Original Article

Statistical methods and errors in family medicine articles between 2010 and 2014-Suez Canal University, Egypt: A cross-sectional study

Hebatallah Nour-Eldein[1]

[1]*Department of Family Medicine, Suez Canal University, Ismailia, Egypt*

Abstract

Background: With limited statistical knowledge of most physicians it is not uncommon to find statistical errors in research articles. **Objectives:** To determine the statistical methods and to assess the statistical errors in family medicine (FM) research articles that were published between 2010 and 2014. **Methods:** This was a cross-sectional study. All 66 FM research articles that were published over 5 years by FM authors with affiliation to Suez Canal University were screened by the researcher between May and August 2015. Types and frequencies of statistical methods were reviewed in all 66 FM articles. All 60 articles with identified inferential statistics were examined for statistical errors and deficiencies. A comprehensive 58-item checklist based on statistical guidelines was used to evaluate the statistical quality of FM articles. **Results:** Inferential methods were recorded in 62/66 (93.9%) of FM articles. Advanced analyses were used in 29/66 (43.9%). Contingency tables 38/66 (57.6%), regression (logistic, linear) 26/66 (39.4%), and *t*-test 17/66 (25.8%) were the most commonly used inferential tests. Within 60 FM articles with identified inferential statistics, no prior sample size 19/60 (31.7%), application of wrong statistical tests 17/60 (28.3%), incomplete documentation of statistics 59/60 (98.3%), reporting *P* value without test statistics 32/60 (53.3%), no reporting confidence interval with effect size measures 12/60 (20.0%), use of mean (standard deviation) to describe ordinal/nonnormal data 8/60 (13.3%), and errors related to interpretation were mainly for conclusions without support by the study data 5/60 (8.3%). **Conclusion:** Inferential statistics were used in the majority of FM articles. Data analysis and reporting statistics are areas for improvement in FM research articles.

Keywords: Reporting, research articles, statistical errors, statistical methods

Introduction

Statistical analysis is a part of the process of writing a scientific article. It is an essential technique that enables a medical researcher to draw meaningful conclusions from their data analysis.[1] Statisticians and methodological experts should be consulted during the study design, analysis, and manuscript writing phases to improve the quality of research and to ensure clear and appropriate application of quantitative methods.[2] On the other hand, many researchers have difficulty or delay in getting a statistical advice or the statistician's involvement in their research from early stages of study design.[3]

The statistical software programs over the past years expanded analytic capabilities and broadened the spectrum of appropriate statistical options.[4] Researchers have to be adequately trained in the application of statistics for biomedical research.[5] It is of great importance to implement statistics accurately and carefully so that the results will be more credible and meaningful.[6] With limited statistical knowledge of most physicians, it is not uncommon to find statistical errors. Statisticians have documented that statistical errors are common, and at least one error could be found in about 50% of the published articles.[7]

Many journals adopt guidelines to improve reporting manuscripts as the Consolidated Standards of Reporting Trials,[8] the

Address for correspondence: Dr. Hebatallah Nour-Eldein, Department of Family Medicine, Suez Canal University, Ismailia, Egypt.
E-mail: hebanour20@hotmail.com

Access this article online

Quick Response Code:

Website: www.jfmpc.com

DOI: 10.4103/2249-4863.184619

How to cite this article: Nour-Eldein H. Statistical methods and errors in family medicine articles between 2010 and 2014-Suez Canal University, Egypt: A cross-sectional study. J Family Med Prim Care 2016;5:24-33.

Transparent Reporting of Evaluations with Nonrandomized Designs[9] Statement, the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE),[10] Guideline specific for reporting statistical analysis, the Statistical Analyses and Methods in the Published Literature (SAMPL Guidelines).[11] These guidelines and others for reporting scientific research were also available at EQUATOR network.[12]

"Because society depends on sound statistical practice, all practitioners of statistics, whatever their training and occupation, have social obligations to perform their work in a professional, competent, and ethical manner." (Ethical Guidelines for Statistical Practice, American Statistical Association, 1999.)[13] These principles should also guide the statistical work of professionals in all other disciplines that use statistical methods.[13]

All assistant lecturers in Suez Canal University have to attend a statistical course and evaluation in the educational curriculum of doctorate degree. Publishing research articles are a mandatory process in postdoctorate promotion. Revising the statistical reporting in the past articles aimed to improve the future manuscripts for publication. This study had two objectives. Objective 1: To determine the types and frequencies of statistical methods in family medicine (FM) research articles. Objective 2: To assess the quantity and character of statistical errors and deficiencies.

Methods

This was a cross-sectional study, the data were collected retrospectively. It was conducted by the researcher between May and August 2015. FM research article selection: Included all original FM articles that were published by FM authors with affiliation to Suez Canal University; all published articles in different National and International Medical Journals between 2010 and 2014. All articles were downloaded in full text as a portable document format. Commentaries, letters to the editor, review articles, and articles with themes that were not related to the scope of FM were excluded.

