%)	27/75 (36.0%)	+14.1	2.12 (0.93–4.83)	0.072	0.185	6/2/1 (+1 NE)

Table 3. *Cont.*

Outcome	2015 *n*/N (%)	2025 *n*/N (%)	Δ pp	MH OR (95% CI)	CMH *p*	BH q	Journals Better/Worse/Tie
Software identified	50/68 (73.5%)	55/66 (83.3%)	+9.8	1.48 (0.61–3.57)	0.387	0.497	7/3/0
Software version	37/100 (37.0%)	47/100 (47.0%)	+10.0	1.62 (0.88–3.00)	0.116	0.248	5/2/3
Nonsignificant = no effect	25/100 (25.0%)	28/97 (28.9%)	+3.9	1.23 (0.63–2.41)	0.546	0.614	3/3/4
Trend/tendency language	17/100 (17.0%)	5/97 (5.2%)	−11.8	0.27 (0.10–0.76)	0.008	0.053	7/2/1
Causal overstatement	15/100 (15.0%)	32/97 (33.0%)	+18.0	2.69 (1.35–5.35)	0.003	0.053	2/8/0
Numerical discrepancy	10/97 (10.3%)	6/85 (7.1%)	−3.3	0.61 (0.21–1.77)	0.356	0.493	5/2/3
Any major problem	12/100 (12.0%)	7/85 (8.2%)	−3.8	0.60 (0.22–1.59)	0.283	0.463	4/4/2

Notes. Each year shows events divided by evaluable articles and the corresponding percentage. Evaluable counts exclude domain-specific codes 8 (not applicable) and 9 (unclear). Δ pp is the 2025 percentage minus the 2015 percentage. ORs and 95% CIs are from uncorrected journal-stratified Mantel–Haenszel estimation and the Robins–Breslow–Greenland variance. *p* values are from the uncorrected CMH score test, and BH q values cover the 18 primary comparisons. The descriptive-statistics OR is at the boundary (0, CI not estimable [NE]). The post hoc stratified exact *p* was 0.0592. Journal counts show better, worse, or tied reporting quality, with NE shown where a journal contrast could not be evaluated. Round-2 binary agreement for the six outcomes with nominal $p < 0.05$ (evaluable pairs, raw agreement, κ, AC1): descriptive statistics (19, 84.2%, 0.34, 0.80), effect estimates (19, 68.4%, 0.30, 0.44), uncertainty (17, 70.6%, 0.41, 0.42), exact *p* values (17, 70.6%, −0.10, 0.61), trend language (16, 100%, 1.00, 1.00), and causal overstatement (16, 100%, 1.00, 1.00). The last outcome had one concordant positive pair and 15 concordant negative pairs. Supplementary Table S2 reports 95% intervals.

Illustrative adjudication decisions demonstrate how the operational rules were applied. For P004, uncertainty reporting was changed from 9 (not verifiable) to 2 (adequate) because the principal regression estimates were accompanied by 95% confidence intervals. For P226, uncertainty reporting was changed from 0 to 2 because the principal regression results included standard errors and confidence intervals, satisfying the codebook criterion for appropriate uncertainty reporting. For P203, the major-problem indicator was changed from 1 to 0 because the identified numerical and reporting discrepancies did not meet the predefined threshold for a serious problem likely to materially affect interpretation. For P160, multiplicity handling was changed from 0 to 2 following documentation of correction for multiple comparisons in the principal analyses. These decisions were based on the unchanged operational definitions in Supplementary Table S3 and were incorporated into the final dataset through the documented adjudication process.

3.3. Statistical Reporting in the Two Samples

Six of the 18 primary uncorrected journal-stratified comparisons had nominal CMH $p < 0.05$. None met the Benjamini–Hochberg criterion $q < 0.05$ (Table 3 and Figure 1). Table 3 also gives Round-2 agreement for the six nominal outcomes.

