Statistical Techniques in General Surgery Literature: What Do We Need to Know?

Phillip J Williams, MD, Patrick Murphy, MD, MSc, FRCSC,
Julie Ann M Van Koughnett, MD, MEd, FRCSC, FACS, Michael C Ott, MD, MEd, FRCSC,
Luc Dubois, MD, MSc, FRCSC, Laura Allen, MSc, Kelly N Vogt, MD, MSc, FRCSC

Effective critical appraisal of the scientific literature is an essential skill for all surgeons to remain up to date in evidence-based surgery. Through the critical appraisal process, a surgeon must determine the validity of a given study and its applicability to practice. In addition to an understanding of general research methodology, familiarity with commonly used statistical techniques is necessary to be able to determine whether study data were analyzed in an appropriate manner to answer the stated research question. Statistical knowledge is particularly important when interpreting the surgical literature, in which, due to numerous factors, observational studies predominate.[1]

Observational research relies heavily on the use of proper statistical techniques to address inherent limitations and biases. The type of statistical methodology chosen may have a significant impact on interpretation of the data and, ultimately, the conclusions drawn from the work.[2,3] Therefore, a working knowledge of statistical techniques is a prerequisite for critical appraisal. As surgical literature continues to grow in both magnitude and complexity, so too does the use of statistical procedures.[4,5] Previous work has described the frequency of statistical techniques used in various medical fields; however, information is limited in the surgical specialties.[4,6-12] Furthermore, surgeons and surgical residents have been shown to be lacking in their statistical knowledge, limiting their ability to interpret studies and critically evaluate the literature.[13-16] This gap in statistical literacy impairs the ability to appropriately incorporate evidence-based medicine into practice.

Residency, with its designated educational time, is an ideal opportunity for surgical trainees to become competent in interpreting statistical data. Given both the variety of subjects competing for space in surgical residency curricula, and the breadth of the field of biostatistics, it is essential that education-related statistics and epidemiologic methods be focused on relevant and high-yield information. Similar pressures exist for practicing surgeons in choosing where to focus their continuing medical education. The purpose of this study was to determine the most frequently used statistical techniques across a representative sample of recent general surgery literature, with the goal of identifying the highest-yield procedures to inform and improve the education of both trainees and surgeons within general surgery.

METHODS

Dataset

Ten representative surgical journals publishing studies related to general surgery were selected based on impact factor, author consensus, and experience (*American Surgeon, Annals of Surgery, British Journal of Surgery, Canadian Journal of Surgery, Journal of Trauma and Acute Care Surgery, Journal of the American College of Surgeons, JAMA Surgery, New England Journal of Medicine, Surgery, World Journal of Surgery*). All studies pertaining to general surgery appearing in these journals over a 12-month period (May 2015 to April 2016) were reviewed for possible inclusion.

Inclusion criteria

Original research articles pertaining to essential content areas of general surgery, as defined by the American Board of Surgery[17] (alimentary tract including bariatric surgery; abdomen and its contents; breast, skin, and soft tissue; endocrine system; solid organ transplantation; surgical oncology; and emergency general surgery), were included in the analysis. Review articles, systematic reviews, and opinion pieces were excluded before they are not primary research.

Article evaluation

Articles meeting the inclusion criteria were reviewed and classified by type of study (randomized controlled trial

Disclosure Information: Nothing to disclose.

Received June 1, 2018; Revised July 13, 2018; Accepted July 13, 2018.

From the Department of Surgery, London Health Sciences Centre (Williams, Murphy, Van Koughnett, Ott, Dubois, Allen, Vogt) and the Department of Clinical Epidemiology and Biostatistics (Dubois, Vogt), Western University, London, Ontario, Canada, and the Department of Surgery, University of Toronto, Toronto, Ontario, Canada (Williams).

