Received: 24 March 2021 | Revised: 20 April 2021 | Accepted: 23 April 2021

DOI: 10.1111/1747-0080.12678

ORIGINAL RESEARCH

Nutrition & Dietetics Journal of Dietitians Australia WILEY

Statistical methods and software used in nutrition and dietetics research: A review of the published literature using text mining

Alison Coenen MNutrDiet, APD[1] | Marijka J. Batterham PhD, AdvAPD[2] | Eleanor J. Beck PhD, FDA[1]

[1]School of Medicine, Faculty of Science, Medicine and Health, University of Wollongong, Wollongong, New South Wales, Australia

[2]School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, New South Wales, Australia

Correspondence
Marijka J. Batterham, School of Mathematics and Applied Statistics, University of Wollongong, NSW, 2522, Australia.
Email: marijka@uow.edu.au

Abstract

Aim: Dietitians must be statistically literate to effectively interpret the scientific literature underpinning the discipline. Despite this, no study has been conducted that objectively identifies common statistical methods and packages specific to current nutrition and dietetics literature. This study aimed to identify statistical methods and software frequently used in nutrition and dietetics research.

Methods: A text mining approach using the bag-of-words method was applied to a random sample of articles obtained from all journals in the 'Nutrition and Dietetics' subject category within the SCImago Journal and Country Rank portal and published in 2018. A list of 229 statistical terms and 19 statistical software packages was developed to define the search terms to be mined. Statistical information from the methods section of included articles was extracted into Microsoft Excel (2016) for data cleaning. Statistical analyses were conducted in R (Version 3.6.0) and Microsoft Excel (2016).

Results: Seven hundred and fifty-seven journal articles were included. Numerical descriptive statistics were the most common statistical method group, appearing in 83.2% of articles (n = 630). This was followed by specific hypothesis tests (68.8%, n = 521), general hypothesis concepts (58.4%, n = 442), regression (44.4%, n = 336), and ANOVA (30.8%, n = 233). IBM SPSS statistics was the most common statistical software package, reported in 41.7% of included articles.

Conclusion: These findings provide useful information for educators to evaluate current statistics curricula and develop short courses for continuing education. They may also act as a starting point for dietitians to educate themselves on typical statistical methods they may encounter.

KEYWORDS

data mining, dietetics, education, evidence-based practice, nutritional sciences, statistics

1 | INTRODUCTION

Research provides the evidence base that guides clinical practice. It informs decision making and ensures the delivery of optimal nutrition care to individuals and the broader community.[1,2] In the context of nutrition, organisations such as the International Confederation of Dietetic Associations recognise the importance of research in International Competency Standards for Dietitian—Nutritionists, where evidence-based practice and application of research is a minimum requirement for entrance into the profession.[3]

To ensure evidence-based nutrition practice, ongoing consultation and critical appraisal of the literature is required.[2,4] Research is fundamental to the application of nutrition science however involvement is inadequate among dietitians and nutritionists who primarily practise outside of research settings.[5-7] Common barriers include a perceived lack of research methodology skills, as well as a lack of time, funding, and administrative support.[7-10] Notwithstanding several efforts to address these obstacles,[9,11,12] little improvement has been documented[13] and few studies have sought to foster the development of specific research skills required for dietitians.

As statistical methods often underpin research outcomes,[14] dietitians must be able to interpret and critique scientific data in any literature they are reviewing. Furthermore, those participating in higher-level research activities must understand statistics to successfully produce, analyse and disseminate findings of their own research.[15] Despite this, not all Australian dietetics education programs include a named statistics subject, although some may include small elements in research theory content preceeding research project subjects.

A need for continuing education programs that focus on research skill development has been identified.[10] For example, dietitians have expressed a desire to participate in relevant statistics courses beyond current entry-level requirements.[6,13] Literature identifying statistical methods and software packages used in medical and public health research exists.[14-16] However, no published study to date has reviewed those commonly used in nutrition and dietetics research literature.

This study aimed to identify statistical methods and software frequently used in current nutrition and dietetics research. It applied a modern technological approach in the form of text mining to derive information from a large collection of journal articles. This provides valuable, objective information to guide statistics curricula and continuing education for dietitians.

