

UC Davis

Dermatology Online Journal

Title

Frequency of statistical mistakes and associated article characteristics: a cross-sectional analysis of dermatology journals

Permalink

<https://escholarship.org/uc/item/9p75464j>

Journal

Dermatology Online Journal, 31(3)

Authors

Woodie, Brad R
Freking, Justin A
Jones, Grace M
et al.

Publication Date

2025

DOI

10.5070/D331365357

Copyright Information

Copyright 2025 by the author(s). This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

Frequency of statistical mistakes and associated article characteristics: a cross-sectional analysis of dermatology journals

Brad R Woodie, BS^{1a}, Justin A Freking, BS¹, Grace M Jones, BS¹, Justin Porter, MD², Sarah E Fleischer, MS³, Annabella G Pauley, MS⁴, Alan B Fleischer Jr, MD⁵

¹ College of Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio, United States, ² Department of Internal Medicine, The Christ Hospital, Cincinnati, Ohio, United States, ³ Department of Research, American Board of Family Medicine, Lexington, Kentucky, United States, ⁴ Marshall University School of Medicine, Huntington, West Virginia, United States, ⁵ Department of Dermatology, University of Cincinnati College of Medicine, Cincinnati, Ohio, United States

Keywords: biostatistics, errors, regression analysis, research design, statistics

Dermatology Online Journal

Vol. 31, Issue 3, 2025

Abstract

Statistical mistakes can undermine research credibility. Identifying common errors may help researchers avoid them in future studies. This study evaluated the frequency and types of statistical mistakes in dermatology journal articles and identified article characteristics that predict these errors. A cross-sectional analysis was conducted on articles published in the 2023 volumes of 8 dermatology journals. Articles were screened for statistical tests, with a target sample of 200 selected pseudorandomly. Multivariable logistic regressions assessed predictors of statistical mistakes, including journal impact factor, statistician involvement, funding source, first author highest degree, and statistical package. Of the 189 articles analyzed, 78% contained at least one statistical mistake. Reporting mistakes were found in 67% and test selection errors in 46%. The absence of statistician involvement (aOR 2.49, P=0.03) and low journal impact factor (aOR 3.82, P=0.02) predicted the presence of at least one mistake. This sample from 8 journals is not representative of all dermatology literature. Original data were not available for testing of test assumptions, so appropriate test selection was determined using statistical conventions. Statistical mistakes are prevalent in dermatology literature. Researchers should review statistical best practices and consider involving a statistician in their work.

Introduction

The validity and reliability of scientific research depend upon rigorous methodology and appropriate statistical analysis. Guidelines such as the Consolidated Standards of Reporting Trials and Strengthening the Reporting of Observational Studies in Epidemiology have been developed to improve the reliability of medical literature.^{1,2} Guides to avoiding common mistakes in various fields have also been published.³⁻⁷ Despite the availability of these resources and computerized statistics software, statistical mistakes persist in medical literature.

Nearly two decades ago, Neville et al (2006) found statistical mistakes in 38% of articles published in the Archives of Dermatology and the Journal of the American Academy of Dermatology.⁸ Recent investigations of dermatology and plastic surgery randomized controlled trials demonstrated that essential elements, such as the significance level, study power, and sample size calculations are often missing.⁹⁻¹¹ In other fields, studies have found statistical mistakes in 45% of cytopathology articles,¹² 94% of radiology articles,¹³ 98% of family medicine articles,¹⁴ and 75% to 93% of doctoral dissertations.^{15,16}

Identifying and preventing common statistical analysis mistakes can help researchers improve the clarity and impact of their work. This study evaluated the frequency and types of statistical mistakes in dermatology journal articles published in 2023. To inform efforts at reducing such errors, this study also identified characteristics of articles that were associated with the presence of statistical mistakes.

Methods

Article selection

Eight journals were chosen in a convenience sample to represent high- and low-impact factor dermatology publications: *Journal of the American Academy of Dermatology (JAAD)*, *British Journal of Dermatology*, *Journal of the American Medical Association (JAMA) Dermatology*, *Journal of the European Academy of Dermatology and Venereology*, *Australasian Journal of Dermatology (AJD)*, *Skin Appendage Disorders*, *Pediatric Dermatology*, and *Cutis*. The Web of Science electronic database was searched for these journals with publication type “article” (thus excluding letters to the editor) and publication year 2023, returning 807 results. Our further inclusion criterion was reporting at least one statistical test; 426 articles performing no statistical tests or solely reporting descriptive statistics were excluded. A minimum sample size of 184 was calculated a priori using a small effect size (Cohen's $d=0.2$), a two-sided alpha of .05, and a power of .90 for logistic regression. Eligible titles were pseudorandomized using Microsoft Excel's “randomize range” function. “Randomize range” generates pseudorandom numbers for each item, then reorders the items based on these numbers. The first 25 qualifying articles from the pseudorandomized list for each journal were selected for a target sample of 200 articles. All the qualifying articles were used in journals that published fewer than 25 qualifying articles.

Possible predictors of mistakes

The following information was obtained from the included articles: impact factor of the publishing journal, funding source, statistician involvement, first author's degree(s), and software used for statistical analysis. Using Clarivate 2023 Journal Impact Factors,¹⁷ journal impact factor was categorized as “high” for *JAAD* (12.8), *JAMA Dermatology* (11.5), *British Journal of Dermatology* (11.0), and *Journal of the European Academy of Dermatology and Venereology* (8.4). Impact factor was categorized as “low” for *AJD* (2.2), *Cutis* (2.1), *Pediatric Dermatology* (1.2), and *Skin Appendage Disorders* (1.4). The funding source was categorized into “industry,” “government,” “other, including nonprofit or community group,” and “none or not stated.” Statistician involvement was assessed by searching author affiliations for departments of statistics, bioinformatics, mathematics, or related fields. In cases of uncertainty, a Google search of the author's name and listed affiliation was performed. We also included statisticians or statistics groups mentioned in article acknowledgments. The first author's highest degrees were categorized into “doctor of medicine only (eg, MD, DO, MBBS),” “doctor of philosophy (PhD) or doctor of medicine with PhD (eg, MD/PhD),” or “master's or bachelor's degree.” Credentials outside of the listed degrees were placed into the best-matched category. Author degrees were obtained from a Google search of the author's name and affiliation if not provided by the journal.