FM research article search: (1) All published articles by FM authors with affiliation to Suez Canal University were available in FM Department database from 1992 to 2013 in 39 medical journals as previously collected in a previous study.[14] The researcher updated the search to include all articles that were published in 2014 on(2) National Journal Websites (The Egyptian journal of Community Medicine, The Medical Journal of Cairo University, and Suez Canal University) and (3) Google and PubMed search for other publications.

The searched articles were published in 24 medical journal (African Safety Promotion, Annals of Burns and Fire Disasters, Eastern Mediterranean Health Journal, Egyptian Journal of Neurology and Psychiatry, Elective Medical Journal, FM and Medical Science Research, International Journal of Health Sciences, International Journal of Medicine and Public Health, Journal of American Science, Journal of FM and Primary Care, Journal of Family and Community Medicine, Journal of the Egyptian Public Health, Journal of Tibah University Medical Sciences, Medical Journal of Cairo University, Middle East Journal of FM, Medical Teacher, Open Access Scientific Reports, Pan African Medical Journal, Peer Journal, Saudi Medical Journal, Suez Canal University Medical Journal, The Arab Journal of Psychiatry, The Egyptian Journal of Community Medicine, and The Egyptian Rheumatologist).

Main outcomes were the types and frequencies of statistical methods in all screened articles; the statistical errors and deficiencies related to study designs, application, and documentation of statistical analyses, data presentation and interpretation in articles with identified inferential statistics.

Statistical methods

Types and frequencies of applied statistical methods were recorded for all the 66 articles and classified into 15 out of 21 categories, earlier used by Emerson and Colditz in 1983 [Table 1].[15] If the same statistical method was repeatedly used in the same article, the method was documented once; however, if more than one statistical technique were used in one article, each of them was considered separately.

Articles containing identified inferential statistical methods beyond descriptive statistics were further classified into basic or advanced analyses according to the sophistication of applied statistical techniques as previously used by Strasak *et al.,* 2007.[16,17] Basic analyses included *t*-test, simple contingency table analysis, nonparametric methods, one-way analysis of variance (ANOVA), correlation, and simple linear regression. Advanced analyses included any method of statistical modeling, multivariate analysis (e.g., multivariate ANOVA, multivariate analysis of covariance MANCOVA), advanced contingency table analysis, epidemiologic statistics, or survival analysis.

Statistical errors

All articles that included identified basic or advanced inferential methods beyond descriptive statistics were included. The articles were screened using a comprehensive 58-item checklist: [Appendix 1]; 46 items were the checklist developed and used in two studies by Strasak *et al.*, 2007[16,17] and the researcher added 10 items specific to regression analysis and one item in the presentation of results regarding the error of not reporting the test statistics, these additional items originated from SAMPL guidelines[11] and previously used in the study by Hassan *et al.,* 2015.[18] Another item was added in the documentation related to reporting the name of the statistical software package used in statistical analysis.[11] In the application of the checklist, the error committed was restricted to obvious ones that could clearly be identified. Unable to assess/not clear was recorded if an article contained insufficient information to assess a specific item of the checklist. Application correct was given to perfect issues. The researcher was adherent to

Table 1: Categories of statistical procedures used to assess the statistical contents of articles

Category	Brief description
1. No statistical methods or descriptive statistics	No statistical content, or descriptive statistics only (e.g., percentages, means standard deviations, standard errors, histograms
2. Contingency tables	Chi-square tests, Fisher's exact test, McNemar's test
3. Multiway tables	Mantel–Haenszel procedure, log-linear models
4. Epidemiological statistics	Relative risk, odds ratio, log odds, measures of association, sensitivity, specificity
5. *t*-tests	One-sample matched pair and two-sample *t*-test
6. Pearson correlation	Classical product moment-correlation
7. Simple linear regression	Least-squares regression with one predictor and one response variable
8. Multiple regression	Includes polynomial regression and stepwise regression
9. Analysis of variance	Analysis of variance, analysis of co-variance, and *F*-tests
10. Multiple comparisons	Procedures for handling multiple inferences on same data sets (e.g., Bonferroni techniques, Scheffe's contrasts, Duncan's multiple range procedures, Newman–Keuls procedure)
11. Nonparametric tests	Sign test, Wilcoxon signed ranks test, Mann–Whitney U test
12. Nonparametric correlation	Spearman's rho, Kendall's tau, test for trend
13. Life table	Actuarial life-table, Kaplan–Meier estimate of survival
14. Regression for survival	Includes Cox regression and logistic regression
15. Other survival analysis	Breslow's Kruskal–Wallis, log-rank, Cox model for comparing survival
16. Adjustment and standardization	Pertains to incidence rates and prevalence rates
17. Sensitivity analysis	Examines sensitivity of outcome to modest changes in parameters of model or in other assumptions
18. Transformation	Use of data transformation (e.g., logarithms) often in regression
19. Power	Loosely defined, includes use of the size of detectable (or useful) difference in determining sample size
20. Cost benefit analysis	The process of combining estimates of cost and health outcomes to compare policy alternatives
21. Other	Anything not fitting the above headings includes cluster analysis, discriminant analysis, and some mathematical modeling

Emerson and Colditz (1983)[15]

the guidelines by Strasak *et al*., 2007,[19] and SAMPL[11] which provided a detailed clarification to the items of the checklist. Some items were more detailed by others.[20-25]

Categorization of study designs within FM research articles was previously sorted in the study by Abdulmajeed *et al.,* 2013:[14] quantitative study designs, observational studies (cross-sectional, case–control, and cohort), and intervention studies (with randomization or without randomization). None of the published FM articles were with a cohort design.

