Ten Points for High-Quality Statistical Reporting and Data Presentation

Pentti Nieminen

Medical Informatics and Data Analysis Research Group, University of Oulu, 90014 Oulu, Finland;
pentti.nieminen@oulu.fi

Featured Application: In this work, an applicable instrument is proposed to quickly evaluate the quality of the statistical reporting and data presentation in research papers.

Abstract: Background: Data analysis methods have become an essential part of empirical research papers, especially in health sciences and medical research. It has previously been reported that a noteworthy percentage of articles have flaws in their statistical reporting. Reporting problems have been a long-term issue, and despite continued efforts to improve the situation, improvements have been far from satisfactory. One explanation is an inadequate assessment of statistical reporting during peer review. This communication proposes a short instrument to assess the quality of data analysis reporting in manuscripts and published papers. Method: A checklist-type instrument was developed by selecting and refining items from previous reports about the quality of statistical reporting in medical journals and from published guidelines for reporting and data presentation. Items were pretested and modified during pilot studies. A total of 160 original medical research articles that were published in 4 journals were evaluated to test the instrument. Interrater and intrarater agreements were examined by comparing quality scores assigned to 40 articles published in a psychiatric journal. Results: The data analysis reporting test consists of nine questions that assess the quality of health research from a reader's perspective. The composed scale has a total score ranging from 0 to 10 and discriminated between journals and study designs. A high score suggested that an article had a good presentation of findings in tables and figures and that the description of analysis methods was helpful to readers. Interrater and intrarater agreements were high. Conclusion: An applicable checklist for quickly testing the statistical reporting quality of manuscripts and published research papers was developed. This instrument aims to improve the quality of empirical research in scientific fields where statistical methods play an important role.

Keywords: data analysis; statistics; reporting; data presentation; publications; medicine; health care

---

1. Introduction

Statistical reporting plays an important role in medical publications [1]. A high proportion of medical articles are essentially statistical in their presentation. For the reader, the main outward demonstration of a publication consists of statistical expressions that summarize the raw data used in the research. Regardless of the quality of the data and the variables chosen to express the results, the overt evidence of the research is produced as the lists of numbers, tables, plots, graphs, and other displays, i.e., the descriptive statistics. The communication of the descriptive statistics may also be combined with statistical inference procedures, such as test statistics and *p*-values [2].

The accurate communication of research results to the scientific community (i.e., other researchers) is essential providing reliable results [3,4]. Failure to report analysis methods or inadequate data presentation could lead to inaccurate inferences, even if the statistical analyses were performed correctly.

Several findings have demonstrated that a noteworthy percentage of articles, even those published in high-prestige journals, have flaws in their statistical reporting [3–8]. Methodological and reporting problems have been reported over several decades, and despite continued efforts to improve the situation [9], the improvements have been far from satisfactory. One explanation is the inadequate assessment of statistical reporting during peer review [8].

Author instructions and editorials alone have been insufficient to change the quality of statistical reporting. Numerous reporting guidelines offer tools or checklists to evaluate the reporting of health research [10]. The EQUATOR Network listed 424 such guidelines in March 2020 [11]. What is often problematic in these checklists is their extent. They try to evaluate all possible aspects of analysis from study design, settings, participants, description of measurements and variables, sample sizes, data analysis methods, specific methods, descriptive data, outcome data, and ancillary analysis. Often it is impossible to evaluate all these points. When reviewing and editing health care and medical papers, I have noticed that a few points will reveal the quality of reporting of data analysis. There is no need to read the whole manuscript and check all the statistical details to ensure that the manuscript contains all the information that readers need to assess the study's methodology and validity of its findings. Editorial processes of medical manuscripts often fail to identify obvious indicators of good and poor data presentation and reporting of results. Editors, reviewers and readers need new tools to quickly identify studies with inadequate descriptions of statistical methods and poor quality of data presentation.

I took on the challenge of compacting and refining the available quality criteria for statistical reporting and data presentation. I had three guiding starting points for assessing statistical reporting. The first key point was that if the data analysis methods are described with enough details, readers will be able to follow the flow of the analyses and to critically assess whether the data analyses provide reliable results. My second key point was the use of tables and figures in data presentation. Tables and figures are a fundamental means of scientific communications and they are included in most submitted and published medical and health care articles [12]. Most readers browse through tables and illustrations, but few read the whole article. Well-prepared tables and figures with proper titles, clear labeling, and optimally presented data will empower readers to scrutinize the data. I presume that the quality of tables and figures indicates the overall quality of data presentation. My third starting key point was that the new tool should be easy to use. I wanted to develop a short evaluation form with a low number of items.

In short, my objective was to develop a handy and reliable instrument to assess the quality of statistical reporting and data presentation that was applicable for a wide variety of medical and health care research forums, including both clinical and basic sciences. In this article, I define nine items for the proposed quality test. In addition, I report findings from an initial evaluation of its functionality and reliability.

2. Materials and Methods

2.1. Prework

My intention was to develop a test or checklist that would complement and not replicate other checklists. The construction of my quality evaluation tool began with a literature search. It included the following steps:

- Search for papers evaluating the use and misuse of statistics in medical articles.
- Search for papers evaluating statistical reporting and data presentation.
- Review medical statistics textbooks.
- Review statistical reporting guidelines, including journal guidelines for authors.
- Assess my experience as a handling editor and referee for medical and medicine-related journals.

After reading and reviewing the existing body of evidence, I discovered that including an assessment of statistical errors into a quality instrument would be very difficult. Several studies have reviewed medical papers and tried to find errors in the selection of statistical procedures [5,8,13]. Most of the statistical problems in medical journals reported in these reviews are related to elementary statistical techniques [14]. The errors in medical papers are probably more a matter of judgment. There is no general agreement on what constitutes a statistical error [6,15,16]. The recommended methods for a particular type of question may not be the only feasible methods, and may not be universally agreed upon as the best methods [17]. In addition, I decided to avoid any judgment of the content, such as the originality, ethics, or scientific relevance.

I found hundreds of journal guidelines for statistical reporting. I observed strong heterogeneity among the guidelines. Some journals have kept these guidelines to the minimum required for a decent manuscript presentation, whereas a large number of journals displayed too many instructions and journal-specific recommendations. Some recommendations regarding the test statistics (reporting of values, degrees of freedom and p-values) or the contents of tables were even contradictory (between journals). Most reporting guidelines said more about the presentation of methods than about the reporting of results. In addition, medical statistics textbooks did not cover the topic of statistical reporting.