Defining mistakes in statistical reporting

The first of two categories of statistical mistakes in this study comprised errors or omissions in reporting. These included: failure to state a significance level (eg, “significance was set at $P<0.05$ ”), failure to state if the significance level is one-sided or two-sided, failure to include a measure of variance for measures of central tendencies (eg, mean $\pm$ standard deviation), failure to report exact P-values, and incorrect interpretation of the P-value. Since χ^2 -family tests and Analysis of Variance family tests are inherently two-sided, we did not require that authors state one-sided or two-sided for these tests. Regarding P-value interpretation, previous studies have found that authors may misinterpret nonsignificant P-values using erroneous statements such as, “trending toward significance,” or “borderline significant.”¹⁸⁻²⁰

Defining mistakes in test selection

The other category of statistical mistakes comprised errors in test selection. When the article authors addressed assumptions of their statistical tests, the corresponding tests were considered correct. Few articles explicitly include this information, so we established criteria to determine appropriate use ([Table 1](#)).

Tests not belonging to a listed test family or shown in this table were evaluated on a case-by-case basis using any references cited by the article authors, performing a literature review, and consulting statistician author S.F.

For sample sizes, we utilized “rules of thumb” to evaluate the assumptions of parametric tests (eg, t-tests, Analysis of Variance, Pearson rank correlation). Larger samples allow more flexibility with violations of normality, so a minimum group size of 30 was chosen for these tests. The Central Limit Theorem supports this approach, as larger sample sizes lead to a normal distribution of sample means, even if the original data is not normal.^{21, 22} In unpaired or paired testing methods, the study design was examined to ensure that variables were independent or dependent, respectively.

The χ^2 test requires each observation to be independent and that the expected frequency in at least 80% of the contingency table cells is at least 5. If the latter condition is violated, the Fisher's exact test should be used, especially if any expected frequency is less than one.²³

For regression, the concept of “minimum” sample sizes is a nuanced and highly debated topic in statistics.²⁴⁻²⁶ In our study, logistic and Cox regressions were considered appropriate if the sample size was at least 100 and included 10 events per independent variable.²⁷⁻³² An “event” refers to the less frequent outcome rather than the total sample size.³² Linear regressions were considered appropriate with a sample size of at least 25 and at least 10 cases per independent variable.^{22,33-35}

Table 1. Criteria Used by Evaluators to Identify Test Selection Mistakes

Test or Test Family	Assumptions that were evaluated	Numerical criteria (if applicable)
χ^2 test	Approximation to the χ^2 distribution	Expected count ≥ 5 in $\geq 80\%$ of cells
	Observations are independent	
	Cells in the contingency table are mutually exclusive	
Fisher's exact test	Observations are independent	
	Cells in the contingency table are mutually exclusive	
Unpaired T-test	Approximate normality of observations	Group size ≥ 30 or article addresses normality ^a
	Groups are independent	
Mann-Whitney U-test/Wilcoxon Rank-Sum	Groups are independent	
Paired T-test	Approximate normality of observations	Group size ≥ 30 or article addresses normality
	Groups are paired	
Wilcoxon signed rank	Groups are paired	
Logistic regression	Approximate normality of errors and model stability	Group size ≥ 100 and ≥ 10 events per variable
Cox regression	Approximate normality of errors and model stability	Group size ≥ 100 and ≥ 10 events per variable
Linear regression	Approximate normality of errors and model stability	Group size ≥ 25 and ≥ 10 events per variable
	Relationship approximately linear	
Analysis of Variance (ANOVA)	Approximate normality of group residuals	Group size ≥ 30 or article addresses normality
	Groups are independent	
Kruskal-Wallis	Groups are independent	
Analysis of Covariance (ANCOVA)	Approximate normality of group residuals	Group size ≥ 30 or article addresses normality
	Groups are independent	
Pearson correlation	Approximate normality of observations	Group size ≥ 30 or article addresses normality
	Monotonic relationship (if plot available)	
Spearman correlation	Monotonic relationship (if plot available)	

^a If article authors addressed a normality assumption qualitatively or with a formal test, this criterion was considered correct.

Article evaluators

Authors BW, GJ, JF, and JP served as primary evaluators in the study. Before data collection, all evaluators completed the courses "Understanding Clinical Research: Behind the Statistics," (University of Cape Town) and "Regression Models" (Johns Hopkins University) available on Coursera (www.coursera.org). Primary evaluators and statistician S.F. discussed and agreed upon the criteria for appropriate statistical reporting and test selection at the initiation of the study. To determine inter-rater reliability, evaluators independently rated 20 articles, which were again pseudorandomly selected. Fleiss' kappa was

0.71, indicating moderate agreement. Each article included in this study was then reviewed by two independent evaluators. Any supplemental content was also reviewed for each article. The scoring of mistakes for each article was discussed between its two reviewers until agreement was reached. When uncertainty remained over the correct scoring, S.F. served as the adjudicator. Of note, multiple uses of the same test within the same article were only counted once. If we judged at least one of the test selection instances to be a mistake, we scored the article's use of that test as a mistake.

Statistical analysis

Descriptive statistics include frequencies and percentages of mistakes by category. Univariable logistic regressions were performed with the article characteristics (statistician involvement, journal impact factor, first author highest degree, funding source, and statistical package). There was no missing data.

With only 42 papers (22%) having no statistical mistakes, we included two binary predictors and one predictor with three nominal levels in our primary multivariable logistic regression to respect the 10 events-per-variable convention. Multivariable logistic regression was performed with predictors being statistician involvement (present or absent), journal impact factor (high or low), and funding source (no funding, industry funding, or other funding). Three regressions were performed with the outcome being any mistake, any reporting mistake, and any test selection mistake. Multicollinearity between predictors was not present as variance inflation factors were less than two, where a value greater than 5 suggests possible multicollinearity.³⁶ We applied the Benjamini-Hochberg correction to P-values to control the false discovery rate from performing multiple comparisons.

Since we did not include the first author's highest degree or statistical package in our primary model, we performed exploratory multivariable logistic regressions for these characteristics. Adjusting for impact factor and statistician involvement, we evaluated the association of the first author's highest degree and statistical package with the presence of mistakes. Misinterpretation of the P-value was also evaluated in an exploratory multivariable logistic regression with impact factor and statistician involvement as predictors. Statistical analysis was conducted in SAS v9.4 (Cary, NC). This study followed the Strengthening the Reporting of Observational Studies in Epidemiology reporting guidelines.