Incompatibility of the applied tests with data type was checked based on Chi-square tests are suitable for categorical data presented in frequencies and percentages. Parametric tests (e.g., *t*-test, ANOVA) are suitable with normally distributed continuous data presented in mean (standard deviation [SD]). Nonparametric tests (e.g., Mann–Whitney U/Wilcoxn rank sum, Kruskal–Wallis H, Wilcoxon signed rank, and Friedman's tests) are suitable in comparison of continuous data not normally distributed expressed in medians and (interquartile range) or mean ranks.[20,21]

Independence: Student's *t*-test and Wilcoxon test were checked for reporting the used variant paired/dependent in comparison of pre- and post-experimental studies and matched controlled studies or unpaired/independent in comparison of two independent samples.[20,22] Furthermore, paired and matched comparisons were checked for the use of (paired *t*-test, Wilcoxon signed rank-test, and Mcnemar test).[22]

Checking the distribution of continuous data (normal or nonnormal) is a prerequisite to the presentation of descriptive statistics and the selection of parametric or nonparametric tests.[20] The assumption of normality is that the normal distribution of variables in case of *t*-test or ANOVA and the distribution of residuals in case of regression. The assumption of homogeneity of variance requires equal population variances per group in case of *t*-test and ANOVA.[11,23]

Skewness of data was checked based on the two tricks by Altman and Bland 1996[24] as the data were likely to be skewed if the mean was smaller than twice the SD and highly skewed if the mean was smaller than SD. The second trick in case of several groups stated that if SD increased as the mean increased was a good indication of positive skewed data.

Adequate cell size was checked in Chi-square test no more than 20% of the cells should have expected frequencies <5. For example, within 2 × 2 tables, no cell should have an expected frequency <5.[25] Fisher exact test is used when this assumption is not met. The expected frequency of a contingency table cell was calculated as expected cell frequency = (row total × column total)/grand total.[25]

In presentation of data, confidence interval (CI) as a measure of precision was checked in reporting effect size measures such as risks (e.g., absolute risks; relative risk differences); rates (e.g., incidence rates; survival rates); ratios (e.g., odds ratios, hazards ratios); and in reporting coefficients in association, correlation, and regression.[11]

Data analysis

The data were extracted from the published articles then entered and analyzed using a Statistical Package for Social Sciences program (SPSS, version 20 IBM, Chicago, IL, USA). Data were presented using descriptive statistics in the form of frequencies and percentages for the qualitative variables.

Results

Statistical methods

The majority of the reviewed articles contained inferential statistical tests (93.9%). More than half of the screened articles contained contingency tables 38/66 (57.6%). Regression analyses (logistic and multiple linear) were recorded in more than one-third of the searched articles 26/66 (39.4%) and a quarter of articles 17/66 (25.8%) mentioned *t*-test. The least used inferential tests were Wilcoxon signed rank test, Kruskal–Wallis H, and McNemar 1/66 (1.5%) for each test. Furthermore, normality test and log transformation were mentioned in only (1.5%). More than one-third of articles contained advanced analyses 29/66 (43.9%) [Table 2].

Table 2: Types and frequencies of statistical methods

Statistical methods	Total articles	
	n=66	100%
Descriptive statistics only	4	6.1
Inferential statistics	62	93.9
Contingency tables	38	57.6
χ^2	37	56.1
Fisher exact	9	13.6
McNemar	1	1.5
Regression	26	39.4
Logistic regression	18	27.3
Multiple linear regression	8	12.1
t-test	17	25.8
t-test (unspecified paired/unpaired)	15	22.7
Paired *t*-test	3	4.5
ANOVA	7	10.6
One-way ANOVA	6	9.1
Advanced ANOVA (ANCOVA-MANCOVA)	2	3.0
Multiple comparisons	4	6.1
Correlation	5	7.6
Pearson correlation	3	4.5
Spearman's rho correlation	2	3.0
Nonparametric tests	5	7.6
Mann–Whitney U-test	2	3.0
Kruskal–Wallis H-test	1	1.5
Wilcoxon signed rank	1	1.5
Friedman	1	1.5
Epidemiological statistics (OR)	1	1.5
Unidentified inferential method/test	2	3.0
CI	17	25.8
Power and sample size calculation	27	40.9
Log transformation	1	1.5
Others (Shapiro normality test)	1	1.5
Sophistication of statistical analyses		
No/descriptive/unidentified methods	6	9.1
Basic analyses	31	47.0
Advance analyses	29	43.9

ANOVA: Analysis of variance; ANCOVA: Analysis of covariance; MANCOVA: Multivariate analysis of covariance; OR: Odds ratio; CI: Confidence intervals

Deficiencies in study design

No mentioned sample size calculation was found in approximately one-third of the articles 19/60 (31.7%). Methods of randomization/allocation to intervention were not clearly stated in 2/60 (3.3%) which represented 2/2 (100.0%) of randomized controlled trial (RCT) articles [Table 3].