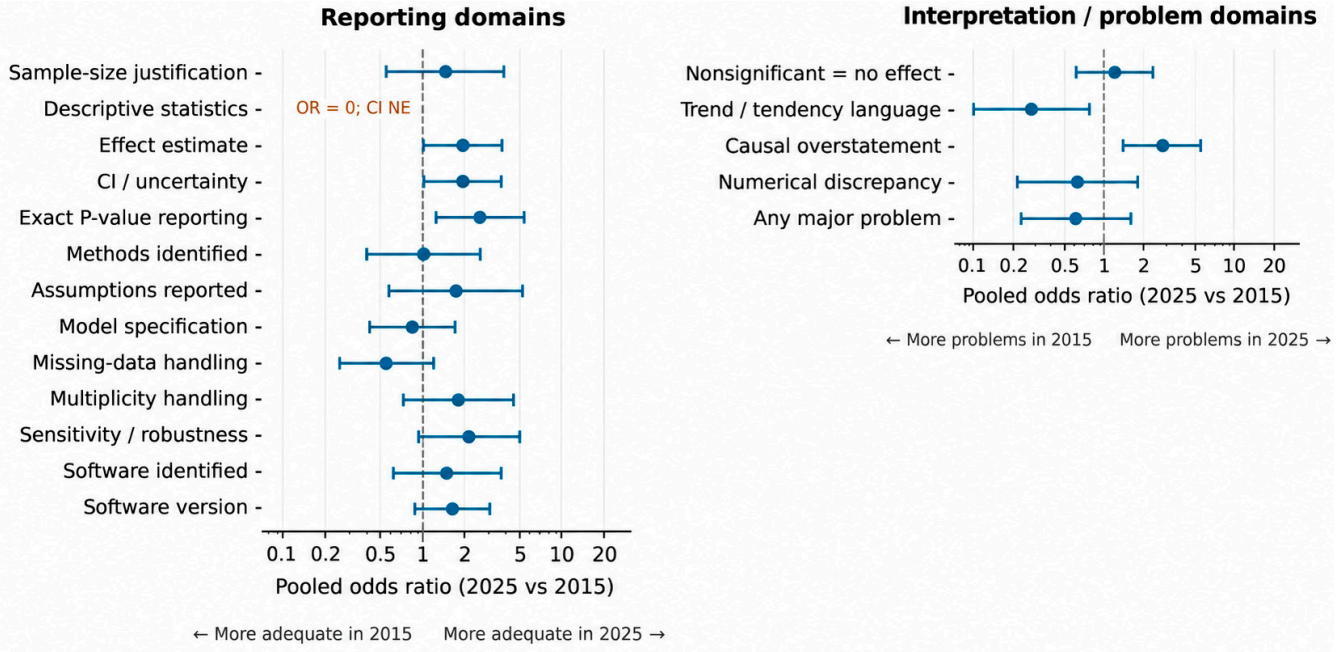


Figure 1. Uncorrected journal-stratified Mantel–Haenszel odds ratios and 95% Robins–Breslow–Greenland confidence intervals for the 2025 versus 2015 samples. The dashed line marks OR = 1 on a logarithmic axis. Descriptive-statistics reporting had a boundary OR of 0 and no estimable log-scale interval. The direction labels are specific to the reporting and interpretation panels. None of the 18 primary comparisons met BH q < 0.05.

Adequate effect-estimate reporting was 63/100 (63.0%) in 2015 versus 74/98 (75.5%) in 2025, with uncorrected MH OR 1.95 (95% CI 1.03–3.72), CMH $p = 0.036$ and BH q = 0.113.

Adequate uncertainty reporting was 48/98 (49.0%) versus 58/92 (63.0%), with an OR of 1.94 (1.03–3.63), $p = 0.038$ and q = 0.113.

Adequate exact *p*-value reporting was 63/92 (68.5%) versus 81/95 (85.3%), with an OR of 2.55 (1.23–5.28), $p = 0.009$ and q = 0.053. All three nominal numerical-reporting contrasts remain exploratory.

The corresponding journal-level descriptive directions after adjudication were five improved, two worsened and three unchanged for effect estimates; seven improved, two worsened and one unchanged for uncertainty; and seven improved, two worsened and one unchanged for exact *p* values. These descriptive patterns are not independent confirmations of pooled inference.

Sample-size justification was 9/99 versus 12/91 (OR 1.48, $p = 0.430$). Fully adequate assumptions reporting was 7/65 versus 10/65 (OR 1.72, $p = 0.332$), and missing-data handling was 25/64 versus 18/69 (OR 0.55, $p = 0.124$); their confidence intervals crossed the null. There was no FDR-confirmed contrast.

Fully adequate multiplicity handling was 13/84 versus 16/67 (OR 1.78, $p = 0.202$); sensitivity/robustness reporting was 16/73 versus 27/75 (OR 2.12, $p = 0.072$). Descriptive-statistics reporting was 100/100 versus 95/100, producing an uncorrected boundary OR = 0 (CI NE), CMH $p = 0.024$ and q = 0.107. Its previously documented post hoc stratified exact $p = 0.0592$ (unstratified Fisher $p = 0.0594$) illustrates asymptotic fragility; the CMH *p* remains in the unchanged 18-test primary BH family.