Results

A total of 189 qualifying journal articles published in 2023 were selected from the eight journals of interest. Cutis had only 21 qualifying articles and Skin Appendage Disorders had only 18 qualifying articles, whereas 25 were pseudorandomly selected from each of the other 6 journals. At least one mistake was found in 147 articles (77.8%), with 127 having at least one statistical reporting mistake (67.2%) and 87 having at least one test selection mistake (46.0%) ([Figure 1](#)). The most common statistical reporting mistake was failing to state if the significance level was one-sided or two-sided ([Table 2](#)). The most employed statistical test was the χ^2 test and the most misused test by percent incorrect was the paired t-test ([Figure 2](#)). Categories of test selection mistakes are shown in [Table 3](#).

Univariable logistic regressions showed that the absence of statistician involvement and low journal impact factor was associated with any statistical mistake, any re-

Table 2. Frequencies of Mistakes in Reporting Statistical Information

Statistical reporting mistake	Count (%)
Did not state significance level (alpha)	35 (18.5)
Did not state if the significance level is one- or two-sided	102 (54.0)
Did not provide a measure of spread for central tendency	9 (4.8)
Did not provide exact P-values	38 (20.1)
Misinterpreted P-value	28 (14.8)
Any statistical reporting mistake	127 (67.2)

porting mistake, and any test selection mistake ([Figure 3](#)). Additionally, industry funding was associated with a lack of test selection mistakes in the univariable analysis.

Multivariable logistic regression was performed to evaluate the roles of statistician involvement, journal impact factor, and funding source in the presence of mistakes ([Figure 4](#)). Absent statistician involvement (adjusted odds ratio aOR 2.49 [1.15-5.39], P=0.03) and low journal impact factor (aOR 3.82 [1.56-9.36], P=0.02) were associated with the presence of a statistical mistake of any type. A low journal impact factor was associated with the presence of a statistical reporting mistake (aOR 3.67 [1.68-8.05], P=0.02). The absence of statistician involvement was associated with the presence of a test selection mistake (aOR 3.12 [1.57-6.21], P=0.02).

After adjusting for impact factor and statistician involvement, the funding source was no longer a significant predictor of a test selection mistake. We confirmed that industry-funded studies were more likely to have statistician involvement and be published in a high-impact journal ([Table 4](#)).

We also performed a separate multivariable logistic regression with impact factor and statistician involvement as predictors of misinterpretations of the P-value, which were present in 28 (15%) articles. Absent statistician involvement (aOR 2.61 [1.01-6.80], P=0.04), but not low journal impact factor (1.91 [.84-4.34], P=0.14), was associated with misinterpretation of the P-value.

Exploratory logistic regressions for the first author's highest degree and statistical package did not yield additional significant predictors of mistakes after adjusting for statistician involvement and journal impact factor ([Figure 5](#) and [Figure 6](#)).

Discussion

Mistakes were common in our sample, with 147 (78%) of the sampled articles containing one or more statistical mistakes. We found that absent statistician involvement and low journal impact factor were associated with the presence of at least one statistical mistake. A low journal impact factor was associated with a statistical reporting mistake, whereas absent statistician involvement was as-

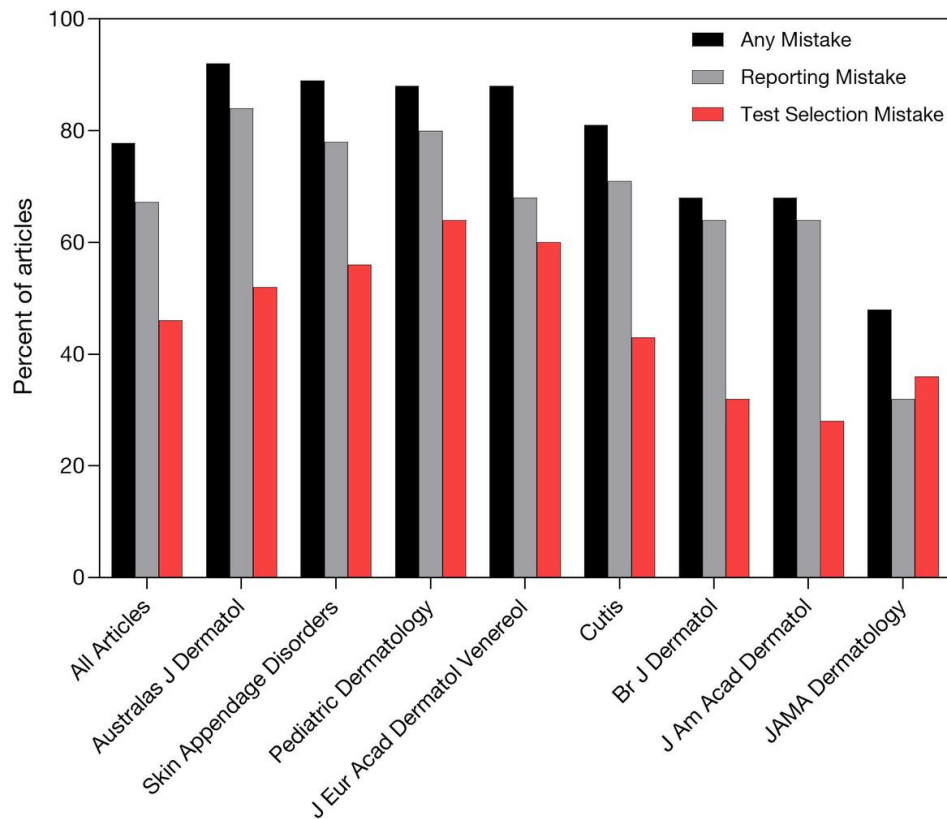


Figure 1. Frequency of mistakes by type and journal. Journal names are shortened to their National Library of Medicine abbreviations. Australas J Dermatol, Australasian Journal of Dermatology; J Eur Acad Dermatol Venereol, Journal of the European Academy of Dermatology and Venereology; Br J Dermatol, British Journal of Dermatology; J Am Acad Dermatol, Journal of the American Academy of Dermatology.