Table 3: Statistical deficiencies and errors in study design

	Articles with inferential statistics	
	n=60	100%
Category		
Study design		
No sample size calculation/power calculation (overall)	19	31.7
Cross-sectional	14	23.3
Case-control	2	3.3
Intervention without randomization	2	3.3
Intervention with randomization	1	1.7
Method of randomization/sampling	2	3.3

Errors in statistical analysis

Wrong analyses were recorded in more than a quarter of articles as 17/60 (28.3%). Failure to proof/report that Student's *t*-test assumptions is not violated in a quarter all articles 15/60 (25.0%), in most of articles with *t*-test 15/17 (88.2%). The assumptions of multiple regression were not reported in 6/60 (10.0%) which represented most of articles with multiple linear regression 6/8 (75.0%) that mentioned the use of multiple regression. Use of Chi-square test instead of Fisher's exact was mentioned in 5/60 (5.8%). Failure to include alpha correction in multiple comparisons was in 4/60 (6.7%) of all articles and these were all articles 4/4 (100.0%) that mentioned the use of multiple comparisons [Table 4].

Errors in documentation

Fifty-nine articles (98.3%) showed failure to define details of a test performed. Failure to state number of tails of significance tests was at 59/60 (98.3%). One-fifth of the articles, i.e., 12/60 (20.0%) showed failure to specify which test was performed on a given set of data when multiple tests were used. In a quarter of articles, there was failure to state if *t*-test was paired or unpaired 15/60 (25.0%) [Table 5].

Errors in data presentation

More than half of the articles, i.e., 32/60 (53.3%) showed no value of test statistics (at least one table in the article contains this error). One-fifth of the articles., i.e., 12/60 (20.0%) presented only *P* value without CIs for main effect size measures. Use of mean (SD) to describes ordinal/nonnormal data 8/60 (13.3%). Numerical imprecision was found in 6/60 (10.0%) [Table 5].

Table 4: Statistical deficiencies and errors in data analysis

	Articles with inferential statistics		
	n=60	100%	Percentage per articles with special errors
Category			
Data analysis			
Use of a wrong statistical test	17	28.3	-
Incompatibility of statistical test with type of data examined	14	23.3	-
Inappropriate use of parametric methods	8	13.3	-
Unpaired tests for paired data or vice versa	5	8.3	-
Special errors with multiple-comparisons (Type I error inflation)			Percentage/7 articles
Failure to include a multiple-comparison correction/*a*-level correction	4	6.7	57.1
Special errors with student's *t*-test			Percentage/17 articles
Failure to proof/report that test assumptions are not violated	15	25.0	88.2
Special errors with Chi-square tests			Percentage/37 articles
Use of Chi-square when expected numbers in a cell are <5	5	8.3	13.5
Special errors with regression analysis			Percentage/8 articles
No description of assumptions of the analysis (e.g., an analysis of residuals confirmed the assumptions of linearity)	6	10.0	75.0
No model validation procedure was given	5	8.3	62.5
For either simple or multiple (multivariable) regression analyses, regression equation was not reported	4	6.7	50.0
Measure of the model's "goodness-of-fit" to the data was not reported (r^2 in simple R^2 in multiple regression)	4	6.7	50.0
For multiple regression: no report of variable selection process (e.g., forward-stepwise; best subset)	3	5.0	37.5

Errors in data interpretation

Errors related to conclusions without support by the study data 5/60 (8.3%), reporting significance without data analysis and missing the discussion of the problem of multiple significance testing were shown in only 2/60 (3.3%) of articles [Table 5].

Discussion

The use of inferential statistics was found in the vast majority of the screened articles giving the advantage and evidence of their analytic character. Although the more frequently recorded deficiencies were related to inadequate documentation of the used statistical methods, the use of wrong statistical test in more than a quarter of the articles was a major finding.

Contingency table analysis was used twice more frequently than *t*-test among the simple tests. These results were relatively in agreement with the British study in Family Practice Articles over 1 year by Rigby *et al.*, 2004[26] and Emerson and Colditz 1983[15] in articles with cross-sectional studies. Contingency tables were less used in prospective and retrospective study designs in other studies.[16,17,27] The use of survival analysis and Chi-square tests followed by nonparametric tests was observed in American surgical articles.[28] The selection of test depends partly on types of study designs.