2 | METHODS

A text mining design based on the bag-of-words method[17] was used to review the frequency of statistical methods and packages reported in a random sample of nutrition-related journal articles published in 2018. This method counts the frequency of individual terms (unigrams) in a corpus (collection of text documents) by removing the structure of words and representing data as a multiset ('bag'), so that multiplicity is retained. The process of text mining here included the following steps: corpus selection, manual data extraction, generation of search terms, data cleaning and concatenation of search terms, quality assurance and statistical analysis to obtain results. As the study was restricted to published literature, ethical approval was not required.

All journals contained within the subject category of 'Nutrition and Dietetics' in SCImago Journal & Country Rank portal[18] at the time of review (April 2019) were considered for inclusion. This portal provides an expansive list of discipline-specific papers as it contains all journals found in the Scopus database. To maximise generalisability across countries and subdisciplines of nutrition and dietetics, all journals within the subject category were eligible unless written in a language other than English or categorised as a book series.

For each journal, one issue published between January and December 2018 was randomly selected using the RANDBETWEEN function in Microsoft Excel (2016). All articles identified within each selected journal issue were manually reviewed by the primary researcher against the following inclusion criteria: (a) human subjects research, (b) available in full text, (c) full text written in English and (d) contained statistical analyses that were described within the methods section. All in vitro, cadaveric and animal model studies were excluded as they are not typically undertaken by dietitians in practice. Titles and abstracts were initially screened, and full texts of all potentially relevant articles were assessed to determine eligibility. Any queries were resolved by consensus with the research team (all authors).

As the aim of the study was to identify frequently used statistical methods, the sample size calculation was based on detecting words expressed in 50% of the sample. After obtaining all eligible articles, 748 were required to detect a proportion of 50% with a 2% error and a 95% confidence interval. To distribute this between the journals and prevent over-representation from any one journal, a maximum of 20 eligible articles were included from each journal. Articles were randomly selected using RANDBETWEEN when this limit was exceeded. All identified articles were managed in Microsoft Excel (2016) and included articles were imported into EndNote X8.

All included journal articles were saved as PDF files and converted into plain text files using Adobe Acrobat Pro DC (2017) to facilitate optical character recognition. Relevant data were manually extracted and entered in a Microsoft Excel (2016) spreadsheet article by one

researcher to create a central database where each row represented the data extracted from one journal article. To enhance the accuracy of identifying data of interest, only information that specifically described statistical analyses within the methods sections was extracted.

Text mining techniques were piloted in R (Version 3.6.0)[19] by the primary researcher using 15% of the extracted data. As many statistical methods are multi-word terms or have various synonymous phrases, a framework containing a pre-specified list of unigrams was used to mine terms of interest. A list of statistical methods, which was developed using a Delphi panel to identify methods used in medical research was used as a starting point.[14] The list was reviewed by all members of the research team. Modifications were made by a biostatistician with expertise in nutrition and dietetics to build on the list of synonyms and include statistical terms relating to meta-analyses. To capture multi-word terms and synonyms, each specific statistical method and its synonyms corresponded with a unigram that was created by concatenating the specific statistical methods. These unigrams were to be used to mine the terms of interest in the database. The final list included 229 statistical terms, which could be mapped to 16 statistical method groups, and 19 statistical packages. The framework of terms to be mined, including identified synonyms, is found in Supplementary File 1.

Data cleaning was performed in Microsoft Excel (2016). The 'find and replace' tool was used to tokenise multi-word statistical phrases to unigrams. All statistical methods that directly corresponded with the unigrams specified in the framework were concatenated (eg, Mann-Whitney *U* test became Mann-Whitney *U* test). Synonymous phrases were searched for (eg, Wilcoxon rank-sum test, another name for the Mann-Whitney *U* test) and replaced with the appropriate unigram (eg, Mann-Whitney *U* test). Misspellings were also corrected when encountered.

Once data cleaning was complete, a source data verification audit was conducted for assurance of data quality[20] and to identify any statistical terms of interest not yet contained in the framework. This involved the senior investigator conducting a manual verification check on a 10% random sample of the database against the original records. Error rate was less than 5%, all identified errors were amended, and the final database was saved as a csv file for analysis.