By the end of the prework process, I had identified and formulated 20 items phrased as questions and grouped them into two domains (description of methods and data presentation). The draft questions were as follows:

Statistical analysis (or Data analysis) subsection in the Material and Methods section:

- Was Statistical analysis subsection included?
- Was there a sample size justification before the study?
- Did authors state how variables were described and summarized?
- Did authors state which methods were used to evaluate the statistical significances?
- Did authors identify the variables for each analysis and mentioned all the statistical methods used?
- Was extended description of some specific procedures provided?
- Did authors verify how data conformed to assumptions of the methods used?
- Was statistical software reported?
- Was missing data addressed?
- Were references to statistical literature provided?
- Was subject attrition or exclusion addressed?
- How any outlying data were treated in the analysis?

Data presentation:

- Was a table included where the basic characteristics of the study subjects were summarized with descriptive statistics?
- Were total and group sample sizes reported for each analysis in all tables and figures?
- Did all tables and figures have a clear and self-explanatory title?
- Were statistical abbreviations explained in all tables and figures?
- Were summary statistics, tests or methods identified and named in all tables and figures?
- Were p-values reported properly (e.g., no expressions like NS, $p < 0.05$, $p = 0.000$)
- Was the total number of reported p-values and confidence intervals in tables and figures less than 100?
- Were all tables and figures appropriate?

In line with my original key principles, the evaluation form should include two sections: (1) Material and Methods section and (2) tables and figures in Results section. In addition, the test form should be only one page long. I developed several pilot versions to structure the assessment of the

quality of statistical reporting in an article or manuscript. Within these frames I reviewed the 20 draft questions for their importance and uniqueness with my health care students, who tested the instrument drafts during their studies. Further testing of the instrument, a review of the literature related to the topic and expert opinions from my colleagues resulted in the generation of 9 items pertaining to the description of statistical and data management procedures and the reporting of results in tables and figures. Following the pilot studies, several items were also reworded, rearranged, and consolidated in the test form for clarity. Because tables and figures should be self-explanatory, I decided to place the assessment of tables and figures at the beginning of the evaluation form. Thus, the assessment of the description of the statistical analysis in the Materials and methods section now follows the assessment of the tables and figures. The updated version of the instrument is included in as Table 1. In the evaluation form, questions 1–4 relate to basic guidelines for preparing effective tables and figures. Questions 5–9 evaluate the description of the data analysis methods.

Table 1. Statistical reporting and data presentation evaluation form.

TABLES AND FIGURES IN RESULTS SECTION:	No	Yes
1.　Was a table included where the basic characteristics of the study participants were summarized with descriptive statistics?	0	1
2.　Was the total number of participants provided in all tables and figures?	0	1
3.　Were summary statistics, tests and methods identified and named in all tables and figures?	0	1
4.　Were tables and figures well prepared?		
0 = more than 50% of the tables and figures had presentation issues 1 = 50% or fewer of the tables and figures had presentation issues 2 = no presentation issues in any of the tables and figures		0 1 2
MATERIALS AND METHODS SECTION:		
5.　Was the statistical analysis (or data analysis) subsection provided in the Materials and Methods section?	0	1
6.　Did authors identify the variables (and methods) for each analysis done in the study?	0	1
7.　Was it verified that the data conformed to the assumptions and preconditions of the methods used to analyze them?	0	1
8.　Were references to statistical literature provided?	0	1
9.　Was the statistical software reported?	0	1
TOTAL SCORE:		

2.2. Ten Items to Assess the Quality of Statistical Reporting and Data Presentation

In the following chapter, I present the items as a sequence of 9 numbered questions and explain some of the reasoning behind the questions.

2.2.1. Item 1: Was a Table Included Where the Basic Characteristics of the Study Subjects Were Summarized with Descriptive Statistics?

Most papers reporting an analysis of health care and medical data will at some point use statistics to describe the sociodemographic characteristics, medical history and main outcome variables of the study participants. An important motive for doing this is to give the reader some idea of the extent to which study findings can be generalized to their own local situation [18]. The production of descriptive statistics is a straightforward matter, but often authors need to decide which statistics to present. The selected statistics should be included in a paper in a manner that is easy for readers to assimilate. When many patient characteristics are being described, the details of the statistics used and the number of participants contributing to analysis are best incorporated in tabular presentation [12,18,19].

I give one point if the main features of important characteristics of the participants were displayed in a table.

2.2.2. Item 2: Was the Total Number of Participants Provided in all Tables and Figures?

The medical literature shows a strong tendency to accentuate significance testing, specifically "statistically significant" outcomes. Most papers published in medical journals contain tables and figures reporting *p*-values [16,20,21]. Finding statistically significant or nonsignificant results depends on the sample size [22,23]. When evaluating the validity of the findings, the reader must know the number of study participants. Sample size is an important consideration for research. A larger sample size leads to a higher level of precision and thus a higher level of power for a given study to detect an effect of a given size. An excessively large number of study participants may lead to statistically significant results even when there is no clinical practicality; an inappropriately small number of study participants can fail to reveal important and clinically significant differences. The total number of participants or sample size of each group should be clearly reported.

I give one point if most of the tables and figures (at least 75%) have provided the total number of participants or number of participants in each subgroup.

2.2.3. Item 3: Were Summary Statistics, Tests, and Methods Identified and Named in all Tables and Figures?

Tables and figures should be able to stand alone [24]. That is, all information necessary for interpretation should be included within the table or figure, legend or footnotes [25]. This means that descriptive statistics, significance tests, and multivariable modeling methods used are named. Many readers will skim an article before reading it closely and identifying data analysis methods in the tables or figure will allow the readers to understand the procedures immediately. It is not necessary to define well known statistical abbreviations, such as SD, SE, OR, RR, HR, CI, r, R2, n, or NA, but the methods producing these statistics should be named.

I give one point if statistical summary statistics, tests, and methods in tables and figures were named.

2.2.4. Item 4: Were Tables and Figures Well Prepared?

High quality tables and figures increase the chance that readers can take advantage of an article's results. In effective scientific writing, it is essential to ensure that the tables and figures are flawless, informative, and attractive. Below is a list of a number of presentation issues that may prevent readers grasping the message quickly and will lower the overall quality of data presentation [12,24,26].

- Messy, inferior, or substandard overall technical presentation of data.

  ○ A table or a figure did not have a clear title.
  ○ The formatting of a table resembled a spreadsheet, and the lines of the same size between each row and each column did not help to clarify the different data presented in the table.
  ○ In a figure, data values were not clearly visible.
  ○ Data values were not defined.
  ○ Obvious errors in presented numbers or data elements.

- Tables or figures included unnecessary features:

  ○ In a figure, nondata elements (gridlines, shading, or three dimensional perspectives) competed with data elements and they did not serve a specific explanatory function in the graph.
  ○ A table or a figure was unnecessary because the data had too few values. Authors could have presented their results clearly in a sentence or two. For example, a sentence is preferred to a pie char.