Correspondence address: Kelly N Vogt, MD, MSc, FRCSC, London Health Sciences Centre, 800 Commissioners Rd, East Room E2-214, London, Ontario, Canada N6A 5W9. email: kelly.vogt@lhsc.on.ca

https://doi.org/10.1016/j.jamcollsurg.2018.07.656
ISSN 1072-7515/18

[RCT], prospective observational, or retrospective observational). Each manuscript was reviewed in its entirety, and the types and number of unique statistical procedures presented were recorded. The articles were also examined for the use of sample size calculations. The data collected were then analyzed for basic descriptive statistics using Excel (Microsoft Corporation; 2011, version 14.7.1).

RESULTS

A total of 980 articles were included in our analysis. Retrospective observational studies were the most common study type (n = 820, 83.7%), followed by prospective cohort studies (n = 93 [9.5%]) and randomized controlled trials (n = 64 [6.5%]) (Table 1). Based on the American Board of Surgery content areas, 28% of the articles were related to the abdomen and its contents, 22% to surgery of the alimentary tract, 19% to surgical oncology, 12% to endocrine surgery, 9% to emergency general surgery, 8% to breast, skin, and soft tissue, and 3% to solid organ transplantation.

The average number of unique statistical techniques used per article was 3.5 ± 1.1 (Table 1). The total number of different statistical procedures identified across all 980 articles included in this analysis was 87 (eTable 1). The most frequently used statistical techniques (appearing in more than 20% of the articles) were the chi-square test (63%), logistic regression (39%), Student's *t*-test (39%), Fisher's exact test (37%), the Mann-Whitney-U-test (34%), Kaplan-Meier time-to-event analysis (30%), log-rank test (25%), and Cox proportional hazard modeling (21%). Of the 10 most commonly identified tests, 50% were nonparametric tests. The most frequently used statistical procedures are outlined in Table 2. Sample size calculation was reported in 13% of all articles, but in 41% of RCTs and prospective cohort studies (Table 1).

DISCUSSION

Our study identifies the most commonly used statistical procedures in the current general surgical literature. The majority of these procedures can be considered to represent standard statistical techniques or concepts that should be feasible to teach surgical residents and practicing surgeons who have not undertaken advanced statistical training. Our results suggest the majority of surgical literature is retrospective and, therefore, an understanding of regression and other methods to account for bias in observational trials is paramount to understanding and applying results in clinical practice. Furthermore, our results have highlighted areas of improvement in surgical literature, such as reporting sample size and other related power calculations.

Statistical procedures identified in this article included a wide range of parametric and nonparametric approaches, as well as a number of regression models and methods to assess goodness-of-model fit. Several correction or adjustment methods, tests of normality, imputation methods, and other modeling approaches were also identified. It should also be noted that many of the procedures and statistics identified are preformed and interpreted within the context of other tests or models listed. The most common procedures identified in this study represent a starting point for the education of surgical residents, which can become a basis for the understanding of more complex procedures.

An understanding of specific statistical techniques is essential for effective critical analysis of the scientific

Table 1. Summary of Data Points Used in Analysis, Each Journal Represents 12-Month Period (May 2015—April 2016)

Journal (IF)	Articles	RCT	Prospective	Retrospective	Statistical tests named per article	Sample size calculation, all, %	Sample size calculation, RCT + prospective, %
BJS (5.9)	106	9	16	81	3.9	12	52
JACS (4.3)	127	11	13	103	3.7	5	29
NEJM (72.4)	4	3	1	0	4.7	75	75
CJS (1.9)	16	3	0	13	1.4	13	67
JTACS (3.4)	12	1	1	10	2.1	0	0
Ann Surg (8.9)	162	22	18	122	4.3	14	58
Am Surg (2.6)	114	0	8	106	2.4	2	25
JAMA Surg (8.0)	76	3	5	68	3.3	4	38
Surg (3.9)	173	6	17	150	4.1	3	26
WJS (2.6)	190	6	14	167	3.5	4	40
TOTAL	980	64	93	820	3.3	13	41

Am Surg, *American Surgeon*; Ann Surg, *Annals of Surgery*; BJS, *British Journal of Surgery*; CJS, *Canadian Journal of Surgery*; IF, impact factor; JACS, *Journal of the American College of Surgeons*; JAMA Surg, *JAMA Surgery*; JTACS, *Journal of Trauma and Acute Care Surgery*; NEJM, *New England Journal of Medicine*; RCT, randomized control trial; Surg, *Surgery*; WJS, *World Journal of Surgery*.