All statistical analyses were conducted by the primary researcher within R (Version 3.6.0) and Microsoft Excel (2016). Data were pre-processed in R using the 'tm', 'readr' and 'qdap' packages. Pre-processing involved transforming all text to lowercase, removing all numbers and punctuation, replacing multiple whitespace characters with a single blank, removing English stop words and applying stemming algorithms to transform terms to their roots (eg, 'multilevelmodeling' and 'multilevelmodels' were reduced to 'multilevelmodel'). A further step to count the number of articles reporting statistical method groups was performed by replacing all specific statistical methods with the corresponding statistical method group. The pre-processed data were exported as Microsoft Excel files and the COUNTIF function was used to tally the total number of articles that reported each statistical method and statistical method group. To ensure only the exact words were counted, spaces were added before and after each unigram. A word cloud was produced using the 'wordcloud' package in R to visually present the most commonly reported inferential statistics within the corpus. The R codes to conduct the analyses are available in Supplementary File 2.

3 | RESULTS

A total of 124 journals were identified within the 'Nutrition and Dietetics' subject category of the SCImago Journal & Country Rank portal. Random selection of one issue within each eligible journal identified 2044 articles. Title and abstract screening removed 841 articles and 1203 full-text articles were assessed for eligibility. A total of 1085 articles were eligible. The final sample included 757 journal articles from 72 journals (Figure 1).

Twenty-three journals provided a large proportion of articles, together representing 49% of the corpus (Table 1). The majority of these 23 journals contributed 20 articles, as this was the upper limit per issue to prevent over-representation from any one journal. A large proportion (46.2%) of articles were published in journals that had a SCImago journal rank indicator within the first quartile, while 38.0% of articles were from journals in the second quartile, 12.8% in the third quartile and only 2.9% in the fourth quartile. The complete list of articles and their corresponding journals are found in Supplementary File 3.

The source data verification audit identified an error rate of 1.3%. This indicates a high level of data accuracy and falls within the acceptable limit of 5% or below, as set by Houston et al.[20]

Numerical descriptive statistics were the most commonly used statistical method group and were reported in majority of articles (83.2%) (Table 2). The most frequently reported numerical descriptive statistic was mean (60.6%), followed by SD (37.1%) percent (23.2%), median (20.7%) and frequency (17.2%).