- General guiding principles for reporting statistical results were not followed:

○   *p*-Values were denoted with asterisks or with a system of letters in tables or figures, and actual *p*-values were not reported. Actual *p*-values should be reported, without false precision, whenever feasible. Providing the actual *p*-values prevents problems of interpretation related to *p*-values close to 0.05 [12,22]. Very small *p*-values do not need exact representation and $p < 0.001$ is usually sufficient.

○   Numbers were not reported with an appropriate degree of precision in tables. In interpreting the findings, the reader cannot pay attention to the numbers presented with several decimals.

○   The standard error of the mean (SE) was used to indicate the variability of a data set.

○   Confidence intervals were not reported with the effect sizes (regression coefficients, ORs, HRs, or IRRs) in regression analyses or meta-analyses. The results of the primary comparisons should always be reported with confidence intervals [27].

○   A table included only *p*-values. *p*-value cannot tell readers the strength or size of an effect, change, or relationship. In the end, patients and physicians want to know the magnitude of the benefit, change or association, not the statistical significance of individual studies [23,28].

I give zero points if more than 50% of the tables and figures included presentation issues.

I give one point if 50% or fewer of the tables and figures included presentation issues.

I give two points if all tables and figures were prepared efficiently and accurately, and the presentation issues mentioned above were avoided.

2.2.5. Item 5: Was a Statistical Analysis (or Data Analysis) Subsection Provided in the Methods Section?

Most general reporting guidelines and recommendations require that original research articles include a Methods section [10,24,29]. In these recommendations, it is stated that the Methods section should aim to be sufficiently detailed such that others with access to the data would be able to reproduce the results. This section should include at least the following subsections: Selection and description of participants, technical information about variables (primary and secondary outcomes, explanatory variables, other variables) and statistical methods. To many researchers, it seems quite obvious that a Statistical analysis (or Data analysis) subsection with a clear subheading should be provided in the Materials and Methods section when the manuscript contains some elements of statistical data analysis. However, in my experience, this is not obvious to all biomedical or health science researchers. I have reviewed several biomedical manuscripts for journals wherein the laboratory experiments were described in depth, but nothing was said about how the reported *p*-values were obtained. When statistical methods are described with enough detail in the Statistical analysis subsection, a knowledgeable reader can judge the appropriateness of the methods for the study and verify the reported methods.

I therefore give one point if the Methods section included a subsection headed with Statistical analysis (or Data analysis or Statistical methods).

2.2.6. Item 6: Did Authors Identify the Variables and Methods for Each Analysis?

Authors know their own work so well that some find it difficult to put themselves in the position of a reader encountering the study for the first time. As a reader, I often request more information about some aspect of the data analysis. Within the Statistical analysis section, authors need to explain which statistical methods or tests were used for each analysis, rather than just listing all the statistical methods used in one place [30]. In a well-written statistical analysis subsection, authors should also identify the variables used in each analysis. Readers should be informed which variables were analyzed with each method. Care must be taken to ensure that all methods are listed and that all tests listed are indeed applied in the study [31]. The statistical section should be consistent with the Results section.

I give one point if authors describe the goal (research question), the variables and the method used for each analysis done in the study.

2.2.7. Item 7: Was It Verified that the Data Conformed to the Assumptions and Preconditions of the Methods Used to Analyze Them?

All basic data analysis methods and multivariable techniques depend on assumptions about the characteristics of the data [32]. If an analysis is performed without satisfying the assumptions, incorrect conclusions may be made on the basis of erroneous results. For example, a widely applied analysis of variance depends at least on three assumptions [33]. In actual data analysis it is unlikely that all the assumptions for the analysis of variance will be satisfied. Some statistical assumptions are essential, while some assumptions are quite lenient. A normal distribution of main variables is a strong requirement in a number of statistical techniques and should be verified and reported. On the other hand, the use of a nonparametric significance test instead of a more powerful parametric test should be justified. If a brief justification is provided, readers may better understand why a specific data analysis method has been applied [14]. Regression modeling has several constraints or preconditions, such as linearity, independence between explanatory variables, and number of participants per variable. In an excellent description of multivariable methods, authors should describe possible limitations or preconditions.

I give one point if authors describe how the data satisfactorily fulfill the underlying assumptions and preconditions of the main analysis methods.

2.2.8. Item 8: Were References to Statistical Literature Provided?

The use of all the statistical methods in a study needs to be documented with a relevant description to help the readers to validate the findings described by the authors. References enable others to identify and trace the methods used in the data analyses. Common statistical methods can be described in brief, but some less common or obscure methods should be explained in detail. The International Committee of Medical Journal Editors (ICMJE) also instructs to give references to established methods [29]. Good scientific writing is includes references and brief descriptions for methods that have been published but are not well known or not commonly used, descriptions of new or substantially modified methods, reasons for using uncommon methods, and an evaluation of the method's limitations [34].

I give one point for providing statistical references.

2.2.9. Item 9: Was the Statistical Software Used in the Analysis Reported?

Identifying the statistical software package used in the data analysis is important because all statistical programs do not use the same algorithms or default options to compute the same statistics. As a result, the findings may vary from package to package or from algorithm to algorithm. In addition, privately developed algorithms may not be validated and updated [12].

I give one point for reporting statistical software if the name and version of the statistical program or package (name, version) is provided.

2.3. Total Score

Users of this tools can calculate a total score by summing all 9 items. The total score ranges from 0 to 10. I have assigned the following labels to the corresponding ranges of the total score:

9–10 Excellent
7–8 Good
5–6 Acceptable
3–4 Weak
0–2 Poor

Although this division is plain and unsophisticated, it does provide useful interpretation of the scores.

2.4. Set of Articles

I used original research articles published between 2017 and 2019 in 4 journals to test the instrument for published studies. I selected two highly visible medical journals (*Lancet* and *JAMA Psychiatry)*, one dental journal (*Journal of Dentistry) (JD*), and one journal from environmental and public health subfields (*International Journal of Environmental Research and Public Health) (IJERPH*) for the evaluation. I chose these journals to cover the range of statistical reporting both in established journals, and in lower visibility journals. I analyzed 40 articles per journal. The starting articles for each journal were chosen randomly from the journal's chronological list of articles, with the only criteria being that there would be at least 39 eligible subsequent articles published that year in the journal in question. The following consecutive 39 articles were also included for the review. Editorials, letters, case reports and review articles were excluded from the evaluation.

2.5. Data Analysis

Cross-tabulation was used to report the differences in the study design and sample size of the evaluated articles by the publication journals. For the total score of statistical reporting and data presentation, the mean value with standard deviation and box plots were used to describe the distribution by study design, sample size, and journal. Analysis of variance (ANOVA) was applied to evaluate the statistical significance of possible differences in the mean values of the total score. The quality score was approximately normally distributed, and the normality assumption of the analysis of variance test was met by the data. The chi-squared test was applied to reveal the statistically significant differences in the distributions of the items of the statistical reporting and data presentation instrument across journals. All of the statistical analyses were executed using IBM SPSS Statistics (version 25) software.