3.4. Interpretation and Reporting Problems

The interpretation outcomes describe the frequency of coded problems among evaluable articles in each selected sample.

Trend/tendency language occurred in 17/100 versus 5/97 articles (OR 0.27, 95% CI 0.10–0.76, $p = 0.008$, q = 0.053), a nominally less frequent adverse practice in 2025.

Causal overstatement was coded in 15/100 versus 32/97 articles (OR 2.69, 95% CI 1.35–5.35, $p = 0.003$, q = 0.053). This nominal contrast may reflect differences in sample composition or in the verifiability or reproducibility of coding.

Nonsignificant-as-no-effect interpretation was 25/100 versus 28/97 (OR 1.23, $p = 0.546$). Any major problem was 12/100 versus 7/85 (OR 0.60, 95% CI 0.22–1.59, $p = 0.283$), following the validated P203 code change. Neither met the nominal threshold.

3.5. Multiple-Testing Adjustment

All 18 primary comparisons had BH-adjusted q ≥ 0.05. The minimum q was 0.0532 (exact *p*-value reporting, trend language, and causal overstatement); q values for effect-estimate and uncertainty reporting were approximately 0.1126, and the descriptive-statistics boundary comparison had q = 0.1072. In the separate post hoc sparse-data sensitivity analysis using the historical 0.5 correction, the minimum BH-adjusted q was 0.0748, with no FDR-significant comparisons. These sensitivity results were reported separately and did not replace the primary estimates.

Accordingly, associations with nominal $p < 0.05$ were interpreted as exploratory rather than FDR-confirmed differences.

3.6. Secondary and Complexity-Adjusted Analyses

Post hoc secondary conditional estimates were finite for 17/18 outcomes (descriptive statistics remained at a boundary). Firth journal-plus-categorical-complexity models were finite for 18/18. Their inferential interpretations differ from standard MH and are not independent confirmation or replacements for the fixed 18-test BH results (Table 4; complete outputs are provided in Supplementary Code S1).

Table 4. Selected secondary models and exploratory heterogeneity checks.

Outcome	Primary MH OR	Conditional Journal OR (95% CI)	Wald *p*	Firth Journal + Complexity OR (95% CI)	PLR *p*	Breslow–Day *p*	Comment
Effect estimate	1.95	2.00 (1.04–3.85)	0.037	1.93 (1.00–3.76)	0.048	0.0497	Post hoc; exploratory
CI/uncertainty	1.94	1.93 (1.03–3.59)	0.039	2.16 (1.14–4.18)	0.018	0.329	Post hoc; exploratory
Exact *p*-value reporting	2.55	2.64 (1.26–5.54)	0.010	2.90 (1.39–6.37)	0.004	0.134	Post hoc; exploratory
Trend/tendency language	0.27	0.27 (0.10–0.76)	0.013	0.32 (0.11–0.82)	0.016	0.184	Post hoc; exploratory
Causal overstatement	2.69	2.83 (1.40–5.75)	0.004	2.73 (1.35–5.72)	0.005	0.040	Post hoc; exploratory
Model specification	0.84	0.82 (0.39–1.73)	0.608	0.74 (0.34–1.56)	0.428	0.001	Post hoc; exploratory
Software identified	1.48	1.47 (0.61–3.52)	0.388	1.53 (0.65–3.70)	0.335	0.397	Post hoc; exploratory

Notes: The primary column reports uncorrected MH estimates from the final 200-article analytical dataset. Conditional journal–model intervals/*p* values are model-based Wald estimates. Firth journal-plus-final-categorical-complexity intervals use the penalized profile, and its *p* values use the penalized likelihood ratio. Breslow–Day checks are exploratory and sparse-cell sensitive. These estimators address different likelihood conventions and must not be substituted for the primary BH-adjusted results.

Firth journal-plus-complexity estimates for selected domains are displayed in Table 4. Their numerical differences from uncorrected MH reflect model choice and adjustment as well as updated coding; complexity itself had undergone 94 prior harmonization recodes. Because otherwise identical adjusted models were not fitted using both the historical and harmonized complexity classifications, the numerical effect attributable specifically to the 94 reclassifications cannot be isolated.