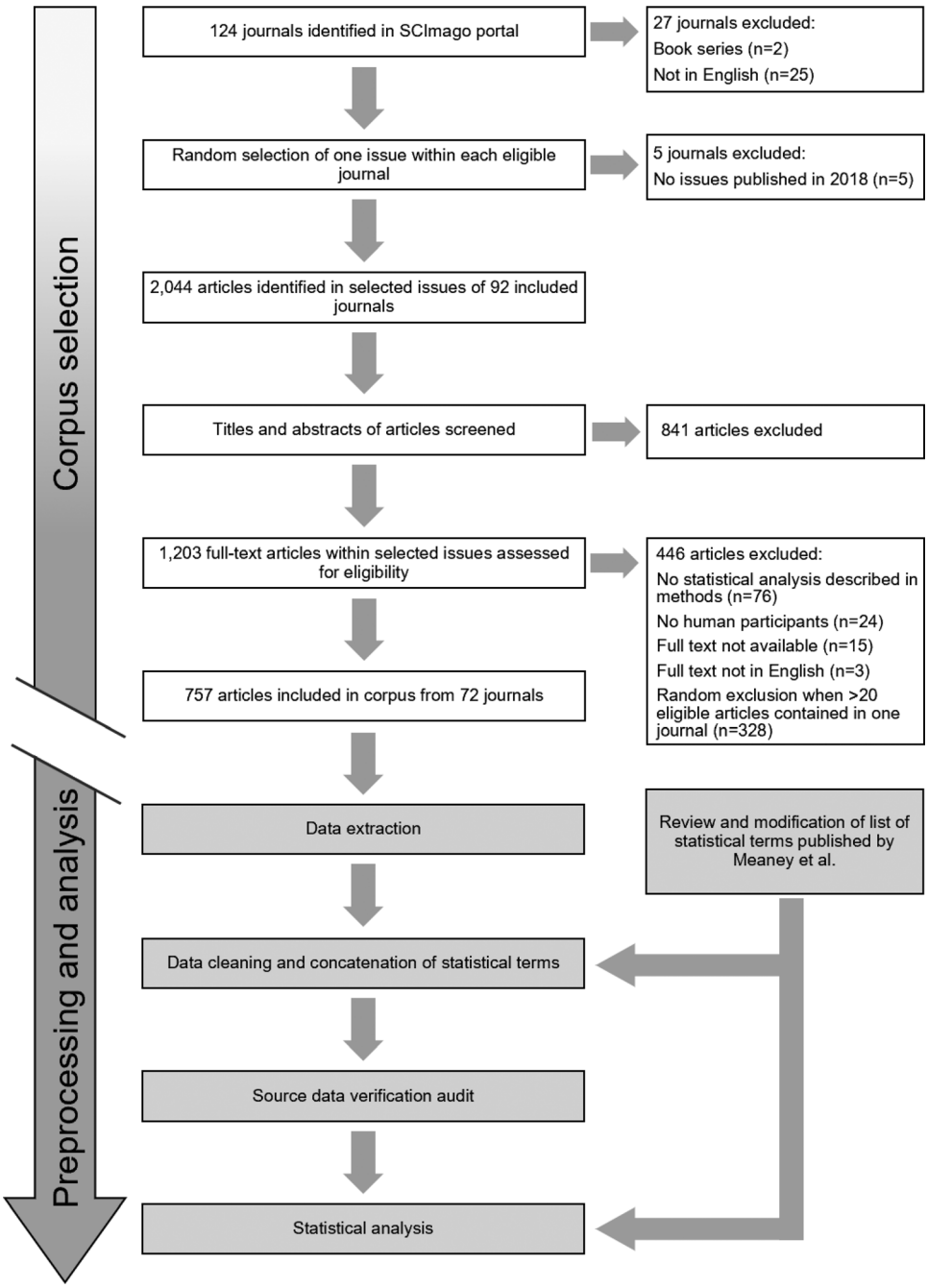

FIGURE 1 Flow chart of study procedures

Graphical descriptive statistics were not reported as often, appearing in 7.0% of articles. The second most frequently encountered statistical methods group was specific hypothesis tests, reported in 68.8% of articles. This was followed by general hypothesis tests (58.4%), regression models (44.4%) and ANOVA models (30.8%). Other common methods were epidemiological measures of risk and effect, epidemiological concepts of classification, and correlated data models, which appeared in 28.0%, 16.1% and 14.7% of articles, respectively. Within these groups were specific statistical methods and various terminology was found to be used by authors to describe individual methods (Table 3).

SPSS was cited as the most commonly utilised package and appeared in 41.7% of the reviewed articles (Table 2). This was followed by SAS (17.2%) and STATA (15.9%). Other packages such as R and Microsoft Excel were less common, yet still used in 7.1% and 4.5% of articles, respectively (Table 2).

P-values (29.7%) and confidence intervals (24.3%) were the most commonly used measures of significance. The chi-square test was the most commonly reported hypothesis test (29.7%), followed by the independent samples *t*-test (22.3%), Mann-Whitney *U* test (15.1%), and the paired samples *t*-test (11.2%). Other common hypothesis tests included Pearson correlation, Spearman correlation, Shapiro-Wilk test for normality, Fisher's exact test, Kolmogorov-Smirnov test, Kruskal-Wallis test and the Wilcoxon signed-rank test. Nutrition professionals are also

TABLE 1 Characteristics of journals contributing ≥2% of the corpus (n = 23/124)

Characteristics	SCImago journal ranking[a]	SCImago journal quartile ranking[a]	Count of articles (%)
American Journal of Clinical Nutrition	2	Q1	20 (2.64)
International Journal of Obesity	3	Q1	20 (2.64)
International Journal of Behavioral Nutrition and Physical Activity	4	Q1	20 (2.64)
Clinical Nutrition	10	Q1	20 (2.64)
Journal of Clinical Lipidology	14	Q1	20 (2.64)
Maternal and Child Nutrition	17	Q1	20 (2.64)
Nutrients	18	Q1	20 (2.64)
Nutrition and Metabolism	20	Q1	20 (2.64)
Nutrition Journal	25	Q1	20 (2.64)
Appetite	26	Q1	20 (2.64)
Obesity Surgery	28	Q1	20 (2.64)
European Journal of Clinical Nutrition	33	Q2	20 (2.64)
Food Quality and Preference	35	Q2	20 (2.64)
Food and Nutrition Research	48	Q2	20 (2.64)
Journal of the International Society of Sports Nutrition	50	Q2	20 (2.64)
Asia Pacific Journal of Clinical Nutrition	55	Q2	20 (2.64)
Journal of Eating Disorders	56	Q2	20 (2.64)
Journal of Nutrition and Metabolism	59	Q2	20 (2.64)
European Journal of Nutrition	27	Q1	19 (2.51)
Clinical Nutrition ESPEN	85	Q3	19 (2.51)
Progress in Nutrition	90	Q3	19 (2.51)
Obesity	9	Q1	16 (2.11)
Nutrition	31	Q1	16 (2.11)