When using quality evaluation forms, several methodological aspects can impact accuracy and reliability. For instance, to obtain an appropriate description of statistical methods, the basic statistical methods must be familiar. Additionally, the level of expertise reviewers have about scientific writing can affect their ability to accurately detect shortcomings in reporting. Therefore, it was important to determine the interrater and test-retest reliability of the proposed tool using raters with low (Rater 1) and moderate (Rater 2) medical statistics experience. I recruited a junior medical researcher to serve as an independent rater (Rater 1). I acted as Rater 2 myself. Rater 1 had minimal experience with medical data analysis but had been involved with medical writing for the previous 2 years. Rater 1 received training in the use of the assessment form, and general guidelines were given on the items of the instrument.

The reliability study started in parallel with a pilot study of the evaluation form. The reliability of the evaluation was checked by comparing Rater 1 ratings with the ratings of Rater 2. We independently read and evaluated 40 articles published in *JAMA Psychiatry*. These articles were selected from the previously described set of 160 articles.

First, agreement between the summary scores was assessed using an intraclass correlation coefficient ICC (with agreement definition, single measures, and mixed model) and Pearson correlation coefficient [35]. Generally, good agreement is defined as an ICC > 0.80. To evaluate the test-retest (or intrarater) performance, I read the 40 articles published in *JAMA Psychiatry* twice. The time interval between the scoring sessions was six months.

Second, the percentage agreement and Cohen's kappa coefficient were used to assess the degree of agreement for each item [35]. The simple percentage agreement is an adequate measure of agreement for many purposes, but it does not account for agreement arising from chance alone. Categorical agreement is often measured with the kappa coefficient, which attempts to account for the agreement that may arise from chance alone. A kappa score in the range of 0.81 to 1 was considered to represent high agreement.

3. Results

3.1. Characteristics of the Evaluated Articles

The articles in the validation sample came from a variety of specialties: general medicine, clinical topics, psychiatry, dentistry, environmental topics, public health, and dentistry. Table 2 shows the basic characteristics of the article set. Observational studies (cross-sectional surveys, longitudinal and case-control studies) were performed at a higher frequency in *JAMA Psychiatry* and *IJERPH*. The proportion of experimental studies (randomized clinical trials, non-randomized intervention studies) was highest in *Lancet* (67.5%). The test article set also included laboratory works that applied data analysis methods scantily and had low intensity of statistical methods. Sample sizes were larger in the more prominent journals.

Table 2. Distribution of study design and sample size of the evaluated articles by the publication journal.

	Lancet n (%)	*JAMA Psy* n (%)	*IJERPH* n (%)	*JD* n (%)	All n (%)
Study design					
Observational studies	11 (27.5)	27 (67.5)	27 (67.5)	13 (32.5)	78 (48.8)
Experimental studies	27 (67.5)	9 (22.5)	6 (15.0)	6 (15.0)	48 (30.0)
Reliability	1 (2.5)	0	2 (5.0)	2 (5.0)	5 (3.1)
Laboratory works	0	0	4 (10.0)	14 (35.0)	18 (11.3)
Meta-analysis	1 (2.5)	4 (10.0)	1 (2.5)	5 (12.5)	11 (6.9)
Sample size					
<99	0	3 (7.5)	14 (35.0)	20 (50.0)	37 (23.1)
100–499	5 (12.5)	12 (30.0)	12 (30.0)	8 (20.0)	37 (23.1)
500–2999	22 (55.0)	7 (17.5)	6 (15.0)	5 (12.5)	40 (25.0)
>3000	12 (30.0)	18 (45.0)	5 (12.5)	3 (7.5)	38 (23.8)
Missing	1 (2.5)	0	3 (7.5)	4 (10.0)	8 (5.0)
Total number of articles	40	40	40	40	160

JAMA Psy = JAMA Psychiatry, IJERPH = International Journal of Environmental Research and Public Health, and JD = Journal of Dentistry.

3.2. Distribution of the Total Quality Score

The total score assessing quality of statistical reporting ranges from 0 to 10. Figure 1 shows the distribution of the quality score among all 160 articles. A total of 14 (8.8%) articles were poor and 42 (26.3%) articles did not reach up to an acceptable level.
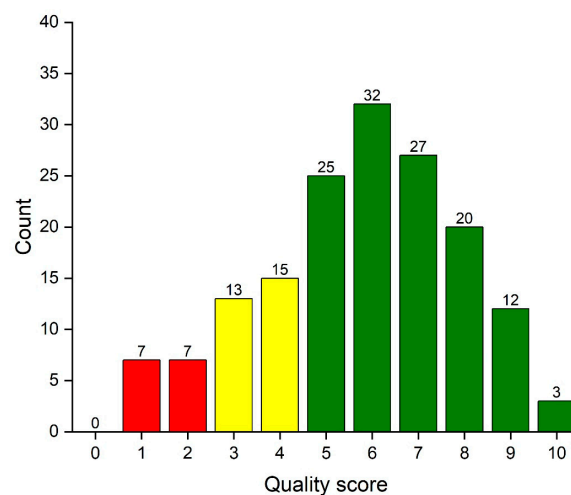
Figure 1. Column graph of the data analysis reporting quality score of 160 articles. Articles with poor quality (0–2) are denoted with red column color, those with weak quality (3–4) with yellow color and at least acceptable (5–10) articles with green color.

The distribution of the data analysis reporting score is summarized in Table 3 by article characteristics. The mean (SD) of quality score in all 160 evaluated reports was 5.7 (SD 2.2). The reporting quality was highest in the meta-analyses and lowest in laboratory works. The data analysis reporting quality was associated with the sample size of the evaluated study.

Table 3. Mean and standard deviation of the data analysis reporting score from 160 articles by study design and sample size.

	Number of Articles	Mean	Standard Deviation
Study design			
• Observational studies	78	5.5	2.2
• Experimental studies	48	6.5	1.8
• Reliability	5	5.4	2.6
• Laboratory works	18	3.9	1.9
• Meta-analysis	11	7.4	2.2
Sample size			
• <99	37	4.7	2.2
• 100–499	37	5.6	2.0
• 500–2999	40	7.2	1.5
• >3000	38	6.0	1.9
• Missing	8	2.9	1.0
All articles	160	5.7	2.2

p-Value of ANOVA < 0.001 for study design and sample size.

3.3. Quality of Statistical Reporting and Data Presentation by the Journal

Table 4 summarizes the distributions of the data analysis reporting test items in tables and figures by the journals. Neglecting to state the number of participants was the most common defect (65.6%). The quality of tables and figures was related to the journal; the failures to provide data and report findings was less common in the *Lancet*.

Table 4. The distributions of the data analysis reporting test items by the journals.