Table 2. Most Frequently Appearing Statistical Procedures by Absolute Number and Relative Proportion

Procedure	Application	Used in observational studies	Used in randomized studies	No. of articles	Proportion of articles, %
Chi-square test	Nonparametric; used to compare frequencies between groups at baseline (ie percent male)	Yes	Yes	613	63
Logistic regression	Model; used to determine odds of an event occurring, controlling for confounders (ie odds of death); often used in the analysis of observational studies	Often	Occasionally for secondary analyses	380	39
Student's *t*-test	Parametric; used to compare group means at baseline (ie age); used for analysis of normally distributed data	Yes	Yes	378	39
Fisher's exact test	Nonparametric; used to compare frequencies between groups at baseline; often used as an alternative to the chi-square test when frequencies are small	Yes	Yes	362	37
Mann-Whitney U test	Nonparametric; used to compare population distributions between groups at baseline using medians (ie length of stay); used for analysis of non-normally distributed data	Yes	Yes	334	34
Kaplan-Meier curve	Nonparametric; used to summarize time-to-event data (ie 5-year disease-free survival)	Yes	Yes	291	30
Log-rank test	Nonparametric; used to compare survival distributions between 2 samples (ie time to event)	Yes	Yes	240	25
Cox proportional Hazard model	Model; used to summarize time-to-Event data, controlling for covariates (ie assess the effect of risk factors on disease-free survival time)	Often	Occasionally for secondary analyses	208	21
Analysis of variance	Parametric; used to compare means at baseline (ie age) when there are >2 groups; used for analysis of normally distributed data	Yes	Yes	94	10
C-statistic	Goodness of fit in multivariable analysis; used to assess the predictive accuracy of a logistic regression model	Often	Occasionally for secondary analyses	88	9

Procedures are listed in descending order of frequency.

literature. Evidence shows there is a deficiency in the statistical literacy of both surgical trainees and practicing surgeons.[13-16] This gap in knowledge could be addressed with formal teaching during residency and as a part of continuing medical education initiatives. In order to optimize this educational process, the material taught should be focused on the most relevant and high-yield techniques identified in this study. In addition to an understanding of the indications for parametric vs nonparametric tests and the role of multivariable analysis, working knowledge of these commonly used techniques would allow one to better interpret the statistical methodology of a large proportion of general surgery literature. These analytic procedures represent information that is both high-yield and practical, making them ideal topics to focus on when designing the educational curriculum of general surgery residency programs or continuing medical education for practicing surgeons.

Although there is limited literature regarding the use of statistics specific to general surgery, the results of this study are consistent with previous work in other specialties that demonstrates there is a small group of commonly appearing procedures that account for a large proportion of all techniques used.[7,9,12,16] Although the complexity and magnitude of the use of statistics in general surgery literature has increased over time, there is significant overlap between the commonly occurring procedures identified in this study and in previous work.[5,7,9,12] These techniques are considered by many to represent standard procedures that commonly appear in introductory statistical texts, and should be conceptually feasible to both teach and learn. Our study was not specifically designed to judge the quality or appropriateness of the statistical techniques used; however, deficiencies were noticed, as highlighted in the summary of articles (Table 1). The ubiquity of retrospective analyses (90% of the literature) is not unexpected and is certainly an improvement from the use of case-series in early surgical literature; however, one should use caution when adapting clinical practice based on retrospective studies.[18] Randomized control trials often fail to confirm the results of historical studies with respect to the efficacy of new interventions.[19] Surgical literature, particularly in North America, is lacking in randomized control trials and high level evidence.[20] Ideally, retrospective studies should inform future prospective studies and randomized trials (if feasible). Surgical literature could be improved through the use of standardized reporting as well as reporting sample size and power calculations to inform readers regarding the strength of the analysis.[21]