Even where alternative model p values are nominally small, they remain exploratory. Conditioning on journal margins and applying a Firth penalty do not repair source-access or coding-reliability limitations.

Exploratory, unadjusted Breslow–Day tests yielded nominal p values below 0.05 for model specification ($p = 0.0013$), software version reporting ($p = 0.0373$), causal overstatement ($p = 0.0398$), and effect-estimate reporting ($p = 0.0497$). These findings were not adjusted for multiplicity and may be unstable in sparse journal-specific strata. They therefore do not establish reliable evidence of between-journal effect modification.

3.7. Sensitivity Analyses

The sensitivity analyses address different questions. Restricting the analysis to verified complete texts changes which articles enter the comparison. Including code 9 under a specified assumption changes which records are evaluable and how unclear records are classified. Lowering the adequacy threshold changes the event from full adequacy to at least partial reporting in the same outcome-specific analysis set (Table 5 and Supplementary Tables S4 and S5).

Table 5. Selected sensitivity-analysis estimates.

Outcome	Primary OR (95% CI)	Full-Text-Only OR (95% CI)	FT Direction	Conservative OR (95% CI)	Conservative Direction	Interpretation
Effect estimate	1.95 (1.03–3.72)	1.28 (0.52–3.15)	More adequate	1.72 (0.92–3.20)	More adequate	Scenario-dependent; exploratory
CI/uncertainty	1.94 (1.03–3.63)	2.34 (0.99–5.52)	More adequate	1.56 (0.86–2.80)	More adequate	Scenario-dependent; exploratory
Exact *p*-value reporting	2.55 (1.23–5.28)	4.96 (1.61–15.28)	More adequate	2.45 (1.26–4.77)	More adequate	Scenario-dependent; exploratory
Trend/tendency language	0.27 (0.10–0.76)	0.25 (0.06–0.95)	Fewer problems	N/A	N/A	Scenario-dependent; exploratory
Causal overstatement	2.69 (1.35–5.35)	4.20 (1.55–11.33)	More problems	N/A	N/A	Scenario-dependent; exploratory
Descriptive statistics	0.00 (CI NE)	NE (CI NE)	NE/boundary	0.00 (CI NE)	NE/boundary	Scenario-dependent; exploratory
Any major problem	0.60 (0.22–1.59)	1.39 (0.33–5.96)	More problems	N/A	N/A	Scenario-dependent; exploratory

Notes: Full-text restriction includes the 126 articles with verified complete text. Conservative code-9 analyses treat unclear as not adequate for the 13 positive domains; code 8 remains excluded. All selected estimates use uncorrected journal-stratified MH. Code-9 favorable/adverse scenarios and the post hoc 0.5 sparse-data convention are presented in Supplementary Code S1. NE denotes non-estimable. None of these sensitivities replaces the primary FDR family.

In the 126-article verified-full-text subset, fully adequate effect-estimate reporting was observed in 49/63 versus 50/63 articles. The OR attenuated from 1.95 (95% CI 1.03–3.72) to 1.28 (0.52–3.15, $p = 0.573$). The corresponding full-text estimates were 2.34 for uncertainty reporting (0.99–5.52, $p = 0.0499$), 4.96 for exact p-value reporting (1.61–15.28, $p = 0.003$), 0.25 for trend language (0.06–0.95, $p = 0.041$), and 4.20 for causal overstatement (1.55–11.33, $p = 0.002$). These estimates are conditional on verified access.

Under conservative code-9-as-not-adequate classification, effect-estimate reporting had OR 1.72 (0.92–3.20), $p = 0.079$; uncertainty reporting had OR 1.56 (0.86–2.80), $p = 0.137$; and exact p reporting had OR 2.45 (1.26–4.77), $p = 0.007$. Alternative assumptions about unverified records change both size and precision.

Under the post hoc threshold of at least partial reporting, the effect-estimate OR was 1.00 (95% CI 0.22–4.50), with 96/100 versus 94/98 articles meeting this threshold. The missing-data-handling OR changed direction, from 0.55 (0.25–1.18) for full adequacy to 1.99 (0.76–5.25) for at least partial reporting (51/64 versus 60/69). Both missing-data intervals include 1 and are compatible with lower and higher odds in 2025. The alternative effect-estimate interval is also too wide to establish equivalence. Full results for the 13 domains and the favorable and adverse code-9 scenarios are in Supplementary Tables S4 and S5 and the machine-readable outputs.