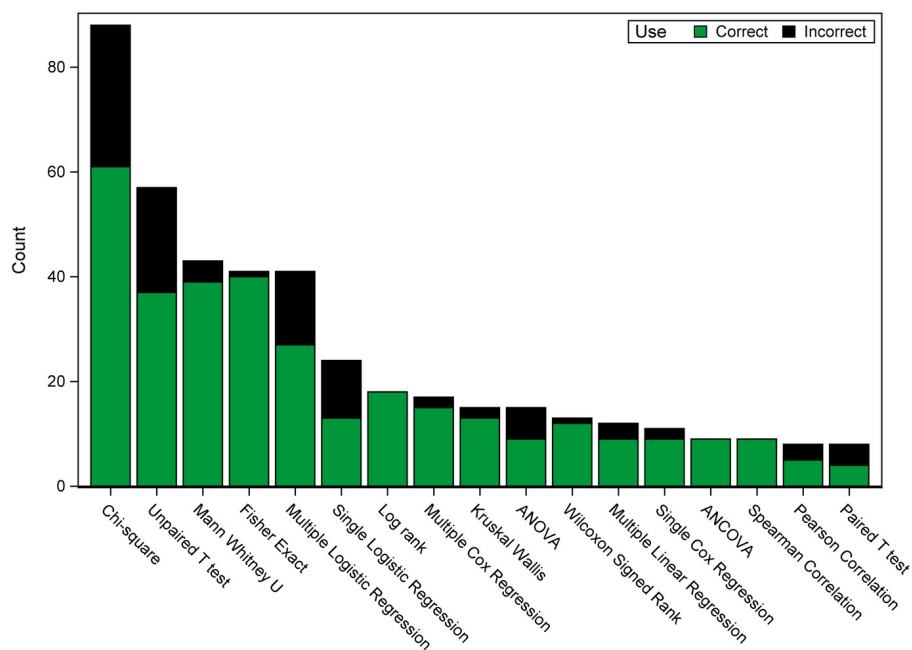


Figure 2. Correct and incorrect use of statistical tests that appeared greater than 5 times in the sample.

Table 3. Categories of Test Selection Mistakes for Tests That Appeared Greater Than 5 Times in the Sample

Test	Category of test selection mistake					Total Test Selection Mistakes	Total Uses
	Nonparametric test should have been used	Paired test should have been used	Insufficient sample size for regression	Other	Results not stated ^c		
χ^2 test ^a	0	0	0	18	9	27	88
Unpaired T-test	16	1	0	0	3	20	57
Mann-Whitney U	0	3	0	0	1	4	43
Fisher's exact	0	0	0	0	1	1	41
Logistic regression	0	0	25	0	2	27	65
Cox regression	0	0	4	0	0	4	28
Log-rank test	0	0	0	0	0	0	19
Linear regression	0	0	4	0	1	5	17
Kruskal-Wallis	0	1	0	0	1	2	15
ANOVA ^b	3	0	0	2	1	6	14
Wilcoxon signed rank	0	0	0	0	1	1	13
ANCOVA	0	0	0	0	0	0	9
Spearman Correlation	0	0	0	0	0	0	9
Pearson correlation	2	0	0	0	1	3	8
Paired T-test	3	0	0	0	1	4	8
Test not specified ^c	0	0	0	19	0	19	19

^a "Other" mistakes for the χ^2 test comprised 17 articles where the Fisher's exact test should have been used and one article where a paired samples test such as McNemar's test should have been used.

^b "Other" mistakes for ANOVA comprised two articles that interpreted significant differences between pairs of group means without performing post-hoc pairwise comparisons.

^c If authors did not state any results for their named test or reported results without naming the test performed, we counted this as a test selection mistake because we could not evaluate the test being used.

sociated with a test selection mistake. After adjusting for other article characteristics, we did not detect a difference in mistakes between funding sources, the highest degree of the first author, or the statistical package used.

Mistakes were more than twice as common in our sample compared to that of Neville et al (2006), which included the 2003 volumes of the *Archives of Dermatology* and *JAAD*.⁸ We included other journals that showed a higher rate of mistakes than *JAMA Dermatology* and *JAAD*, but the error rate among only these two journals was still higher at 58%. We also found a wider range of tests in use. New to our 2023 sample were types of re-

gression analysis, which were commonly used and misused.

Test selection mistakes, which were identified in 87 (46%) articles, may be more serious than reporting mistakes because inappropriate use of a statistical test can lead to invalid findings. Statistician involvement was the only studied characteristic that was protective against test selection mistakes after adjusting for other characteristics. In this study, the χ^2 test, t-tests, and logistic regression were the most used and misused tests. These tests are typically taught in introductory statistics courses, but some researchers may no longer be familiar