[a]Journal rankings were obtained from SCImago Journal & Country Rank portal in April 2019.

likely to frequently encounter the following ANOVA models: ANOVA (20.9%), RMANOVA (7.1%) and ANCOVA (6.6%), as well as logistic regression (20.2%) and linear regression models (15.9%). The most frequent epidemiological statistics were sensitivity (10.6%), odds ratio (10.2%) and prevalence (6.1%). Effect size was reported in 8.9% of the corpus. Multilevel models (8.5%) were the only advanced statistical model that appeared in more than 5% of the corpus. Inferential tests that were observed in 10 or more articles within the corpus are visually depicted in a word cloud (Figure 2). The complete counts of all statistical terms mined can be found in Supplementary File 4.

4 | DISCUSSION

Nutrition professionals require statistical literacy skills to effectively interpret the scientific literature that underpins the discipline. This is the first study that comprehensively reviews the literature to identify statistical methods and packages commonly used in nutrition and dietetics research.

As hypothesised, numerical descriptive statistics were observed in most articles reviewed. This is in agreeance with the findings of previous studies in other health-related disciplines,[14,16,21] with some studies reporting their appearance in almost all articles. Our study also agrees with previous studies in medical research,[14] where hypothesis tests, regression and ANOVA models were the most commonly reported types of inferential statistics.

Classical statistical techniques, such as the chi-square test and *t*-tests, appeared most frequently in the sample of articles. These methods are typically taught throughout introductory and intermediate statistics[16,22] which indicates that they may be the most important for

TABLE 2 Frequency of articles reporting use of statistical methods groups and packages (n = 757)

	Count of articles (%)	Examples of common terms in each statistical method group
Statistical methods groups		
Numerical descriptive statistics	630 (83.2)	Mean, SD, percent
Hypothesis test (specific tests)	521 (68.8)	Chi-square test, *t*-tests
Hypothesis test (general concepts)	442 (58.4)	*P*-value, confidence interval
Regression	336 (44.4)	Logistic, linear regression
ANOVA	233 (30.8)	ANOVA, ANCOVA, RMANOVA
Epidemiology (risk estimation)	212 (28.0)	Odds ratio, effect size, prevalence
Epidemiology (classification and diagnostic accuracy)	122 (16.1)	Sensitivity, ROC curve, likelihood ratio
Correlated data analysis	111 (14.7)	Multilevel model, LMM, GEE
Missing data	62 (8.2)	Missing data, multiple imputation
Graphical descriptive statistics	53 (7.0)	Funnel plot, Q-Q plot, histogram
Multivariate statistics	37 (4.9)	Cronbach α, PCA, factor analysis
Survival analysis	34 (4.5)	Cox regression, Kaplan-Meier
Causal inference	20 (2.6)	Structural equation model
Computation	20 (2.6)	Bootstrap, resampling
Machine learning	17 (2.2)	Splines, discriminant analysis
Time series	3 (0.4)	Autocorrelation
Statistical packages		
SPSS	316 (41.7)	
SAS	130 (17.2)	
STATA	120 (15.9)	
R	54 (7.1)	
Microsoft Excel	34 (4.5)	
GraphPad	25 (3.3)	
Statistica	10 (1.3)	
Other[a]	31 (4.1)	

Abbreviations: ANCOVA, analysis of covariance; GEE, generalized estimating equations; LMM, linear mixed model; PCA, principal component analysis; RMANOVA, repeated-measures ANOVA; ROC, receiver operating characteristic; Q-Q plot, quantile-quantile plot.

[a]Other software used in more than one article included RevMan (n = 7), Epidata (n = 6), GPower (n = 6), Systat (n = 4), MATLAB (n = 3), MINITAB (n = 3) and Python (n = 2).

nutrition researchers to understand, and most relevant for inclusion in dietetics education programs. An interesting finding related to the numerous variations in describing common inferential tests. It may be that the slight variation in terminology adds to the confusion of the reader. A finding from our research is that using recognised descriptions of statistical terms may be helpful, but also that researchers may need to be familiar with common synonyms used to refer to frequently used statistical methods (Table 3). Perhaps part of teaching statistics needs to be a recognition of the variation in 'labels' and that researchers can access simple lists of synonyms.