Strengths of our analysis include the large number of articles and breadth of journals included and the contemporary nature of the literature studied. Previous work similar to our study was last published in 1987[16] and, given that the nature of general surgery publications have changed significantly over the last 30 years, this study fills a significant knowledge gap on the use of statistics in contemporary general surgery literature.[4,5]

Our study has several limitations. We did not assess the appropriateness of the statistical tests identified, which has been previously identified as a common source of error within medical literature, nor did we assess the quality of the literature examined.[1,22] Furthermore, we chose 10 surgical journals based on expert consensus, and it is possible we did not include a more relevant source of surgical literature; however, the impact factors of all included journals was high. Additionally, although we report on the number of unique statistical tests used, we acknowledge that these are only the tests reported in the article, and additional procedures may have been applied in the full analysis of the study data. Finally, we specifically excluded summary articles due to the lack of statistical techniques (consensus guidelines, systematic reviews, editorials). Meta-analyses are also summary articles and therefore excluded. Although well-conducted meta-analyses of randomized controlled trials are the highest level of evidence and serve to collate the literature on a specific topic, their interpretation is more straightforward and outlined well elsewhere.[23]

CONCLUSIONS

Given the significant volume and complexity of information that general surgeons are required to master in the course of their education, it is essential that the information they are taught is practical and high-yield. Through the analysis of a large representative body of contemporary general surgery literature, our study identifies the individual statistical procedures that best fulfill this need. This information can be used to more optimally design the educational curriculum of general surgery residency programs and continuing medical education initiatives for practicing surgeons.

Author Contributions

Study conception and design: Williams, Murphy, Van Koughnett, Ott, Dubois, Vogt
Acquisition of data: Williams, Murphy
Analysis and interpretation of data: Williams, Murphy, Van Koughnett, Ott, Dubois, Allen, Vogt
Drafting of manuscript: Williams, Murphy, Allen, Vogt
Critical revision: Williams, Murphy, Van Koughnett, Ott, Dubois, Allen, Vogt

REFERENCES

1. Wu R, Glen P, Ramsay T, et al. Reporting quality of statistical methods in surgical observational studies: protocol for systematic review. Syst Rev 2014;3:70.
2. MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess 2000;4:1–154.
3. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185–1190.
4. Kurichi JE, Sonnad SS. Statistical methods in the surgical literature. J Am Coll Surg 2006;202:476–484.
5. Wells CI, Robertson JP, O'Grady G, et al. Trends in publication of general surgical research in New Zealand, 1996–2015. ANZ J Surg 2017;87:76–79.
6. Avram MJ, Shanks CA, Dykes MH, et al. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg 1985;64:607–611.
7. Hokanson JA, Luttman DJ, Weiss GB. Frequency and diversity of use of statistical techniques in oncology journals. Cancer Treat Rep 1986;70:589–594.
8. Goldin J, Zhu W, Sayre JW. A review of the statistical analysis used in papers published in Clinical Radiology and British Journal of Radiology. Clin Radiol 1996;51:47–50.
9. Cardiel MH, Goldsmith CH. Type of statistical techniques in rheumatology and internal medicine journals. Rev Invest Clin 1995;47:197–201.
10. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983;309:709–713.
11. Oliver D, Hall JC. Usage of statistics in the surgical literature and the 'orphan P' phenomenon. Aust NZ J Surg 1989;59:449–451.
12. Hokanson JA, Stiernberg CM, McCracken MS, et al. The reporting of statistical techniques in otolaryngology journals. Arch Otolaryngol Head Neck Surg 1987;113:45–50.
13. Guller U, Oertli D. Sample size matters: a guide for surgeons. World J Surg 2005;29:601–605.
14. Ridgway PF, Guller U. Interpreting study designs in surgical research: a practical guide for surgeons and surgical residents. J Am Coll Surg 2009;208:635–645.
15. Anderson BL, Williams S, Schulkin J. Statistical literacy of obstetrics-gynecology residents. J Grad Med Educ 2013;5:272–275.
16. Reznick RK, Dawson-Saunders E, Folse JR. A rationale for the teaching of statistics to surgical residents. Surgery 1987;101:611–617.
17. American Board of Surgery. Specialty of General Surgery Defined 2016. Available at: http://www.absurgery.org/default.jsp?aboutsurgerydefined. Accessed January 2017.
18. Centre for Evidence Based Medicine. Centre for Evidence Based Medicine 2017. Available at: http://www.cebm.net. Accessed December 17, 2017.
19. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. Am J Med 1982;72:233–240.
20. Forrester JA, Forrester JD, Wren SM. Trends in country-specific surgical randomized clinical trial publications. JAMA Surg 2018;153:386–388.
21. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg 2014;12:1495–1499.
22. Thiese MS, Ronna B, Robbins RB. Misuse of statistics in surgical literature. J Thorac Dis 2016;8:E726–E730.
23. Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011] The Cochrane Collaboration 2011. Available at: www.handbook.cochrane.org. Accessed August 16, 2018.