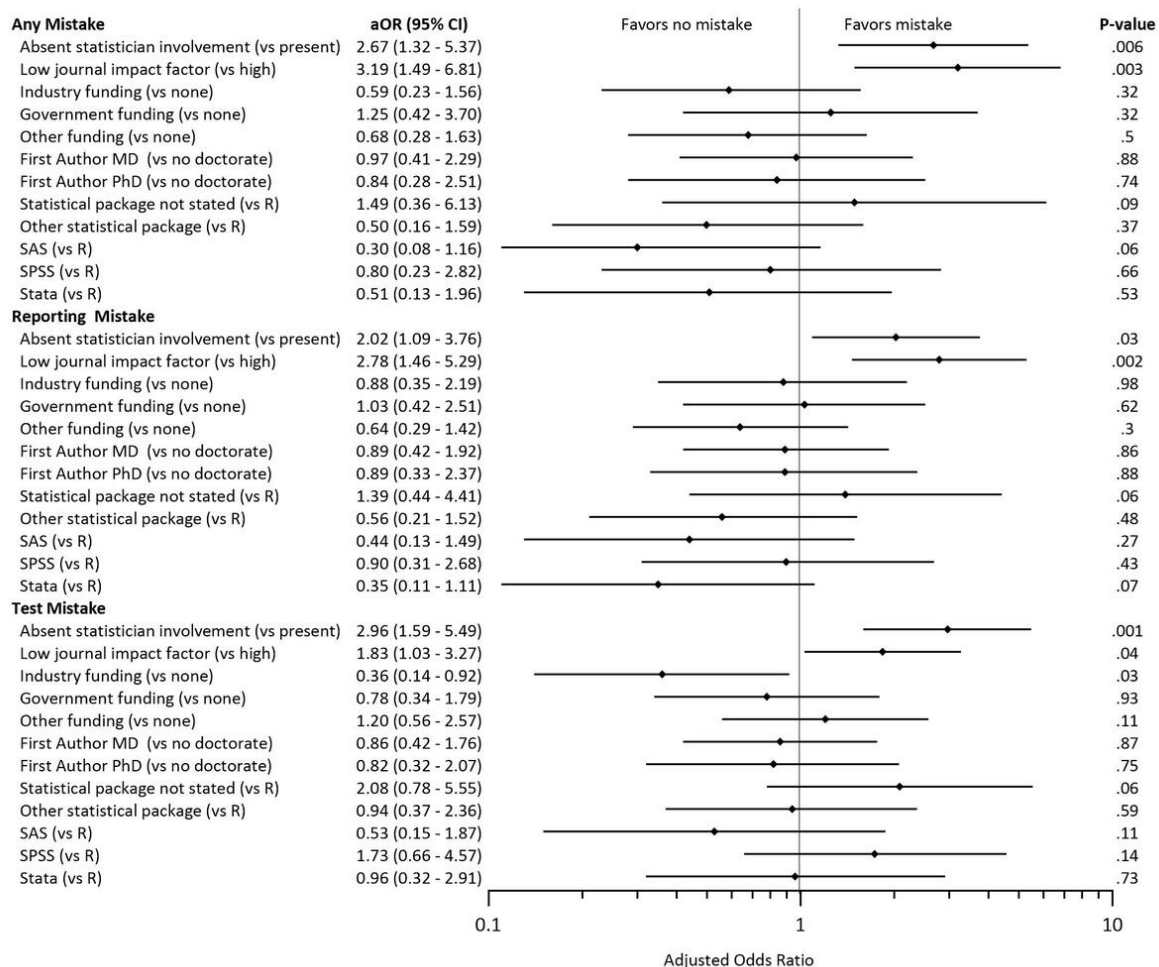


Figure 3. Unadjusted univariable logistic regressions for article characteristics and the presence of a statistical mistake. aOR: adjusted odds ratio. P-value: Benjamini-Hochberg adjusted p-values with significance at $p < .05$. The reference group for “no statistician involvement” is “statistician involvement,” the reference group for “low journal impact factor” is “high journal impact factor,” the reference group for funding sources is “no funding or not stated,” the reference group for first author’s highest degree is “bachelor’s or master’s degree,” and the reference group for statistical package is “R.”

with the assumptions made by these tests. Using our sample size criteria for regressions, we identified insufficient sample size for logistic regression in 38% of its uses and insufficient sample size for linear regression in 24% of its uses. The “10 events per variable” convention was proposed by Peduzzi et al (1995) and aims to avoid positive and negative biases that occur with fewer than 10 events per variable.^{31,32}

Depending on the characteristics of the data, some have found as few as 5 events per variable to be sufficient,²⁴ whereas others recommend as many as 50.²⁹ Given the nuance and debate surrounding sample sizes for regression, investigators should consider performing priori power and sample size calculations. If authors are unfamiliar with the tests they are conducting, they should also consider consulting an expert in statistics.

Though industry funding appeared protective against test selection mistakes in univariable analysis, the effect was not detected after adjusting for statistician involve-

ment and journal impact factor. This may be related to differences in study execution when industry funding is involved. Regulatory agencies require the statistical analysis plan to be finalized before study subject enrollment and statisticians are involved to ensure that the analysis is developed correctly. Also, industry-funded randomized controlled trials in this study were more often seen in high-impact journals.

Incomplete statistical reporting, which was seen in 127 (67%) articles, limits the reader’s ability to assess a study’s validity. Article authors may have obtained the necessary quantities, but ultimately did not include them in their manuscript. For reference, Neville et al (2006) found only 26% of dermatology articles to have reporting mistakes.⁸ More recently, McClean and Silverberg (2015) found that a significance level was not stated in 28% of dermatology randomized controlled trials.⁹ Though definitions of mistakes differ between studies, others have found reporting mistakes in 40% of cytopathology arti-

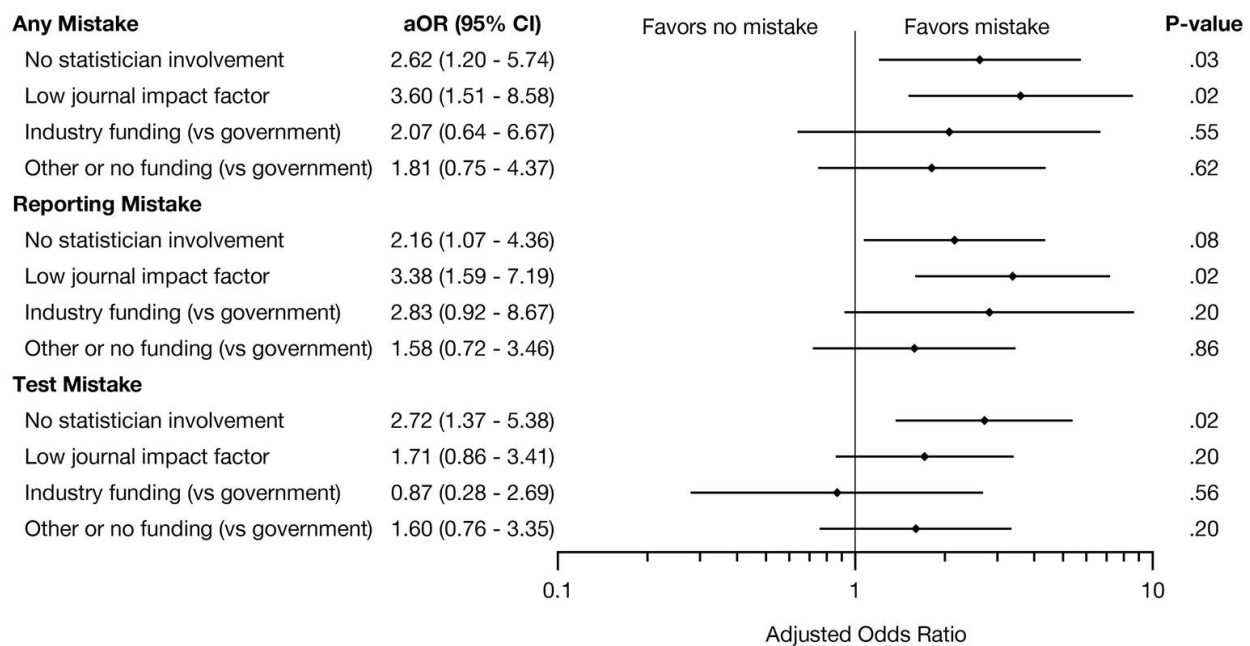


Figure 4. Multivariable logistic regressions for article characteristics with the outcome as the presence of a statistical mistake. Benjamini-Hochberg adjusted P-values with significance at $p < 0.05$. The reference group for “no statistician involvement” is “statistician involvement,” the reference group for “low journal impact factor” is “high journal impact factor,” and the reference group for “industry funding” and “other funding” is “no funding or not stated.” aOR: adjusted odds ratio.