Item	Lancet n (%)	JAMA Psy n (%)	IJERPH n (%)	JD n (%)	All n (%)	*p*-Value of the Chi-Squared Test
Tables and figures in results section:						
Basic characteristics reported in a table	35 (87.5)	27 (77.5)	27 (67.5)	12 (30.0)	101 (63.1)	<0.001
Total number of participants provided	27 (67.5)	9 (22.5)	12 (30.0)	7 (17.5)	55 (34.4)	<0.001
Statistics, tests and methods identified	34 (85.0)	17 (42.5)	20 (50.0)	20 (50.0)	91 (56.9)	<0.001
Presentation issues: • several • 50% or less • no issues	2 (5.0) 16 (40.0) 22 (55.0)	6 (15.0) 14 (35.0) 20 (50.0)	10 (25.0) 18 (45.0) 12 (30.0)	10 (25.0) 20 (50.0) 10 (25.0)	28 (17.5) 68 (42.5) 64 (40.0)	0.029
Materials and methods section						
Statistical analysis subsection provided	37 (92.5)	37 (92.5)	30 (75.0)	28 (70.0)	132 (82.5)	0.011
Variables identified	23 (57.5)	35 (87.5)	22 (55.0)	22 (55.0)	102 (63.7)	0.005
Assumptions verified	14 (35.0)	14 (35.0)	11 (27.5)	15 (37.5)	54 (33.8)	0.858
Software reported	32 (80.0)	31 (77.5)	29 (72.5)	31 (77.5)	123 (76.9)	0.911
References to statistical literature	14 (35.0)	25 (62.5)	14 (35.0)	11 (27.5)	64 (40.0)	0.008
Software reported	32 (80.0)	31 (77.5)	29 (72.5)	31 (77.5)	123 (76.9)	0.911
Total number of articles	40	40	40	40	160	

JAMA Psy = JAMA Psychiatry, IJERPH = International Journal of Environmental Research and Public Health and JD = Journal of Dentistry.

Table 4 also compares the prevalence of adequate description of methods in the four journals. The failure to verify that the data conformed to the assumptions and preconditions of the methods used was very common (66.2%). A total of 28 articles (17.5%) did not provide a statistical analysis subsection; 60.0% (96 articles) did not provide references to statistical literature; and 23.1% (37 articles) did not report the statistical software.

Figure 2 shows how the quality score varies by journals. The high-impact general medical journal *Lancet* had mean score of 6.9 (SD 1.7) while the mean score in the leading psychiatric journal *JAMA Psy* was 6.1 (SD 1.4). The quality core also identified the journals (*IJERPH* and *JD*) that publish more laboratory studies with smaller sample sizes. Articles published in these journals had lower scores, with a mean score of 5.2 (SD 2.3) in *IJERPH* and a mean score of 5.0 (SD 2.3) in *JD*. The mean values were statistically significantly different (*p*-value of ANOVA < 0.001).
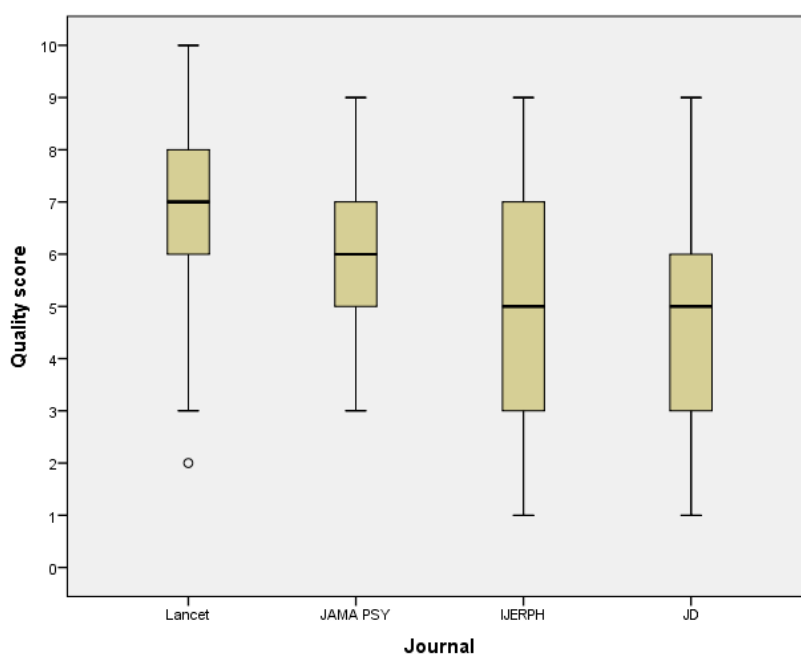
Figure 2. Box plot graph of the data analysis reporting quality score of 160 articles published in four journals. There were 40 articles from each journal. *JAMA Psy = JAMA Psychiatry*, *IJERPH = International Journal of Environmental Research and Public Health,* and *JD = Journal of Dentistry*.

3.4. Interrater and Test-Retest Reliability

The interobserver agreement measured by the ICC and Pearson correlation coefficient were 0.88 and 0.89, respectively. The test-retest reliability was excellent with identical mean scores for the first and second evaluations (Pearson's correlation coefficient was 0.95, and the intraclass correlation coefficient was 0.95).

The interrater reliability between Rater 1 and Rater 2 was also analyzed for each item. The overall percentage agreement ranged from 80% to 100% and kappa values ranged from 0.53 to 1. Individual item analyses showed very high agreement (100%) for items 5 (providing a statistical analysis subsection) and 9 (reporting software). The most common disagreement among Raters 1 and 2 was in the description of statistical methods. For Item 6 ("Did authors identify the variables with the methods for each analysis done in the study?"), the agreement was 92.5% and kappa was 0.53. For item 3 ("Were summary statistics, tests and methods identified and named in all tables and figures?") these values were 80% and 0.60, respectively.

The test-retest evaluation showed good agreement for each item: Agreement percentage ranged from 90% to 100% and the kappa coefficients ranged from 0.72 to 1. For items 1, 2, 5, 8, and 9,

the reliability measures provided complete agreement. This preliminary intrarater validation analysis revealed that disagreement arose especially for item 6.

4. Discussion

My purpose was to help authors, editors, reviewers, and readers to evaluate the quality of statistical reporting and data presentation in research papers when data analysis procedures are invoked to clarify findings and to draw conclusions from raw data. I developed a test for quickly assessing the quality by selecting the checklist items from readers' and reviewers' perspectives. The composed scale from 0 to 10 assessed the selected reporting characteristics in the reviewed articles. A high value suggested that an article had a good presentation of findings in tables and figures and that the description of analysis methods was helpful to readers. This indicates an emphasis on analysis and reporting in the study and indicates that the methodological quality is high and understandable. The items included in the index also gave detailed information to reviewers and authors about the defects with respect to specific parts of the reporting of data analysis findings.