Despite the increased use of more complex statistical methods by researchers,[14,15] it was surprising that a relatively low occurrence of advanced statistical techniques was observed. Multilevel modelling was the most frequently encountered advanced statistical method and only appeared in 8.5% of the articles. Similar findings were reported in a smaller study that investigated statistical methods used in public health research.[16] The low occurrence of advanced methods could suggest their lack of importance or relevance in nutrition and dietetics literature, as many nutrition research questions can be answered using simple statistical techniques.[23] However, it has also been postulated that this may be due to the historic lack of training in these

TABLE 3 Common synonyms for frequently reported statistical methods in nutrition and dietetics research

Statistical method	Synonyms	Common Use
ANOVA models		
ANCOVA	Analysis of covariance	Comparing means of multiple groups while adjusting for covariates
ANOVA	Analysis of variance	Comparing means of multiple groups
RMANOVA	Repeated measures analysis of variance	Compare means of one or more variables at multiple time points
Correlated data analysis		
Multilevel model	Random effect model, random parameter model, random coefficient model, random intercept model, hierarchical linear model, hierarchical model, linear mixed-effects model, nested data model	Generally used for clustered or grouped data for example considering patients grouped/nested within a hospital
Intraclass correlation coefficient	Intracluster correlation coefficient	
Epidemiology (classification and diagnostic accuracy)		
ROC curve	Receiver operating characteristic curve	Visualising sensitivity and specificity
Relative risk	Risk ratio	Compares probabilities of outcomes in exposed and unexposed groups
Hypothesis tests		
Chi-square test	χ^2 test	Comparing proportions
Independent samples *t*-test	Independent *t*-test, Student's *t*-test, unpaired *t*-test, unpaired Student's *t*-test, independent measures *t*-test independent two-sample *t*-test, two-sample *t*-test	Comparing two means - parametric
Mann-Whitney *U* test	Mann-Whitney Wilcoxon test, Wilcoxon rank-sum test, rank-sum test, Mann-Whitney nonparametric test	Comparing two means - nonparametric
Paired samples *t*-test	Paired *t*-test, dependent *t*-test, repeated measures *t*-test, paired Student's *t*-test, related *t*-test	Comparing means of two measures on the same subject/sample
Fisher's exact test	Fisher's test	Comparing proportions when there are low cell counts
Bonferroni correction	Bonferroni adjustment, Bonferroni post-hoc test, Bonferroni method	Adjusting for multiple comparisons
Tukey test	Tukey-Kramer test, Tukey's range test, Tukey's post-hoc test, Tukey's adjustment, Tukey correction, Tukey multiple comparison test, Tukey-Kramer adjustment, Tukey's HSD (honestly significant difference) test	Adjusting for multiple comparisons
Kruskal-Wallis test	Kruskal-Wallis H nonparametric test, Kruskal-Wallis H test, One-way ANOVA on ranks	Comparing means of multiple groups-nonparametric
Wilcoxon signed Rank Test	Signed rank test, Wilcoxon matched-pairs test, Wilcoxon matched-pairs signed-rank test	Comparing means of two measures on the same subject/sample—nonparametric
Egger's test	Egger's asymmetry test, Egger's regression test	Assessing publication bias in meta-analysis

(Continues)

TABLE 3 (Continued)

Statistical method	Synonyms	Common Use
Cohen's kappa	Cohens kappa, Cohen's k, kappa statistics, kappa correlation, κ kappa coefficient	Inter-rater reliability
Mantel-Haenszel test	Mantel-Haenszel χ^2 test, Cochran-Mantel-Haenszel test	Analysis of stratified 2 by 2 tables
Regression models		
Logistic regression	Logit regression, logit model	Predicting a binary outcome
Survival analysis		
Cox regression	Proportional hazards regression, Cox proportional hazards regression	Modelling variables predicting time to an event

Abbreviations: ANCOVA, analysis of covariance; RMANOVA, repeated-measures ANOVA; ROC, receiver operating characteristic.

FIGURE 2 A word cloud to provide a visual representation of the inferential tests most frequently observed in the corpus, where larger font sizes represent words that appeared more often than words with smaller font size

methods.[16] Technological advances and the ready availability of open-source software packages such as 'R'[19] in the past 20 years have enabled researchers to apply advanced modelling and computationally intensive statistical methods to data analysis. Continuing education opportunities in advanced statistical methods such as these would ideally be available to assist and upskill nutrition researchers.