4. Discussion

The clearest descriptive result was the infrequency of fully adequate reporting in several methodological domains in both selected samples. Six of 18 outcomes had nominal associations, and none met the primary BH-FDR threshold. The 11 human-adjudicated coding corrections attenuated the effect-estimate and uncertainty associations. Sampling, access to sources, article composition and coding reproducibility constrain the interpretation of the remaining contrasts.

4.1. Numerical Reporting

The adequate-only ORs were 1.95 for effect estimates and 1.94 for uncertainty. The effect-estimate OR fell to 1.28 in the availability-selected subset and to 1.00 under the alternative threshold. The subset changes the population under study. The lower threshold measures a different aspect of reporting and yields a high event frequency in both samples. The resulting point estimates and intervals demonstrate dependence on source availability and outcome definition.

4.2. Analytical Transparency

Among evaluable articles, fully adequate sample-size justification (9/99 vs. 12/91), assumptions reporting (7/65 vs. 10/65), missing-data handling (25/64 vs. 18/69), and multiplicity handling (13/84 vs. 16/67) were uncommon in both selected samples. The denominators reflect outcome-specific evaluability.

Related article-level audits also found limited reporting of sample-size calculations in their own settings. Hogue et al. [7] recorded 35 mentions in 241 preclinical animal studies from 99 journals (14.7% after weighting for journal clustering). Héroux et al. [9] recorded a calculation in 15 of 556 articles assessed for that item (2.7%) within an audit of 580 original *Journal of Neurophysiology* articles published in 2019–2020. The present definition of full adequacy also permits an explicit design-based rationale for a fixed sample. Differences in audit populations, denominators and criteria preclude a direct comparison of rates.

Incomplete reporting of analytical decisions restricts assessment of methodological appropriateness and the potential influence of alternative approaches [12–15,26]. STROBE similarly recommends transparent descriptions of statistical methods, missing-data han-

dling, and sensitivity analyses [27], although overall STROBE adherence was not assessed in this study.

More complex analyses may require additional information about model structures, variable selection, repeated measurements, and missingness assumptions. Analytical sophistication can therefore coexist with limited reporting transparency, but the present audit cannot establish a causal relationship between complexity and reporting deficiencies.

4.3. Statistical Complexity and Coding

The 94 statistical-complexity reclassifications illustrate the influence of operational definitions on an analytical covariate rather than changes in the underlying studies. Complexity describes analytical structure, not methodological quality. Accordingly, the secondary complexity-adjusted models remain exploratory and potentially sensitive to classification uncertainty. They cannot establish whether statistical complexity explains the observed differences between the sampled years.

4.4. Interpretation and Heterogeneity

Causal overstatement was coded in 15.0% of evaluable 2015 articles and 33.0% of evaluable 2025 articles. This contrast and the lower frequency of trend language were nominal findings (both $q = 0.0532$). Differences in the mix of study designs or in how authors described causal questions could contribute to the observed contrast. For example, observational studies and secondary analyses may present different opportunities for causal overstatement. These are hypotheses.

The archived study-design descriptions were free text and were not independently classified. We therefore cannot establish year-specific proportions of observational studies or secondary analyses. High statistical complexity was recorded for 19/100 and 35/100 articles, respectively. In the same exploratory journal fixed-effects specification, adding categorical complexity moved the causal-overstatement OR from 3.00 to 2.98. Complexity describes analytical structure. The small change in this estimate cannot establish a mechanism or substitute for design classification [16–18].

The use of the same ten journals in both years limits variation in journal identity but does not account for changes in article types, analytical approaches, editorial practices, or candidate-pool composition. Breslow–Day heterogeneity tests were exploratory, unadjusted for multiplicity, and potentially unstable in sparse journal-specific strata. Their nominal results therefore do not support definitive claims about between-journal differences.

4.5. Limitations

The journal panel and candidate pools were selected pragmatically. Incomplete search and screening records prevent assessment of pool coverage and whether the two samples were assembled comparably. The 2015 *Journal of the Neurological Sciences* pool contributed all ten candidates. Historical records do not establish whether issue year and first-online year were applied consistently. Equal allocation and journal stratification address differences in journal identity while leaving potential differences in article design, content and source availability within journals.