Table 4. Industry-Funded Studies Had Higher Odds of Involving a Statistician and Being Published in a Journal With a High Impact Factor

Comparison	Statistician involvement		High-impact journal	
	OR (95% CI)	P-value	OR (95% CI)	P-value
Industry funding vs government funding	3.85 [1.15-12.86]	<.001	6.12 [1.51-24.82]	.05
Industry funding vs no funding or none stated	4.02 [1.70-9.57]	<.001	20.1 [5.60-72.5]	<.001
Industry funding vs other funding sources	7.23 [2.23-23.44]	<.001	6.50 [1.27-33.3]	.05

cles¹² and 42% of radiology articles.¹³ Though the current study did not evaluate the reporting of power calculations, they are often omitted from published randomized controlled trials in dermatology and plastic surgery.⁹⁻¹¹ In this study, only high journal impact factor protected against reporting mistakes.

We found misinterpretations of nonsignificant P-values using statements such as “trend toward significance” and “borderline significant” in 28 (15%) of the sampled articles. In doing this, the authors imply that more data drawn from the same population would yield a statistically significant result. Interestingly, simulations show that adding more data to a study with a “borderline significant” P-value does not necessarily lead to statistical significance.²⁰ Statistician involvement, but not journal impact factor, was protective against this mistake. Researchers should consider performing power calculations for their planned statistical tests before data collection, reviewing the proper interpretation of

nonsignificant P-values, and consulting a statistician. To convey practical significance in addition to statistical significance, authors should consider presenting the magnitude of their observed differences using confidence intervals and effect size measurements.^{37,38}

This study has several limitations that should be considered. Lack of access to original data limits our ability to verify statistical assumptions, and our criteria for correct test selection relies on rules of thumb. Furthermore, we cannot judge in this study whether identified “mistakes” impact an article in a clinically meaningful way. Statistical reporting mistakes may only limit a reader’s ability to critically evaluate an article while having no effect on the article’s conclusions. Also, certain test selection “mistakes” may still result in the “true” finding. A future study using actual or simulated data could better elucidate the impact of common test selection mistakes. Another limitation concerning the “statistician involvement” characteristic is the possibility that statisticians were involved

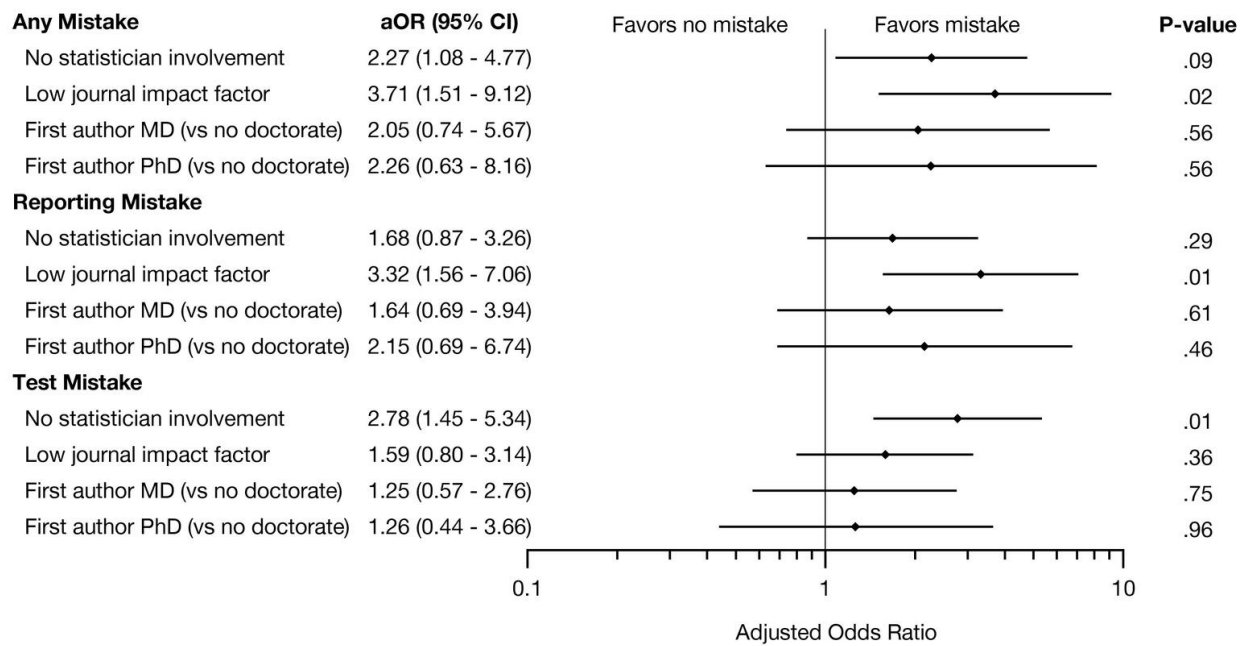


Figure 5. Multivariable logistic regression for the first author's highest degree, adjusted for statistician involvement and journal impact factor. aOR: adjusted odds ratio. P-value: Benjamini-Hochberg adjusted p-values with significance at $p < .05$. The reference group for "no statistician involvement" is "statistician involvement," the reference group for "low journal impact factor" is "high journal impact factor," and the reference group for first author's highest degree is "bachelor's or master's degree."