SPSS (IBM SPSS Statistics, IBM Corp., Armonk, NY), SAS (SAS Institute Inc., Cary, NJ) and STATA (STATA Corp., College Station, TX) were the three most prevalent statistical software packages. These packages have also been found to dominate the software used to conduct statistics within the public health literature.[15,16] Therefore, exposure of students to at least one of these packages

through practical laboratory-based experiences in statistic courses would seem prudent for any introductory statistical training.

The overarching strength of this study is its innovative and computationally efficient text mining approach. This enabled a large sample of journal articles published in 2018 to be reviewed, providing a representative sample of current statistical methods used in nutrition and dietetics research not previously achieved. Despite this, the study had some limitations. Firstly, the extraction of statistical information from only the methods sections meant that information contained within other sections was not captured. While it is recommended that all applied statistical methods are described in a paper's methods, descriptive statistics are commonplace in results sections, particularly within figures. This may explain why the descriptive studies reported in our review were proportionately lower than other similar studies. As previous studies have reported descriptive statistics appearing in 95% or more of articles reviewed, it seems plausible that they may have also occurred in almost all articles within this review.

This study provides a representative overview however, its cross-sectional design is unable to establish any emerging trends over time. As trends towards increasingly complex statistical techniques have been observed elsewhere,[14] it is important for nutrition researchers to be familiar with emergent methods. Follow-up studies or retrospective analysis to expand on our findings may provide further insight into any changed or emerging trends within the nutrition and dietetics literature.

While the use of a word cloud was an effective way to visually communicate commonly identified statistical methods, it is limited in its ability to uncover relationships. We are in agreeance with previous studies that highlight the importance of considering how study design and methodology is associated with statistical techniques[21,22] and future research may also consider collecting data on study design and methodology.

Lastly, readers should be aware of limitations relating to the text mining approach itself. One main disadvantage of text mining is the issue of polysemy (one word having multiple meanings) and synonymy (multiple words having the same meaning). Examples of polysemy within our database were single words such as 'average', 'power' and 'sensitivity'. These terms run the risk of their occurrence being overestimated as they may be used in a different context than statistical analysis. Best efforts were made to reduce this possibility by extracting only the information pertaining to statistical analysis within the methods sections. However, it is still possible these terms were used in a different context. Synonymy was also prevalent in the database, as authors used many variations in language to describe a statistical method (Table 3). If not addressed, this may cause some statistical methods to be underrepresented. The generation of a framework was a notable strength of our study, as we were able to replace all corresponding phrases within a synonym set with a specific unigram that represented all phrases of the same statistical method. This significantly reduced the possibility of underestimating statistical methods, however, given the complexity of the English language, it is likely that not all possible synonyms were identified in the database.

In conclusion, this study presented an innovative text mining approach to identify the most frequently reported statistical methods in nutrition and dietetics research. These findings provide useful information for educators to evaluate current statistics curricula and develop short courses for continuing education. They may also act as a starting point for nutrition professionals to educate themselves on typical statistical methods they may encounter.

CONFLICT OF INTEREST

Marijka Batterham is Statistics Editor for Nutrition & Dietetics. This manuscript has been managed throughout the review process by the Journal's Editor-in-Chief. The Journal operates a blinded peer review process and the peer reviewers for this manuscript were unaware of the authors of the manuscript. This process prevents authors who also hold an editorial role to influence the editorial decisions made. There are no further conflicts of interest to declare.

AUTHOR CONTRIBUTIONS

MB initiated the project. AC conducted data collection, extraction, cleaning, and statistical analysis with guidance from MB and EB. MB modified list of searches terms and conducted quality assurance data audits. AC prepared the initial draft of the manuscript with input from MB and EB. All authors approved the final version of the manuscript.