Complete full-text verification was unavailable for 74/200 articles (37 per year), and 71/200 had at least one unclear outcome code. Equal counts of source-limited articles do not establish equal verifiability by outcome or journal. For example, 15 any-major-problem classifications were unclear in 2025 and none in 2015. Assigning all unclear cases to the adverse category changed the OR from 0.60 to 2.01. The estimate depends on the assumed classifications. Twenty-three discrepancies in six source-limited articles remain unresolved.

The full-text restriction changes the study population and neither sensitivity approach can determine the direction or extent of access-related selection or measurement bias.

One author coded the main 200-article sample. The independent 20-article Round-2 assessment found modest or low agreement in several headline domains, and causal-overstatement agreement was based on only one concordant positive pair. The separate 33-item adjudication resolved selected discrepancies. It provides no estimate of reproducibility for the final sample as a whole. The assessment forms omitted the publication-year column, but publication-year information remained visible in the source PDFs (Section 2.5).

Dichotomization combined absent and partially adequate reporting, discarding their ordinal distinction. The post hoc alternative-threshold analysis demonstrated that some estimates depend materially on this decision. The software version outcome also requires caution: under the locked coding rule, code 0 may include records with unverified software identity, making its denominator of 200 conceptually different from the 134 evaluable records used for software identification.

The analysis plan was finalized after descriptive inspection rather than prospectively registered. The historical 0.5 correction, alternative-threshold analyses, conditional and Firth models, and exact boundary check were post hoc or exploratory. None provides independent confirmation of findings that failed the primary FDR criterion.

Finally, reporting completeness and analytical validity are different constructs. Incomplete reporting does not establish that an analysis was incorrect, just as complete reporting does not establish methodological validity.

4.6. Implications and Future Research

The observed reporting gaps suggest several practical priorities for clinical neurology journals. Author and statistical-reviewer checklists could request (i) sample-size or precision justification; (ii) principal effect estimates with confidence intervals and appropriately reported *p* values; (iii) explicit model specifications, adjustment variables, and software identification and version; (iv) the extent and management of missing data; (v) relevant assumptions and diagnostic checks; (vi) multiplicity considerations; (vii) clear identification of prespecified and exploratory sensitivity analyses; and (viii) interpretation consistent with the underlying study design [12–15,19,26]. These recommendations address deficiencies identified in the sampled articles but have not themselves been evaluated as editorial interventions.

The article-level audit contributes an operational coding framework, reproducible analytical data and sensitivity analyses for a fixed panel of ten clinical neurology journals. The checklist addresses the reporting gaps observed in the sampled articles.

Future investigations would benefit from prospectively documented sampling, complete source accessibility, independent duplicate coding, predefined publication-year attribution, and validated study-design classification. These improvements would support more reliable comparisons of reporting practices across publication years and study types.

5. Conclusions

This audit describes statistical reporting and interpretation in two selected publication-year samples from ten clinical neurology journals. None of the 18 primary 2015–2025 comparisons met the BH-FDR threshold (minimum q = 0.0532). Nominal associations were exploratory.

Fully adequate reporting among evaluable articles remained uncommon in both samples. The 2015 and 2025 frequencies were 9.1% and 13.2% for sample-size justification, 10.8% and 15.4% for assumptions, 39.1% and 26.1% for missing-data handling, and 15.5% and 23.9% for multiplicity. In the conservative code-9 scenario, the corresponding assumptions

rates were 8.0% and 10.6% and multiplicity rates were 13.7% and 16.7%. Section 4.6 provides a practical editorial checklist addressing these deficiencies.

Uncertainty about pool coverage, the composition of the two samples, incomplete full-text verification and modest coding agreement limits inference to the selected articles and the operational definitions used. The comparison cannot establish a continuous temporal trend, population-wide improvement, or an effect of editorial policies.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/medsci14080804/s1: Supplementary Table S1: Documented journal-year sampling pools, replacements, and final selected articles; Supplementary Table S2: Final Round-2 inter-rater reliability by coding domain; Supplementary Table S3: Operational definitions and final coding rules; Supplementary Table S4: Post hoc reporting-threshold sensitivity for 13 ordinal domains; Supplementary Table S5: Extreme-case sensitivity analyses for unclear classifications; Supplementary Table S6: Bibliographic characteristics of the 200 articles included in the main audit (Supplementary_Tables_S1–S6.docx); Supplementary Data S1: Article-level coded analytical data for the 200 selected articles, an original-code analytical comparator, and the operational codebook (coded_articles_200.json; coded_articles_final_200.csv; Operational_Codebook_19.csv). Supplementary Code S1: Python reproduction scripts, dependency requirements, instructions, and machine-readable primary, sensitivity, and secondary analysis results (Supplementary_Data_and_Code.zip).