The problems in the reproducibility of biomedical research have received considerable attention in recent years [8,36–38]. One explanation is the poor presentation of statistical findings and inadequate reporting of data analysis techniques by authors. Low quality has been a long-term issue in medical research papers [4,8,36]. There seems to be a cultural component to these practices [39]. The reporting practices of senior scientists are transmitted to junior scientists. Authors will often copy inadequate reporting practices from previously published papers in their field and ignore the recommendations of current general guidelines. Another contributing factor is the disregard for medical biostatistics in biomedical and clinical education [36,40–43]. Many researchers clearly recognize the importance of data analysis concepts, but they continue to misunderstand elementary data analysis techniques and statistical concepts [44].

To clinicians and scientists, the literature is an important means for acquiring new information to guide health care research and clinical decision making. Poor reporting is unethical and may have serious consequences for clinical practice, future research, policy making, patient care, and ultimately for patients [10]. Securing the quality of the publications is an important activity of journals in their editorial policy. Journals need to be more proactive in providing information about the quality of what journals publish [10,45]. Journals should consider strategies and actions to verify that authors realize full responsibility for the statistical reporting quality of their manuscripts. Use of the short checklist proposed in this paper with reporting guidelines would be an important step towards quality-secured research.

Peer review provides the foundation for the scientific publishing system. During the editorial process peer reviewers are required to comment on whether the methods and findings are clearly reported [46]. This approach heavily relies on the statistical expertise of subject reviewers. In general, peer reviewers are competent in a specific range of statistical methods, but they may not necessarily be aware of more general statistical issues and best practices [16,36,47]. It has been claimed that due to the inadequate assessment of statistical reporting and data presentation during peer review and editorial processes, the quality of biomedical articles has not improved [8,36,47–49].

Medical journals often ask their subject reviewers if they are able to assess all statistical aspects of the manuscript themselves or whether they recommend an additional statistical review [47]. Leading medical journals, such as *Lancet*, *BMJ*, *Annals of Medicine* and *JAMA* have adopted statistical review. Despite demonstration of widespread statistical and data presentation errors in medical articles, increasing the use of statistical reviewers has been slow [50]. A recent survey found that only 23% of the top biomedical journals reported that they routinely employed statistical review for all original research articles [51]. Introduction of specialist statisticians to the peer review process has made peer review more specialized. In addition, statistical reviewing is time intensive, limited by both reviewer supply and expense.

In biomedical journals, there is no single model for statistical review in peer review strategies [51–53]. Some journals recruit statistical methodologists to the editorial board, some draw their statistical reviewers from an external pool. If all papers cannot be statistically reviewed, editors have to select which manuscripts should undergo statistical scrutiny. There are also models where subject reviewers are assisted to comment on the statistical aspects of a manuscript [47,48,54]. However, these checklists cover extensively all aspects of data analysis and are not straightforward for non-statistical reviewers to get an overall impression of the statistical quality. Some form of simplified checklist could be handy for editors and reviewers to spot the issues that might indicate more serious problems in the reporting of scientific articles. I hope that editors and reviewers could use the short quality test proposed in this paper for deciding when the presentation in a manuscript is clearly inadequate and they should recommend rejecting the manuscript. I agree with Kyrgidis and Triaridis [48] that if the reviewer cannot find the basic information and description related to the data analysis, the reviewer does not need to read the whole article. After checking tables and figures and reading through the statistical analysis subsection in the methods section, the reviewer can reject the manuscript on good grounds. When the proposed simple quality test shows that the statistical reporting and data presentation are appropriate, the whole article needs to be read and further reviewed.

In recent years, several journals have tried to improve peer review processes [55]. Their efforts have been focused on introducing openness and transparency to the models of peer review [56]. New strategies in peer review might help to address persistent statistical reporting and data presentation issues in the medical literature [55]. Software algorithms and scanners have been developed to assess internal consistency and validity of statistical tests in academic writing [57]. However, their use is still rare and limited to flag specific potential errors. The open peer review, where all peer reviews are made openly available brings into use new models where the quality of a paper may be assessed by the whole scientific community. The pre- and post-publication peer review models include commenting systems for the readership. Readers could use tools, such as proposed in this paper, to give feedback to authors. Subsequently, authors prepare a second version of the manuscript reflecting the comments and suggestions proposed by the scientific community.

My validation set of articles included four journals with different levels of prestige and visibility. The reporting of data analysis information was more detailed and useful for the reader in the more visible journals (*Lancet* and *JAMA Psychiatry*). This is in line with previous studies [5,6]. High-impact journals have more rigorous review process, including extensive statistical reviews. Several researchers in scientific communication have recommended sending out all manuscripts with numerical data for statistical review [8,58,59]. In addition, the strict monitoring of revisions made in manuscripts is important.

The proposed quality test is short. It includes only nine items. These items measure the data presentation quality and the description of data analysis methods. It does not include items about study design, justification of sample size, potential sources of bias, common statistical methodological issues, or interpretation of results. Several reporting guidelines and checklists are available to ensure that authors have paid attention to these other aspects [10,11]. However, these checklists have had little impact on the quality of statistical reporting, and they have not been sufficient for ensuring that issues related to poor statistical reporting and inadequate data presentation are eliminated from manuscripts [3,8,60]. In addition, the general checklists do not include detailed recommendations for reporting how statistical analyses were performed and how to present data.

The article set for pilot testing the proposed instrument included only published studies. Editors and reviewers can use the proposed tool to judge whether manuscripts should be accepted for publication. Educators can utilize it when educating researchers about how to improve statistical reporting and data presentation. A future study is required to apply the instrument at the initial submission of manuscripts to scientific journals and to test how applicable it is in the peer review process. In addition, content and construct validity properties [61] need to be evaluated in future studies.

It should be noted that the instrument was specifically developed for health care, biomedical and clinical studies, and limitations may arise if it is used in nonmedical fields that apply statistical methods.

5. Conclusions

In summary, I have developed an applicable checklist for quickly testing the statistical reporting quality of manuscripts and published research papers. One aim of this work was to help authors prepare manuscripts. Good research deserves to be presented well, and good presentation is as much a part of the research as the collection and analysis of the data. With this instrument authors could test the effectiveness of their presentations (tables and figures) for their readers. In addition, they could check that they have not described statistical methods too briefly or superficially.

Competent editors and scientific reviewers should be able to identify errors in basic statistical analysis and reporting. However, there is accumulative evidence that inadequate reporting of basic statistics is a persistent and ubiquitous problem in published medical studies. Editors and reviewers can improve the quality of reporting by identifying submitted manuscripts with poor-quality statistical reporting and data presentation. I hope that the editors and reviewers of scientific journals will incorporate the proposed test instrument in their editorial process. For a manuscript to be considered for full review, its statistical presentation has to be good enough for readers, subject reviewers, and statistical reviewers to understand what statistical methods have been applied and what data are presented in tables and figures. Manuscripts with poor-quality statistical reporting (total score ranges from 0 to 2) should be rejected. Giving detailed comments in review reports to improve the reporting may not be straightforward in those cases.