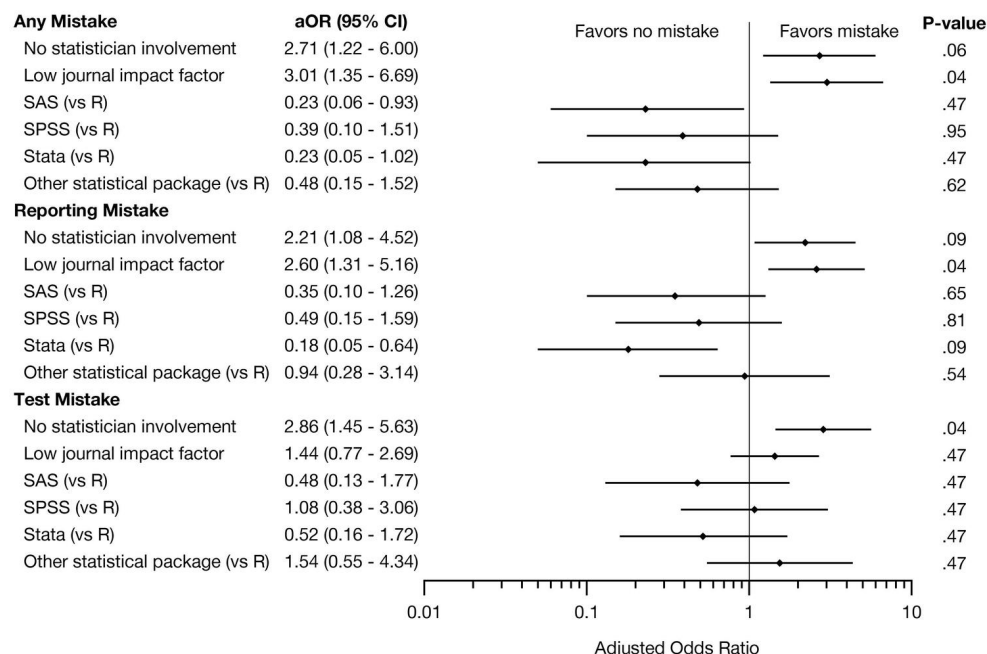


Figure 6. Multivariable logistic regression for the statistical package, adjusted for statistician involvement and journal impact factor. aOR: adjusted odds ratio. P-value: Benjamini-Hochberg adjusted p-values with significance at $p < .05$. The reference group for "no statistician involvement" is "statistician involvement," the reference group for "low journal impact factor" is "high journal impact factor," and the reference group for statistical package is "R."

in the research process but were not formally credited or that authors may have had prior statistical training

not reflected in their current affiliations. We selected only high and low-impact journals to maximize our study's

power to detect the effect of journal impact factor on mistakes if one was truly present. However, excluding “middle-impact” journals, which contain the majority of publications, limits the generalizability of our findings.

The statistician George Box said, “All models are wrong, but some are useful.” In other words, statistical models fail to perfectly describe reality, but better models can still yield valuable insights. To improve the quality of statistical analyses in dermatology journal articles, researchers should consider referencing best practices for statistical reporting and test selection. Researchers should also consider enlisting the help of an expert in statistics, especially when their sample size is small or when performing tests with which they are unfamiliar.

Conclusion

Statistical mistakes are prevalent in dermatology literature, and the most common statistical tests are fre-

quently misused. A low journal impact factor was associated with a statistical reporting mistake, whereas a lack of statistician involvement was associated with a test selection mistake. Funding source, first author’s highest degree, and statistical package were not associated with statistical mistakes. Researchers should review common statistical errors and consider involving a statistician in their work.

.....

Potential conflicts of interest

Alan Fleischer is a consultant for Bluefin and Incyte (fees). He is an investigator for Amgen, Bayer, Biogen, CellDex, Galderma, Incyte, Leo, and UCB (research support). He is a speaker for Imedic Healthcare Solutions (Hyderabad, India). The remaining authors have no potential conflicts of interest.

References

- Schulz KF, Altman DG, Moher D, the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med*. 2010;8:18. doi:[10.1186/1741-7015-8-18](https://doi.org/10.1186/1741-7015-8-18). PMID:20334633
- Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61:344-349. doi:[10.1016/j.jclinepi.2007.11.008](https://doi.org/10.1016/j.jclinepi.2007.11.008). PMID:18313558
- Silverberg JL. Study designs in dermatology: Practical applications of study designs and their statistics in dermatology. *J Am Acad Dermatol*. 2015;73:733-740. doi:[10.1016/j.jaad.2014.07.062](https://doi.org/10.1016/j.jaad.2014.07.062). PMID:26475533
- George BJ, Beasley TM, Brown AW, et al. Common scientific and statistical errors in obesity research. *Obesity*. 2016;24:781-790. doi:[10.1002/oby.21449](https://doi.org/10.1002/oby.21449). PMID:27028280
- Turkiewicz A, Luta G, Hughes HV, Ranstam J. Statistical mistakes and how to avoid them – lessons learned from the reproducibility crisis. *Osteoarthritis Cartilage*. 2018;26:1409-1411. doi:[10.1016/j.joca.2018.07.017](https://doi.org/10.1016/j.joca.2018.07.017). PMID:30096356
- Singh R, Mukhopadhyay K. Survival analysis in clinical trials: basics and must know areas. *Perspect Clin Res*. 2011;2(4):145-148. doi:[10.4103/2229-3485.86872](https://doi.org/10.4103/2229-3485.86872). PMID:22145125
- Carmona-Bayonas A, Jimenez-Fonseca P, Fernández-Somoano A, et al. Top ten errors of statistical analysis in observational studies for cancer research. *Clin Transl Oncol*. 2018;20:954-965. doi:[10.1007/s12094-017-1817-9](https://doi.org/10.1007/s12094-017-1817-9). PMID:29218627
- Neville JA, Lang W, Fleischer AB Jr. Errors in the Archives of Dermatology and the Journal of the American Academy of Dermatology from January Through December 2003. *Arch Dermatol*. 2006;142:737-740. doi:[10.1001/archderm.142.6.737](https://doi.org/10.1001/archderm.142.6.737). PMID:16785376
- McClean M, Silverberg JL. Statistical reporting in randomized controlled trials from the dermatology literature: a review of 44 dermatology journals. *Br J Dermatol*. 2015;173(1):172-183. doi:[10.1111/bjd.13907](https://doi.org/10.1111/bjd.13907). PMID:25989239
- Weinberg T, Wang G, Lam K, Kitchen J, Chan AW. Reporting of sample-size calculations for randomized trials in dermatology: comparison of publications with registries. *Br J Dermatol*. 2019;180:929-930. doi:[10.1111/bjd.17332](https://doi.org/10.1111/bjd.17332). PMID:30339744
- Ayeni O, Dickson L, Ignacy TA, Thoma A. A systematic review of power and sample size reporting in randomized controlled trials within plastic surgery. *Plast Reconstr Surg*. 2012;130:78e-86e. doi:[10.1097/PRS.0b013e318254b1d1](https://doi.org/10.1097/PRS.0b013e318254b1d1). PMID:22743957
- Bahar B, Pambuccian SE, Barkan GA, Akdaş Y. The use and misuse of statistical methods in cytopathology studies: review of 6 journals. *Lab Med*. 2019;50:8-15. doi:[10.1093/labmed/lmy036](https://doi.org/10.1093/labmed/lmy036). PMID:30020520
- Karadeniz PG, Uzabacı E, Kuyuk SA, et al. Statistical errors in articles published in radiology journals. *Diagn Interv Radiol*. 2019;25:102-108. doi:[10.5152/dir.2018.18148](https://doi.org/10.5152/dir.2018.18148). PMID:30582574
- Nour-Eldein H. Statistical methods and errors in family medicine articles between 2010 and 2014-Suez Canal University, Egypt: a cross-sectional study. *J Fam Med Prim Care*. 2016;5:24-33. doi:[10.4103/2249-4863.184619](https://doi.org/10.4103/2249-4863.184619). PMID:27453839
- Allam RM, Noaman MK, Moneer MM, Elattar IA. Assessment of statistical methodologies and pitfalls of dissertations carried out at National Cancer Institute, Cairo University. *Asian Pac J Cancer Prev APJCP*. 2017;18:231-237. PMID:28240524
- Masic I, Jankovic S, Begic E. PhD Students and the most frequent mistakes during data interpretation by statistical analysis software. *Stud Health Technol Inform*. 2019;262:105-109. PMID:31349277
- The Clarivate Analytics Impact Factor. Web of Science Group. 2024. Accessed August 27, 2024. <https://clarivate.com/webofsciencegroup/essays/impact-factor/>
- El Tecle NE, Urquiaga JF, Griffin ST, et al. Misinterpretations of null hypothesis significance testing results near the P-value threshold in the neurosurgical literature. *World Neurosurg*. 2022;159:e192-e198. doi:[10.1016/j.wneu.2021.12.030](https://doi.org/10.1016/j.wneu.2021.12.030). PMID:34915206