Author Contributions: Conceptualization, R.R.A. and A.K.M.; methodology, R.R.A., I.M.K., V.G.A., A.K.M., S.N.M., A.K., S.E.A. and N.V.G.; software, R.R.A.; validation, I.M.K., V.G.A., S.N.M., A.K., S.E.A. and N.V.G.; investigation, R.R.A., I.M.K. and V.G.A.; formal analysis, R.R.A., A.K.M. and N.V.G.; writing—original draft preparation, R.R.A., S.N.M., I.M.K., A.K. and S.E.A.; writing—review and editing, R.R.A., I.M.K., V.G.A., A.K.M., S.N.M., A.K., S.E.A. and N.V.G.; visualization, R.R.A., S.N.M., V.G.A., N.V.G., A.K., S.E.A. and A.K.M.; supervision, R.R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author(s). The two Python scripts in Supplementary Code S1 (reproduce_reporting.py and analysis_methods.py), together with the supplied requirements file and coded_articles_200.json, provide the reported journal-stratified MH/CMH/BH primary comparisons and post hoc 0.5-corrected, source-availability, classification-threshold and secondary analyses. The original private source-verification files are not required to rerun these statistical computations, but the computations cannot independently verify unavailable article content.

Acknowledgments: The authors gratefully acknowledge Rovshan L. Hasanov and Zamina M. Sadikhova for their independent external assessment of the audit framework, including participation in the calibration and inter-rater reliability exercises. Their careful application of the operational codebook and documentation of coding discrepancies contributed to the refinement and harmonization of the audit procedures. The authors also sincerely thank Khanlar I. Hasanov and Madina R. Tagiyeva for their independent reassessment of 33 selected article-domain coding items and participation in the subsequent discrepancy-resolution process. Special thanks are extended to Rana K. Shiraliyeva for serving as the third adjudicator and providing independent determinations on three items that remained unresolved after the initial assessments and discussions. These contributors did not participate in the primary coding of the 200-article study sample, inferential statistical analysis, interpretation of the study-level statistical findings, or manuscript preparation. During preparation of this manuscript, the authors used OpenAI ChatGPT-6.1 Sol for the purposes described in Section 2.11, reviewed and edited its outputs, and take full responsibility for the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lang, T.A.; Altman, D.G. Basic statistical reporting for articles published in biomedical journals: The "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. *Int. J. Nurs. Stud.* **2015**, *52*, 5–9. [CrossRef] [PubMed]
2. International Committee of Medical Journal Editors. *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*; Updated January 2026; International Committee of Medical Journal Editors: Philadelphia, PA, USA, 2026.
3. Wasserstein, R.L.; Lazar, N.A. The ASA Statement on p-values: Context, process, and purpose. *Am. Stat.* **2016**, *70*, 129–133. [CrossRef]
4. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a world beyond "$p < 0.05$.". *Am. Stat.* **2019**, *73*, 1–19. [CrossRef]
5. Harrington, D.; D'Agostino, R.B., Sr.; Gatsonis, C.; Hogan, J.W.; Hunter, D.J.; Normand, S.L.T.; Drazen, J.M.; Hamel, M.B. New Guidelines for Statistical Reporting in the Journal. *N. Engl. J. Med.* **2019**, *381*, 285–286. [CrossRef] [PubMed]
6. Gore, S.M.; Jones, I.G.; Rytter, E.C. Misuse of statistical methods: Critical assessment of articles in BMJ from January to March 1976. *Br. Med. J.* **1977**, *1*, 85–87. [CrossRef] [PubMed]
7. Hogue, O.; Harvey, T.; Crozier, D.; Sonneborn, C.; Postle, A.; Block-Beach, H.; Somasundaram, E.; May, F.J.; Snyder Braun, M.; Pasadyn, F.L.; et al. Statistical practice and transparent reporting in the neurosciences: Preclinical motor behavioral experiments. *PLoS ONE* **2022**, *17*, e0265154. [CrossRef] [PubMed]