The use of normality tests was mentioned in only 1.5% of articles could explain in part by the inappropriate presentation of skewed data in mean (SD) and inappropriate use of parametric methods for skewed data. Checking the normality was lower than in another study.[26] Multiple comparisons were used only in 4/7 of the reported ANOVA tests; these results were higher than findings by Olsen in 2003[20] and partly consistent with the results of ignoring or misusing the method of multiple pair-wise comparisons in ANOVA in the analysis of Chinese articles.[29] The presentation of unidentified method is an error and deficiency in both documentation and presentation. However, these unidentified methods were excluded from further assessment.

Basic analyses were used slightly more in articles than advanced analyses. These results were nearly consistent with other studies.[28,30] Pet *et al.,* 2014,[30] mentioned that the sophistication of statistical methods are going to be increased over time and avoiding use of advanced techniques may miss many possible important inferences from the same data. The difference in selection of inferential statistical tests depends on study designs, the main study hypothesis, type of data, and independence of variables.[22]

One of two RCTs was with no sample size calculation. This point is crucial to detect treatment effects.[19,21] If no sample size calculation was used the study size must be justified, for example, all available patients in two centers were included and a sample size calculation was not relevant. Although the method of randomization/allocation to intervention was not clearly stated in 3.2% of all articles which represented all searched RCT articles. A full explanation of the method of randomization and sampling should be mentioned as all inferential statistical techniques are valid only for random samples.[19]

Table 5: Statistical errors and deficiencies in documentation, data presentation, and interpretation

	Articles with inferential statistics	
	n=60	100%
Category		
Documentation		
Failure to specify/define all applied tests clearly and correctly	59	98.3
Failure to state number of tails	59	98.3
Failure to state if test was paired or unpaired	15	25.0
Failure to state which values of *P* indicate statistical significance	13	21.7
Failure to specify all tests was performed on a given set of data	12	20.0
"Where appropriate" statement	2	3.3
Name of the statistical software used in the analysis was not mentioned	9	14.0
Presentation		
Reporting *P* value without test statistics	32	53.3
No CI for main effect size measures presented	12	20.0
CI given for each group rather than for the contrast	1	1.7
Use of mean (SD) to describe ordinal/nonnormal data	8	13.3
"Mean" but no indication of variability of data	2	3.3
Failure to define ± notion for describing variability; use of unlabeled error bars	1	1.7
Numerical results and *P* values given to too many (or too few) decimal places (e.g., $P<0.000000$)	6	10.0
P=NS, $P<0.05$, $P>0.05$ etc., instead of reporting exact *P* values	1	1.7
Interpretation		
Drawing conclusions not supported by the study data	5	8.3
Significance claimed without data analysis or statistical test mentioned	2	3.3
Missing discussion of the problem of multiple significance testing if occurred	2	3.3
Failure to consider CI's when interpreting "NS" differences	1	1.7

CI: Confidence interval; NS: Nonsignificant; SD: Standard deviation

Unfortunately, incompatibility of statistical test with the type of data examined and the inappropriate use of parametric methods on skewed data was higher than in British and Australian Clinical Articles,[16,17] and the latter item was higher than others.[5,27] The use of unpaired tests for paired data was nearly similar to others.[5,17] The improper use of Pearson's Chi-square test instead of the McNemar test was found in analysis of correlated and dependent categorical variables, and this may lead to misleading conclusions and recommendations.[31,32]

Failure to proof or report that the *t*-test assumptions and not including appropriate multiple comparison α-level correction was lower than in other studies.[5,16,17] Correcting the alpha level by dividing 0.05 by times of multiple comparisons maintains a "family wise" error rate of 5% likelihood of Type I error.[33] Most of the errors related to application and reporting regression models in the current study were related to multiple linear regression. The check of the assumptions in regression analysis was not mentioned in most of the articles using multiple linear regression, this error was higher than in another study.[18] Chi-square was incorrectly used when expected cells <5 in 8.1% of the articles. These results were similar to the Indian study[5] and lower than in the articles of New England Journal of Medicine and Chinese articles.[16,34]

All the errors related to statistical analysis could be due to the use of new statistical software by nonexperts. Hoekstra *et al.*, 2012,[35] set four possible explanations for failing to check for violations of assumptions such as lack of knowledge of the assumptions, methods of checking the assumptions, the problem of possible violation of an assumption, and lack of knowledge of an alternative if an assumption was violated.

Multiple and different deficiencies in documentation of the used statistical methods were nearly similar to others.[16,17,27] Failure to state number of tails was higher than other studies.[16,17] Hypothesis tests whether one- or two-sided with *P* value were the most unreported while fail to mention the name of software by which analyzed the data was lower than other study.[4,30] Deficiencies in documentation mean nonadherence to the guidelines of reporting statistics.