eTable 1. Complete List of All Statistical Procedures Identified from the Articles Included in the Analysis

1	Kaplan-Meier statistic
2	Log-rank test
3	Chi-square test
4	Fisher's exact test
5	Student's *t*-tests
6	Logistic regression
7	Linear regression
8	Cox proportion hazards model
9	Mann-Whitney U test
10	Gray's test (see also Fine-Gray regression)
11	Sensitivity analysis
12	c-statistic
13	Poisson regression
14	Wilcoxon sign rank test
15	Shapiro-Wilk test
16	Spearman rank correlation coefficient
17	Yates' correction
18	Pearson's coefficient
19	Kruskal-Wallis test
20	Binomial test
21	Hosmer-Lemeshow test
22	Jonckheere-Terpstra test
23	Analysis of variance
24	Tukey's test
25	Cochran-Armitage trend test
26	Bonferonni correction
27	Mood's test
28	Kolmogorov-Smirnov test
29	Efron's pseudo R^2
30	Rosenbaum's test
31	Wald test
32	Linear-by-linear association
33	Restricted cubic regression splines
34	Sidak correction
35	Friedman test
36	Brier score
37	Royston-Parmar
38	Akaike information criterion
39	Quantile regression
40	Cuzick test
41	McNemar's test
42	Stuart-Maxwell test
43	Fisher Z transformation
44	Cox-Hinkley-Miettinen-Nurminen method
45	Cohen's kappa statistic
46	Propensity score
47	Cox-Snell R^2
48	Bootstrapping
49	Multivariate imputation by chained equations
50	Markov chain Monte Carlo
51	Little's MCAR test
52	Firth's regression
53	Durbin-Watson test
54	Cumulative sum statistic
55	Nagelkerke R^2
56	Difference-in-difference estimation
57	Breslow test
58	Cochran-Mantel-Haenszel test
59	Generalized linear model
60	Generalized estimating equation
61	Cronbach's alpha
62	Aalen's additive regression
63	Lasso regression
64	Schoenfeld residuals test
65	Mann-Kendall test
66	Bursac-Hosmer selection
67	Fine-Gray regression (see also Gray's test)
68	Youden index
69	Sign test
70	Benjamini-Hochberg correction
71	Cochran's Q test
72	Kendall's tau coefficient
73	Fleming-Harrington test
74	Levene's test
75	Cook's distance
76	Tamhane's test
77	Scheffe's test
78	Wilson score
79	Dunnett's test
80	Games-Howell test
81	Random Survival Forest
82	Dunn's test
83	Park test
84	Hotelling's T-square test
85	Cohen's d
86	LOESS curve fitting
87	D'Agostino-Pearson test

See Appendix 1 for citations of all 980 articles included in this analysis.