ORCID

Marijka J. Batterham https://orcid.org/0000-0002-9520-6508
Eleanor J. Beck https://orcid.org/0000-0002-3448-6534

REFERENCES

1. Stein K. Propelling the profession with outcomes and evidence: building a robust research agenda at the academy. *J Acad Nutr Diet*. 2017;117(10S):S62-S78.
2. Allman-Farinelli M. Research and dietetic practice: an inevitable linkage. *Nutr Diet*. 2008;65(4):242-243.
3. International Confederation of Dietetic Associations. International Competency Standards for Dietitian-Nutritionists 2016. Available from: https://www.internationaldietetics.

org/Downloads/International-Competency-Standards-for-Dietitian-N.aspx.
4. Tan SY, Hemmelgarn M, Baumgardner K, Tucker RM. Attitudes towards and experiences with research: differences between dietetics students and professionals in Australia and the United States. *Nutr Diet.* 2017;74(4):388-395.
5. Howard AJ, Ferguson M, Wilkinson P, Campbell KL. Involvement in research activities and factors influencing research capacity among dietitians. *J Hum Nutr Diet.* 2013;26(Suppl 1):180-187.
6. Morley-Hauchecorne C, Lepstourel JA. Self-perceived competence of clinical dietitians to participate in research: a needs assessment. *Can J Diet Pract Res.* 2000;61(1):6.
7. Slawson DL, Clemens LH, Bol L. Research and the clinical dietitian: perceptions of the research process and preferred routes to obtaining research skills. *J Am Diet Assoc.* 2000;100(10):5.
8. Harrison JA, Brady AM, Kulinskaya E. The involvement, understanding and attitudes of dietitians towards research and audit. *J Hum Nutr Diet.* 2001;14(4):11.
9. King C, Byham-Gray L, O'Sullivan Maillet J, Scott Parrott J, Splett P, Roberts MM. Dietitians and research: facilitating involvement. *Top Clin Nutr.* 2014;29(3):227-238.
10. Pager S, Holden L, Golenko X. Motivators, enablers, and barriers to building allied health research capacity. *J Multidiscip Healthc.* 2012;5:53-59.
11. Desbro B, Leveritt M, Palmer M, Hughes R. Evaluation of a curriculum initiative designed to enhance the research training of dietetics graduates. *Nutr Diet.* 2014;71:6.
12. Johnson F, Black AT, Koh JC. Practice-based research program promotes dietitians' participation in research. *Can J Diet Pract Res.* 2016;77(1):43-46.
13. Boyd M, Byham-Gray L, Touger-Decker R, Marcus AF, King C. Research interest and research involvement among US registered dietitian nutritionists. *Top Clin Nutr.* 2016;31(3):267-277.
14. Meaney C, Moineddin R, Voruganti T, O'Brien MA, Krueger P, Sullivan F. Text mining describes the use of statistical and epidemiological methods in published medical research. *J Clin Epidemiol.* 2016;74:124-132.
15. Karran J, Moodie E, Wallace M. Statistical method use in public health research. *Scand J Public Health.* 2015;43:776-782.
16. Hayat MJ, Powell A, Johnson T, Cadwell BL. Statistical methods used in the public health literature and implications for training of public health professionals. *PloS One.* 2017;12(6):e0179032-e.
17. Soguero-Ruiz C, Hindberg K, Rojo-Alvarez JL, et al. Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. *J Biomed Health Informatics.* 2016;20(5):1404-1415.
18. SCImago Journal & Country Rank: Scimago Lab; 2019 [April 26, 2019]. Available from: https://www.scimagojr.com/journalrank.php?category=2916.
19. *R: A Language and Environment for Statistical Computing: R Core Team*, R Foundation for Statistical Computing, Vienna, Austria; 2019 [October 28, 2019]. Available from: https://www.R-project.org.
20. Houston L, Probst Y, Martin A. Measuring data quality through a source data verification audit in a clinical research setting. *Stud Health Technol Inform.* 2015;214:107-113.
21. Roush J, Farris J, Bordenave L, Sesso S, Benson A, Millikan C. Commonly used statistical methods in the journals associated with physical therapy and physiotherapy. *J Phys Ther Educ.* 2015;29:5–9 5p.
22. Myoung Jin K, Sung-Bae RP. Statistical techniques and software employed in the journal of sport management between 2006 and 2015. *Int J Sports Sci Coach.* 2017;11(2):3-19.
23. Batterham M. Statistical requirements for reporting nutrition research. *Nutr Diet.* 2011;68(3):3.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Coenen A, Batterham MJ, Beck EJ. Statistical methods and software used in nutrition and dietetics research: A review of the published literature using text mining. *Nutrition & Dietetics.* 2021;78(3):333–342. https://doi.org/10.1111/1747-0080.12678