Clear statistics should be reported, either through labels in the table or as a footnote.[21] Reporting *P* value only without test statistics in at least one table was in 53.3% of the articles. These results were in agreement with the study by Hassan *et al.*, 2015.[18] It is recommended to report observed values of test statistics (e.g., *t*-test, χ^2-test) with tabulated values and *P* value.[23] From the reported observed test statistics, tabulated values and its degrees of freedom, it is possible to compute the observed *P* value with most statistical packages and check the congruence of the results.[36]

Inappropriate reporting of mean (SD) to describe ordinal/nonnormal data for nonparametric tests was higher than in other studies[5,16,18] this could be related to no checking of the assumption of normality. No reporting of CI for main effect size measures was lower than in other articles.[16-18] This deficiency could be due to difference in the study designs and the used statistical tests. CIs provide an alternate approach to quantifying the role of chance in research.[37]

Numerical results and *P* values given to too many (or too few) decimal places were shown in nearly one-tenth of the articles. This error was not detected in other studies.[16,17] Too many digits clutter a table and make it more difficult for the eye to brain connection to extract the relevant trends.[19] *P* = nonsignificant (NS) $P < 0.05$, $P > 0.05$, etc., instead of reporting exact *P* values was lower than in prestigious journals in other studies.[16-18]

Drawing conclusions not supported by the study data was in a number of the articles were mostly due to the conclusions based on wrong test of significance. Significance claimed without data analysis or statistical test mentioned, and missing discussion of the problem of multiple significance testing was shown in few of the reviewed articles these results varied in other studies.[5,16,17,26] The variation could be due to difference in skills of interpretations by authors, their statistical background, and ignoring the interpretation of NS results in the examined articles.

The researcher received formal training in statistics and research; a member in FM research continuous quality improvement and had experience in teaching, the assumptions of most common statistical tests and errors in FM research.

Strengths and limitations

Strengths: This is the first study about statistics in FM research (Suez Canal University-Egypt) and will provide a base for continuous quality improvement in FM research. Most of the reported statistical errors by this study provide a teaching tool in FM research education. Limitations: The reviewed articles were published in a wide range of medical journals and were not classified in this article into PubMed indexed or not, National or International Journals. Items of study design evaluation in the checklist were more specific to longitudinal studies than those listed in STROBE one, but the checklist was applicable, more comprehensive, and covers many other statistical areas. Although most of FM articles were shared publication with authors from other specialties, some journals/authors did not provide adequate author information to identify the share of statisticians.

Conclusion

The use of inferential statistical tests was reported in the majority of FM articles. Omission and inadequate documentation of the statistical methods; failure to mention test statistics in the results with only *P* values and the incorrect use of statistical tests in statistical analysis. Frequency and quality of using statistical methods in FM research articles are nearly comparable to other research articles in different disciplines. This study calls for future education interventions based on the detected statistical errors to improve the quality of statistics in FM research. Adherence to statistical guidelines and review by all professionals, editors, and journals are also recommended.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Suresh K, Thomas SV, Suresh G. Design, data analysis and sampling techniques for clinical research. Ann Indian Acad Neurol 2011;14:287-90.
2. Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AH. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. BMC Res Notes 2011;4:304.
3. Altman DG, Goodman SN, Schroter S. How statistical expertise is used in medical research. JAMA 2002;287:2817-20.
4. Arnold LD, Braganza M, Salih R, Colditz GA. Statistical trends in the journal of the American Medical Association and implications for training across the continuum of medical education. PLoS One 2013;8:e77301.
5. Sridharan K, Gowri S. Reporting quality of statistics in Indian journals: Analysis of articles over a period of two years. J Scientometric Res 2015;4:10-3. Available from: http://www.jscires.org/downloadpdf.asp?issn=2320-0057;year=2015;volume=4;issue=1;spage=10;epage=13;aulast=Sridharan; type=2. [Last accessed on 2015 Jun 15].
6. Qin N, Zhang J, Zhang W, Dai J, Chen W. Some tips about statistics on medical research. J Thorac Dis 2015;7:E177-8.
7. Curran-Everett D, Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society. J Appl Physiol 2004;97:457-9.
8. Schulz KF, Altman DG, Moher D; CONSORT Group. Consort 2010 statement: Updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010;152:726-32.
9. Fuller T, Peters J, Pearson M, Anderson R. Impact of the transparent reporting of evaluations with nonrandomized designs reporting guideline: Ten years on. Am J Public Health 2014;104:e110-7.
10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Ann Intern Med 2007;147:573-7.
11. Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: The SAMPL guidelines. In: Smart P, Maisonneuve H, Polderman A, editors. Science Editors' Handbook. European Association of Science Editors; 2013. Available from: http://www.equator-network.org/reporting-guidelines/sampl/. [Last accessed on 2015 May 03].
12. The EQUATOR Network | Enhancing the Quality and Transparency of Health Research. Available from: http://www.equatornetwork.org/. [Last accessed on 2015 Nov 24].
13. ASA's Ethical Guidelines for Statistical Practice-American. American Statistical Association; 1999. Available from: http://www.amstat.org/about/ethicalguidelines.cfm. [Last accessed on 2015 Jun 05].
14. Abdulmajeed AA, Ismail MA, Nour-Eldein H. Research publications in medical journals (1992-2013) by family medicine authors – Suez Canal University-Egypt. J Family Med Prim Care 2014;3:368-73.
15. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983;309:709-13.
16. Strasak AM, Zaman Q, Marinell G, Peffifer KP, Ulmer H. The use of statistics in medical research: A comparison of the New England journal of medicine and nature medicine. Am Stat 2007;61:47-55.
17. Strasak AM, Zaman Q, Marinell G, Peffifer KP, Ulmer H. The use of statistics in medical research: A comparison of Wiener Klinische Wochenschrift and Wiener Medizinische Wochenschrift. Austrian J Stat 2007;36:141-52. Available from: http://www.stat.tugraz.at/AJS/ausg072/072Strasak.pdf. [Last accessed on 2015 May 01].
18. Hassan S, Yellur R, Subramani P, Adiga P, Gokhale M, Iyer MS, *et al.* Research design and statistical methods in Indian medical journals: A retrospective survey. PLoS One 2015;10:e0121268.

19. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research – A review of common pitfalls. Swiss Med Wkly 2007;137:44-9.
20. Olsen CH. Review of the use of statistics in infection and immunity. Infect Immun 2003;71:6689-92.
21. Worthy G. Statistical analysis and reporting: Common errors found during peer review and how to avoid them. Swiss Med Wkly 2015;145:w14076.
22. Study Design and Choosing a Statistical Test. BMJ Online. Available from: http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/13-study-design-and-choosing-statisti. [Last accessed on 2015 May 01].
23. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. Br Med J (Clin Res Ed) 1983;286:1489-93.
24. Altman DG, Bland JM. Detecting skewness from summary information. BMJ 1996;313:1200.
25. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1991.
26. Rigby AS, Armstrong GK, Campbell MJ, Summerton N. A survey of statistics in three UK general practice journal. BMC Med Res Methodol 2004;4:28.
27. Yim KH, Nahm FS, Han KA, Park SY. Analysis of statistical methods and errors in the articles published in the Korean journal of pain. Korean J Pain 2010;23:35-41.
28. Zaman Q, Azam M, Pfeiffer KP, Strasak AM. Statistical methods and complexity of data analysis in recent surgical research. Hum Physiol 2011;35:2961-3. Available from: http://www.elixirpublishers.com. [Last accessed on 2015 Aug 01].
29. Jin Z, Yu D, Zhang L, Meng H, Lu J, Gao Q, *et al.* A retrospective survey of research design and statistical analyses in selected Chinese medical journals in 1998 and 2008. PLoS One 2010;5:e10822.
30. Pet S, Naik VD, Petal P. Use of statistical methods and complexity of data analysis in recent research publications in basic medical sciences. Natl J Community Med 2014;5:253-6. Available from: http://www.njcmindia.org. [Last accessed on 2015 Jun 07].
31. Hoffman JI. The incorrect use of Chi-square analysis for paired data. Clin Exp Immunol 1976;24:227-9.
32. Adedokun OA, Burgess WD. Analysis of paired dichotomous data: A gentle introduction to the McNemar test in SPSS. J Multidiscip Eval 2012;8:125-31. Available from: http://www.jmde.com/. [Last accessed on 2015 May 01].
33. Schatz P, Jay KA, McComb J, McLaughlin JR. Misuse of statistical tests in archives of clinical neuropsychology publications. Arch Clin Neuropsychol 2005;20:1053-9.
34. Wu S, Jin Z, Wei X, Gao Q, Lu J, Ma X, *et al.* Misuse of statistical methods in 10 leading Chinese medical journals in 1998 and 2008. Sci World J 2011;11:2106-14.
35. Hoekstra R, Kiers HA, Johnson A. Are assumptions of well-known statistical techniques checked, and why (not)? Front Psychol 2012;3:137.
36. García-Berthou E, Alcaraz C. Incongruence between test statistics and *P* values in medical papers. BMC Med Res Methodol 2004;4:13.
37. Eldridge S. Good practice in statistical reporting for family practice. Fam Pract 2007;24:93-4.