Readers can use the proposed data analysis reporting test as an initial indicator of research quality. They can easily check whether, based on numerical data analyses, a published research article is readable, understandable and accessible to healthcare professionals who might wish to use similar techniques in their future work or to those who are not experts in a particular subfield. A high score indicates that the authors have focused on reporting and readers should be able to express an opinion about the results after reading the main text.

Further refinement of the data analysis reporting instrument will continue, and I invite feedback.

Author Contributions: P.N. has undertaken all the works of this paper.

Funding: This research received no external funding.

Acknowledgments: I am grateful to Jasleen Kaur who provided useful comments on the draft of the instrument items. I wish to thank all health since students at the University of Oulu who participated in the piloting of the instrument.

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

References

1. Sato, Y.; Gosho, M.; Nagashima, K.; Takahashi, S.; Ware, J.H.; Laird, N.M. Statistical Methods in the Journal—An Update. *N. Engl. J. Med.* **2017**, *376*, 1086–1087. [CrossRef] [PubMed]
2. Nieminen, P.; Miettunen, J.; Koponen, H.; Isohanni, M. Statistical methodologies in psychopharmacology: A review. *Hum. Psychopharmacol. Clin. Exp.* **2006**, *21*, 195–203. [CrossRef] [PubMed]
3. Norström, F. Poor quality in the reporting and use of statistical methods in public health—The case of unemployment and health. *Arch. Public Health* **2015**, *73*, 56. [CrossRef] [PubMed]
4. Diong, J.; Butler, A.A.; Gandevia, S.C.; Héroux, M.E. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. *PLoS ONE* **2018**, *13*, e202121. [CrossRef] [PubMed]
5. Altman, D.G. Poor-quality medical research: What can journals do? *JAMA* **2002**, *287*, 2765–2767. [CrossRef]
6. Nieminen, P.; Carpenter, J.; Rucker, G.; Schumacher, M. The relationship between quality of research and citation frequency. *BMC Med. Res. Methodol.* **2006**, *6*, 42. [CrossRef]
7. McClean, M.; Silverberg, J. Statistical reporting in randomized controlled trials from the dermatology literature: A review of 44 dermatology journals. *Br. J. Dermatol.* **2015**, *173*, 172–183. [CrossRef]

8. Dexter, F.; Shafer, S.L. Narrative Review of Statistical Reporting Checklists, Mandatory Statistical Editing, and Rectifying Common Problems in the Reporting of Scientific Articles. *Anesth. Analg.* **2017**, *124*, 943–947. [CrossRef]

9. Vandenbroucke, J.P.; Von Elm, E.; Altman, D.G.; Gøtzsche, P.C.; Mulrow, C.D.; Pocock, S.J.; Poole, C.; Schlesselman, J.J.; Egger, M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *PLoS Med.* **2007**, *4*, 1628–1654. [CrossRef]

10. Moher, D.; Altman, D.G.; Schultz, K.F.; Simera, I.; Wager, E. *Guidelines for Reporting Health Research. User's Manual*; John Wiley & Sons, Ltd.: Chichester, UK, 2014.

11. EQUATOR Network. Available online: https://www.equator-network.org/ (accessed on 30 March 2020).

12. Lang, T.A.; Michelle, S. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors and Reviewers*, 2nd ed.; American College of Physicians: Philadelphia, PA, USA, 2006.

13. Strasak, A.M.; Zaman, Q.; Pfeiffer, K.P.; Gobel, G.; Ulmer, H. Statistical errors in medical research—A review of common pitfalls. *Swiss Med. Wkly.* **2007**, *137*, 44–49.

14. Lee, S. Avoiding negative reviewer comments: Common statistical errors in anesthesia journals. *Korean J. Anesthesiol.* **2016**, *69*, 219–226. [CrossRef] [PubMed]

15. Altman, D.G. Statistical reviewing for medical journals. *Stat. Med.* **1998**, *17*, 2661–2674. [CrossRef]

16. Nieminen, P.; Virtanen, J.I.; Vähänikkilä, H. An instrument to assess the statistical intensity of medical research papers. *PLoS ONE* **2017**, *12*. [CrossRef] [PubMed]

17. Bland, M. *An Introduction to Medical Statistics*; Oxford University Press: Oxford, UK, 2015; ISBN 0199589925.

18. Pickering, R.M. Describing the participants in a study. *Age Ageing* **2017**, *46*, 576–581. [CrossRef] [PubMed]

19. 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]

20. Nieminen, P.; Kaur, J. Reporting of data analysis methods in psychiatric journals: Trends from 1996 to 2018. *Int. J. Methods Psychiatr. Res.* **2019**, *28*. [CrossRef]

21. Nieminen, P.; Vähänikkilä, H. Use of data analysis methods in dental publications: Is there evidence of a methodological change? *Publications* **2020**, *8*, 9. [CrossRef]

22. Thiese, M.S.; Ronna, B.; Ott, U. P value interpretations and considerations. *J. Thorac. Dis.* **2016**, *8*, E928–E931. [CrossRef]

23. Motulsky, H. *Intuitive Biostatistics*, 3rd ed.; Oxford University Press: New York, NY, USA, 2014; ISBN 9780199946648.

24. American Medical Association. *AMA Manual of Style: A Guide for Authors and Editors*, 11th ed.; Oxford University Press: New York, NY, USA, 2020; ISBN 978-0190246556.

25. Oberg, A.L.; Poland, G.A. The process of continuous journal improvement: New author guidelines for statistical and analytical reporting in VACCINE. *Vaccine* **2012**, *30*, 2915–2917. [CrossRef]

26. American Psychological Assiciation. *Publication Manual of the American Psychological Assiciation*, 6th ed.; American Psychological Assiciation: Washington, DC, USA, 2017; ISBN 978-1-4338-0561-5.

27. Gardner, M.J.; Altman, D.G. Confidence intervals rather than P values: Estimation rather than hypothesis testing. *Br. Med. J. (Clin. Res. Ed.)* **1986**, *292*, 746–750. [CrossRef]

28. Spector, R.; Vesell, E.S. Pharmacology and statistics: Recommendations to strengthen a productive partnership. *Pharmacology* **2006**, *78*, 113–122. [CrossRef] [PubMed]

29. ECMJE Recommendations for He Conduct, Reporting, Editing, and Publication of Scolarly Work in Medical Journals. Available online: http://www.icmje.org/recommendations/ (accessed on 25 May 2020).