19. Nead KT, Wehner MR, Mitra N. The use of "trend" statements to describe statistically nonsignificant results in the oncology literature. *JAMA Oncol.* 2018;4:1778-1779. doi:[10.1001/jamaoncol.2018.4524](https://doi.org/10.1001/jamaoncol.2018.4524). PMID:30347042
20. Wood J, Freemantle N, King M, Nazareth I. Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data. *BMJ.* 2014;348:g2215. doi:[10.1136/bmj.g2215](https://doi.org/10.1136/bmj.g2215). PMID:24687314
21. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. *Korean J Anesthesiol.* 2017;70:144-156. doi:[10.4097/kjae.2017.70.2.144](https://doi.org/10.4097/kjae.2017.70.2.144). PMID:28367284
22. Wilson Van Voorhis CR, Morgan BL. Understanding power and rules of thumb for determining sample sizes. *Tutor Quant Methods Psychol.* 2007;3:43-50. doi:[10.20982/tqmp.03.2.p043](https://doi.org/10.20982/tqmp.03.2.p043)
23. Camilli G. The relationship between Fisher's exact test and Pearson's chi-square test: a bayesian perspective. *Psychometrika.* 1995;60:305-312. doi:[10.1007/BF02301418](https://doi.org/10.1007/BF02301418)
24. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox Regression. *Am J Epidemiol.* 2007;165:710-718. doi:[10.1093/aje/kwk052](https://doi.org/10.1093/aje/kwk052). PMID:17182981
25. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol.* 2014;14:137. doi:[10.1186/1471-2288-14-137](https://doi.org/10.1186/1471-2288-14-137). PMID:25532820
26. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol.* 2016;76:175-182. doi:[10.1016/j.jclinepi.2016.02.031](https://doi.org/10.1016/j.jclinepi.2016.02.031). PMID:26964707
27. Long JS, Freese J. Estimation, testing, fit, and interpretation. In: *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. Stata Press; 2006:65.
28. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol.* 2009;9:56. doi:[10.1186/1471-2288-9-56](https://doi.org/10.1186/1471-2288-9-56). PMID:19635144
29. Bujang MA, Sa'at N, Sidik TMITAB, Joo LC. Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *Malays J Med Sci MJMS.* 2018;25:122-130. doi:[10.21315/mjms2018.25.4.12](https://doi.org/10.21315/mjms2018.25.4.12). PMID:30914854
30. Væth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Stat Med.* 2004;23:1781-1792. doi:[10.1002/sim.1753](https://doi.org/10.1002/sim.1753). PMID:15160408
31. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J Clin Epidemiol.* 1995;48:1503-1510. doi:[10.1016/0895-4356\(95\)00048-8](https://doi.org/10.1016/0895-4356(95)00048-8). PMID:8543964
32. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* 1996;49:1373-1379. doi:[10.1016/S0895-4356\(96\)00236-3](https://doi.org/10.1016/S0895-4356(96)00236-3). PMID:8970487
33. Jenkins DG, Quintana-Ascencio PF. A solution to minimum sample size for regressions. *PLOS ONE.* 2020;15:e0229345. doi:[10.1371/journal.pone.0229345](https://doi.org/10.1371/journal.pone.0229345). PMID:32084211
34. Bujang MA, Sa'at N, Sidik TMITAB. Determination of minimum sample size requirement for multiple linear regression and analysis of covariance based on experimental and non-experimental studies. *Epidemiol Biostat Public Health.* 2017;14. doi:[10.2427/12117](https://doi.org/10.2427/12117)
35. Harrell FE. Multivariable modeling strategies. In: Bickel P, Diggle P, Feinberg SE, Gather U, Olkin I, Zeger S, eds. *Regression Modeling Strategies*. Springer International Publishing; 2015:72-73. doi:[10.1007/978-3-319-19425-7_4](https://doi.org/10.1007/978-3-319-19425-7_4)
36. Chatterjee S, Simonoff JS. Collinearity. In: *Handbook of Regression Analysis*. Wiley; 2013:26-29. doi:[10.1002/9781118532843](https://doi.org/10.1002/9781118532843)
37. Dunkler D, Haller M, Oberbauer R, Heinze G. To test or to estimate? P-values versus effect sizes. *Transpl Int Off J Eur Soc Organ Transplant.* 2020;33:50-55. doi:[10.1111/tri.13535](https://doi.org/10.1111/tri.13535). PMID:31560143
38. Lytsy P. P in the right place: Revisiting the evidential value of P-values. *J Evid-Based Med.* 2018;11:288-291. doi:[10.1111/jebm.12319](https://doi.org/10.1111/jebm.12319). PMID:30398018