8. Olarte Parra, C.; Bertizzolo, L.; Schroter, S.; Dechartres, A.; Goetghebeur, E. Consistency of causal claims in observational studies: A review of papers published in a general medical journal. *BMJ Open* **2021**, *11*, e043339. [CrossRef] [PubMed]
9. Héroux, M.; Diong, J.; Bye, E.; Fisher, G.; Robertson, L.; Butler, A.; Gandevia, S. Poor statistical reporting, inadequate data presentation and spin persist despite Journal awareness and updated Information for Authors. *F1000research* **2023**, *12*, 1483. [CrossRef] [PubMed]
10. Deriu, F.; Martinez, G.; Loi, N.; Ventura, L.; Ginatempo, F.; Dvir, Z.; Manca, A. Reporting quality of TMS studies in neurological conditions: A critical appraisal of the main gaps, challenges and clinical implications. *J. Neurosci. Methods* **2021**, *362*, 109293. [CrossRef] [PubMed]
11. Tumma, S.; Sonneborn, C.; Padate, A.; Hogue, O. Statistical guidance provided to authors by clinical neurology and neuroscience journals. *Eur. J. Neurol.* **2023**, *30*, 3675–3681. [CrossRef] [PubMed]
12. Mansournia, M.A.; Nazemipour, M. Recommendations for accurate reporting in medical research statistics. *Lancet* **2024**, *403*, 611–612. [CrossRef] [PubMed]
13. Ordak, M. Implementation of SAMPL guidelines: Recommendations for improving statistical reporting in biomedical journals. *Clin. Med.* **2025**, *25*, 100304. [CrossRef] [PubMed]
14. Rovetta, A.; Piretta, L.; Mansournia, M.A. p-Values and confidence intervals as compatibility measures: Guidelines for interpreting statistical studies in clinical research. *Lancet Reg. Health-Southeast Asia* **2025**, *33*, 100534. [CrossRef] [PubMed]
15. Hopewell, S.; Chan, A.W.; Collins, G.S.; Hróbjartsson, A.; Moher, D.; Schulz, K.F.; Tunn, R.; Aggarwal, R.; Berkwits, M.; Berlin, J.A.; et al. CONSORT 2025 Statement: Updated Guideline for Reporting Randomized Trials. *JAMA* **2025**, *333*, 1998–2005. [CrossRef] [PubMed]
16. Dahabreh, I.J.; Bibbins-Domingo, K. Causal Inference About the Effects of Interventions From Observational Studies in Medical Journals. *Jama* **2024**, *331*, 1845–1853. [CrossRef] [PubMed]
17. Cashin, A.G.; Hansford, H.J.; Hernán, M.A.; Swanson, S.A.; Lee, H.; Jones, M.D.; Dahabreh, I.J.; Dickerman, B.A.; Egger, M.; Garcia-Albeniz, X.; et al. Transparent Reporting of Observational Studies Emulating a Target Trial—The TARGET Statement. *Jama* **2025**, *334*, 1084–1093. [CrossRef] [PubMed]
18. Labrecque, J.A.; Kezios, K.L. How and when to use causal and associational language. *BMJ* **2026**, *392*, e085749. [CrossRef] [PubMed]
19. Lozada-Martinez, I.D.; Delgado Arias, L.M.; Perea Rojas, D.M.; Rojas Torres, I.L. Meta-Research in Neuroscience: An Urgent Call to Strengthen the Reliability and Translation of Knowledge into Evidence-Based Neurological Practice. *J. Clin. Med.* **2025**, *14*, 8552. [CrossRef] [PubMed]
20. Ioannidis, J.P.A.; Fanelli, D.; Dunne, D.D.; Goodman, S.N. Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLoS Biol.* **2015**, *13*, e1002264. [CrossRef] [PubMed]
21. Gwet, K.L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48. [CrossRef] [PubMed]
22. Feinstein, A.R.; Cicchetti, D.V. High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [CrossRef] [PubMed]
23. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300. [CrossRef]

24. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **1993**, *80*, 27–38. [CrossRef]
25. Heinze, G.; Schemper, M. A solution to the problem of separation in logistic regression. *Stat. Med.* **2002**, *21*, 2409–2419. [CrossRef] [PubMed]
26. Srinivasan, A.; Berkowitz, J.; Friedrich, N.A.; Kivelson, S.; Tatonetti, N.P. Large Language Model Analysis of Reporting Quality of Randomized Clinical Trial Articles: A Systematic Review. *JAMA Netw. Open* **2025**, *8*, e2529418. [CrossRef] [PubMed]
27. von Elm, E.; Altman, D.G.; Egger, M.; Pocock, S.J.; Gøtzsche, P.C.; Vandenbroucke, J.P.; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *PLoS Med.* **2007**, *4*, e296. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.