Appendix 1: Checklist for statistical evaluation of medical articles

	Assessment		
	Error committed	Unable to assess/not clear	Application correct
Design of study			
Errors and deficiencies related to randomization/blinding and selection of control groups			
1. Failure to use/report randomization (e.g., in a controlled trial/experiment)	O	O	O
2. Method of randomization/allocation to intervention not clearly stated (e.g., table of random numbers used)	O	O	O
3. Failure to report initial equality of baseline characteristics/comparability of study groups	O	O	O
4. Use of an inappropriate control group (heterogeneous, clearly not comparable material)	O	O	O
Errors and deficiencies related to the design of the study			
5. Failure to report number of participants/observations (sample size)	O	O	O
6. Failure to report possible withdrawals from the study	O	O	O
7. No a priori sample size calculation/neglect of effect-size estimation; power calculation	O	O	O
8. Inappropriate testing for equality of baseline characteristics (e.g., for initial statistical equality of groups)	O	O	O
Data analysis			
Use of a wrong or suboptimal statistical test			
9. Incompatibility of statistical test with type of data examined	O	O	O
10. Unpaired tests for paired data (e.g., repeated observations analyzed as independent data) or vice versa	O	O	O
11. Inappropriate use of parametric methods (e.g., for data that are obviously nonnormal or skewed)	O	O	O
12. Use of an inappropriate test for the hypothesis under investigation	O	O	O
Multiple testing/multiple comparisons (Type I error inflation)			
13. Failure to include a multiple-comparison correction	O	O	O
14. Inappropriate *post hoc* subgroup analysis ("shopping for statistically significant differences")	O	O	O
Special errors with Student's *t*-test			
15. Failure to test and report that test assumptions were proven and met	O	O	O
16. Unequal sample sizes for paired *t*-test	O	O	O
17. Improper multiple pair wise comparisons (without adjustment of alpha-level) of >2 groups	O	O	O
18. Use of an unpaired *t*-test for paired data or vice versa	O	O	O
Special errors with Chi-square tests			
19. No Yates-continuity correction reported if small numbers	O	O	O
20. Use of Chi-square when expected numbers in a cell are <5	O	O	O
21. No explicit statement of the statistical null-hypothesis tested	O	O	O
22. *P* values obviously wrong	O	O	O
Special errors with regression analysis			
23. No description of assumptions of the analysis (an analysis of residuals confirmed the assumptions of linearity)	O	O	O
24. No description of how any outlying values were treated in the analysis if relevant	O	O	O
25. No report of how any missing data were treated in the analyses	O	O	O
26. For either simple or multiple (multivariable) regression analyses, regression equation was not reported	O	O	O
27. For multiple regression analyses: no report of variable selection process by which the final model was developed (e.g., forward-stepwise; best subset)	O	O	O
28. No reporting of the regression coefficients (beta weights) of each explanatory variable	O	O	O
29. Measure of the model's "goodness-of-fit" to the data was not reported	O	O	O
30. No model validation procedure was given	O	O	O
31. For primary comparisons analyzed with simple linear regression analysis, results were not presented graphically	O	O	O
32. Regression line (or the interpretation of the analysis) beyond the minimum and maximum values of the data was extended in the plot	O	O	O
Documentation			
Improper description of statistical tests			
33. Failure to specify/define all applied tests clearly and correctly	O	O	O
34. Wrong names for statistical tests	O	O	O
35. Referring to unusual/obscure methods without explanation or reference	O	O	O
36. Failure to specify which test was performed on a given set of data when more than one test was done	O	O	O
37. "Where appropriate" statement	O	O	O

Contd...

Appendix 1: Contd...

	Assessment		
	Error committed	Unable to assess/not clear	Application correct
Failure to define details of a test performed			
38. Failure to state number of tails	O	O	O
39. Failure to state if test was paired or unpaired	O	O	O
40. Failure to state in advance which values of *P* indicate statistical significance	O	O	O
41. Name of the statistical software or programs used was not mentioned	O	O	O
Presentation			
Inadequate (graphical or numerical) description/presentation of basic data (location, dispersion)			
42. Mean but no indication of variability of the data (failure to describe variability)	O	O	O
43. Giving SE instead of SD to describe/summarize study data	O	O	O
44. Failure to define±notion for describing variability of the sample; unlabeled error bars	O	O	O
45. Use of arithmetic mean and SD to describe nonnormal or ordinal data	O	O	O
46. SE on undefined (or too small) sample sizes	O	O	O
Inappropriate/poor reporting of results			
47. Reporting *P* value without test statistics	O	O	O
48. Results given only as *P* values, no CIs given for main effect size measures (and in regression model for each explanatory variable)	O	O	O
49. CI given for each group rather than for the contrast	O	O	O
50. Numerical results and *P* values given to too many (or too few) decimal places (e.g., $P<0.000000$)	O	O	O
51. "P=NS," "$P<0.05$," "$P>0.05$" (or other arbitrary thresholds) instead of reporting exact *P* values	O	O	O
Interpretation			
Wrong interpretation of results			
52. "NS" treated/interpreted as "no effect/no difference"	O	O	O
53. Marginal statistical significance (e.g., $P=0.1$) treated as genuine effect	O	O	O
54. Drawing conclusions not supported by the study data	O	O	O
55. Significance claimed (or *P* values stated) without data analysis or statistical test mentioned	O	O	O
Poor interpretation of results			
56. Failure to consider CI's when interpreting "NS" differences (especially in small studies)	O	O	O
57. Disregard for Type II error when reporting NS results	O	O	O
58. Missing discussion of the problem of multiple significance testing if occurred	O	O	O

SD: Standard deviation; SE: Standard error; CI: Confidence interval; NS: Nonsignificant