30. Lang, T.A.; Altman, D.G. Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines. In *Guidelines for Reporting Health Research: A User's Manual*; John Wiley & Sons, Ltd.: Oxford, UK, 2014; pp. 264–274. ISBN 9780470670446.

31. Simundic, A.-M. Practical recommendations for statistical analysis and data presentation in Biochemia Medica journal. *Biochem. Med.* **2012**, *22*, 15–23. [CrossRef] [PubMed]

32. Mayo, D. *Statistical Inference as Severe Testing. How to Get beyond the Statistical Wars*; Cambridge University Press: Cambridge, UK, 2018; ISBN 978-1-107-66464-7.

33. Indrayan, A.; Malhotra, R.K. *Medical Biostatistics*, 4th ed.; CRC Press: Cambridge, UK, 2018; ISBN 1498799531.

34. Indrayan, A. Reporting of Basic Statistical Methods in Biomedical Journals: Improved SAMPL Guidelines. *Indian Pediatr.* **2020**, *57*, 43–48. [CrossRef] [PubMed]

35.	Gwet, K.L. *Handbook of Inter-Rater Reliability*, 4th ed.; Advances Analytics, LLS: Gaithersburg, MD, USA, 2014; ISBN 9780970806284.

36.	Glasziou, P.; Altman, D.G.; Bossuyt, P.; Boutron, I.; Clarke, M.; Julious, S.; Michie, S.; Moher, D.; Wager, E. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* **2014**, *383*, 267–276. [CrossRef]

37.	*The Academy of Medical Sciences Reproducibility and Reliability of Biomedical Research: Improving Research Practice*; The Academy of Medical Sciences: London, UK, 2015.

38.	Vetter, T.R.; McGwin, G.; Pittet, J.F. Replicability, Reproducibility, and Fragility of Research Findings-Ultimately, Caveat Emptor. *Anesth. Analg.* **2016**, *123*, 244–248. [CrossRef]

39.	Smaldino, P.E.; McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **2016**, *3*, 9. [CrossRef]

40.	Shetty, A.C.; Al Rasheed, N.M.; Albwardi, S.A. Dental professionals' attitude towards biostatistics. *J. Dent. Oral Hyg.* **2015**, *7*, 113–118.

41.	Hannigan, A.; Hegarty, A.C.; McGrath, D. Attitudes towards statistics of graduate entry medical students: The role of prior learning experiences. *BMC Med. Educ.* **2014**, *14*, 70. [CrossRef]

42.	Batra, M.; Gupta, M.; Dany, S.S.; Rajput, P. Perception of Dental Professionals towards Biostatistics. *Int. Sch. Res. Not.* **2014**, *2014*, 1–6. [CrossRef]

43.	Altman, D.G.; Goodman, S.N.; Schroter, S. How statistical expertise is used in medical research. *JAMA* **2002**, *287*, 2817–2820. [CrossRef]

44.	Belia, S.; Fidler, F.; Williams, J.; Cumming, G. Researchers misunderstand confidence intervals and standard error bars. *Psychol. Methods* **2005**, *10*, 389–396. [CrossRef]

45.	Moher, D. Reporting guidelines: Doing better for readers. *BMC Med.* **2018**, *16*, 18–20. [CrossRef]

46.	Glonti, K.; Cauchi, D.; Cobo, E.; Boutron, I.; Moher, D.; Hren, D. A scoping review on the roles and tasks of peer reviewers in the manuscript review process in biomedical journals. *BMC Med.* **2019**, *17*, 1–14. [CrossRef] [PubMed]

47.	Greenwood, D.C.; Freeman, J.V. How to spot a statistical problem: Advice for a non-statistical reviewer. *BMC Med.* **2015**, *13*, 1–3. [CrossRef] [PubMed]

48.	Kyrgidis, A.; Triaridis, S. Methods and Biostatistics: A concise guide for peer reviewers. *Hippokratia* **2010**, *14*, 13–22. [PubMed]

49.	Patel, J. Why training and specialization is needed for peer review: A case study of peer review for randomized controlled trials. *BMC Med.* **2014**, *12*, 1–7. [CrossRef] [PubMed]

50.	Horbach, S.P.J.M.; Halffman, W. The changing forms and expectations of peer review. *Res. Integr. Peer Rev.* **2018**, *3*, 1–15. [CrossRef]

51.	Hardwicke, T.E.; Goodman, S.N. How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey. *MetArXiv Prepr.* **2020**, 1–29.

52.	Hardwicke, T.E.; Frank, M.C.; Vazire, S.; Goodman, S.N. Should Psychology Journals Adopt Specialized Statistical Review? *Adv. Methods Pract. Psychol. Sci.* **2019**, *2*, 240–249. [CrossRef]

53.	Overgaard, K.; Roberts, J.; Schaeffer, M.B.; Pessin, R. the Trenches. *Sci. Ed.* **2015**, *38*, 2015.

54.	Curtis, M.J.; Alexander, S.; Cirino, G.; Docherty, J.R.; George, C.H.; Giembycz, M.A.; Hoyer, D.; Insel, P.A.; Izzo, A.A.; Ji, Y.; et al. Experimental design and analysis and their reporting II: Updated and simplified guidance for authors and peer reviewers. *Br. J. Pharmacol.* **2018**, *175*, 987–993. [CrossRef]

55.	Fresco-Santalla, A.N.A.; Hernández-Pérez, T. Current and evolving models of peer-review. *Ser. Libr.* **2014**, *67*, 373–398. [CrossRef]

56.	Vercellini, P.; Buggio, L.; Vigano, P.; Somigliana, E. Peer review in medical journals: Beyond quality of reports towards transparency and public scrutiny of the process. *Eur. J. Intern. Med.* **2016**, *31*, 15–19. [CrossRef]

57.	Nuijten, M.B.; Hartgerink, C.H.J.; van Assen, M.A.L.M.; Epskamp, S.; Wicherts, J.M. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* **2016**, *48*, 1205–1226. [CrossRef] [PubMed]

58.	Vahanikkila, H.; Tjaderhane, L.; Nieminen, P. The statistical reporting quality of articles published in 2010 in five dental journals. *Acta Odontol. Scand.* **2015**, *73*, 76–80. [CrossRef] [PubMed]

59.	Lukic, I.; Marusic, M. Appointment of statistical editor and quality of statistics in a small medical journal. *Croat. Med. J.* **2001**, *42*, 500–503. [PubMed]

60. Pouwels, K.B.; Widyakusuma, N.N.; Groenwold, R.H.H.; Hak, E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. *J. Clin. Epidemiol.* **2016**, *69*, 217–224. [CrossRef]

61. Mokkink, L.B.; Terwee, C.B.; Patrick, D.L.; Alonso, J.; Stratford, P.W.; Knol, D.L.; Bouter, L.M.; de Vet, H.C.W. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* **2010**, *63*, 737–745. [CrossRef]