Journal of Clinical Epidemiology 74 (2016) 124−132

Journal of Clinical Epidemiology

Text mining describes the use of statistical and epidemiological methods in published medical research

Christopher Meaney[a,*], Rahim Moineddin[a], Teja Voruganti[b], Mary Ann O'Brien[a], Paul Krueger[a], Frank Sullivan[a,c]

[a]*Department of Family and Community Medicine, University of Toronto, 500 University Avenue (Suite 346), Toronto, Ontario M5G1V7, Canada*
[b]*Institute of Health Policy, Management and Evaluation, 155 College Street, University of Toronto, Toronto, Ontario M5T3M6, Canada*
[c]*North York General Hospital, 4001 Leslie St., Toronto, Ontario M2K 1E1, Canada*

Accepted 26 October 2015; Published online 19 December 2015

Abstract

Objective: To describe trends in the use of statistical and epidemiological methods in the medical literature over the past 2 decades.

Study Design and Setting: We obtained all 1,028,786 articles from the PubMed Central Open-Access archive (retrieved May 9, 2015). We focused on 113,450 medical research articles. A Delphi panel identified 177 statistical/epidemiological methods pertinent to clinical researchers. We used a text-mining approach to determine if a specific statistical/epidemiological method was encountered in a given article. We report the proportion of articles using a specific method for the entire cross-sectional sample and also stratified into three blocks of time (1995−2005; 2006−2010; 2011−2015).

Results: Numeric descriptive statistics were commonplace (96.4% articles). Other frequently encountered methods groups included statistical inferential concepts (52.9% articles), epidemiological measures of association (53.5% articles) methods for diagnostic/classification accuracy (40.1% articles), hypothesis testing (28.8% articles), ANOVA (23.2% articles), and regression (22.6% articles). We observed relative percent increases in the use of: regression (103.0%), missing data methods (217.9%), survival analysis (147.6%), and correlated data analysis (192.2%).

Conclusions: This study identified commonly encountered and emergent methods used to investigate medical research problems. Clinical researchers must be aware of the methodological landscape in their field, as statistical/epidemiological methods underpin research claims.

Keywords: Statistical methods; Epidemiological methods; Text mining; Bibliometrics; PubMed; Medical research

1. Introduction

Statistical and epidemiological methods often underpin the claims made in published research articles [1,2]. Effective interpretation of medical research requires an understanding of statistical and epidemiological methodology [3−5]. Like other areas of science, medical statistics and epidemiology is an evolving field, where novel and complex methods are continually being developed and incorporated into different disciplines to help solve challenging problems. An understanding of commonly used and emergent methods is important for individuals whose decisions require an understanding of biomedical research, such as practicing physicians, researchers, and students [3−5].

Previous articles have considered the use of statistical methods in medical research. In 1983, Emerson and Colditz [6] performed a review of articles from the New England Journal of Medicine (NEJM) and suggested that knowledge of numeric descriptive statistics, basic hypothesis testing tools, and methods from categorical data analysis allowed a reader the ability to critically appraise nearly three quarters of published articles. [6] In 2005, Horton [7] performed a similar review of a sample of NEJM articles and found evidence of increasingly sophisticated methods being used in the journal. Both reviews concluded that "an acquaintance with a few basic statistical techniques cannot give full statistical access to the research appearing in the journal" [6,7]. Altman and Goodman [8] reviewed NEJM articles from the 1970s and 1980s and found evidence to suggest that classical methods (e.g., *t*-tests; ANOVA; Pearson correlation; contingency

Funding: SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund—Research Excellence; the University of Toronto.

* Corresponding author. Tel.: +1-416-978-5602; fax: +1-416-978-8179.

E-mail address: christopher.meaney@utoronto.ca (C. Meaney).

http://dx.doi.org/10.1016/j.jclinepi.2015.10.020
0895-4356/

What is new?

Key findings?

- The methodological landscape encountered by clinical researchers continues to evolve.

What this adds to what was known?

- Text mining offers a computationally efficient way to process and extract information from ever increasing amounts of medical research literature (e.g., the PubMed Central Open-Access corpus).
- The most frequently encountered statistical and epidemiological methods in medical research include numeric measures of location/spread for continuous data (e.g., mean, median, standard deviation, variance, and interquartile range), counts/percentages for categorical data, regression methods (logistic regression, linear regression, and Cox regression), ANOVA, classical hypothesis tests (e.g., *t*-test, Fisher exact test), inferential statistical machinery (e.g., confidence intervals and *P*-values), concepts about missing data, epidemiological measures of disease burden (e.g., incidence and prevalence), epidemiological measures of association between exposures and outcomes (e.g., odds ratio, risk ratio, hazard ratio), and concepts related to classification and diagnostic accuracy of medical procedures (e.g., sensitivity and specificity).

What is the implication and what should change now?

- Clinical researchers who are acquainted with common and emergent statistical and epidemiological methods can assess and use the greatest proportion of medical research.

tables; epidemiological statistics; and linear, logistic, and Cox regression) were commonplace. They speculated that novel methods such as the bootstrap, procedures based on Gibbs sampling, generalized additive models, generalized estimating equations, random effect models, and classification methods (e.g., CART, neural networks) would increase in use. Taback and Krzyzanowska [9] conducted a review investigating the use of statistical methods published in the abstracts of four high-impact medical journals (*NEJM*, *BMJ*, *JAMA*, *Lancet*) in July 2003. Their review found evidence to suggest that the most commonly used methods were descriptive summary statistics along with classical regression procedures. More complex methods appeared in the article body compared to the abstract itself. A large bibliometric review was conducted by Nietert et al. [10] and focused specifically on methods applicable to general internal medicine (GIM). Using bibliometric methods, Nietert et al. conducted their review on 127,469 articles from GIM. They found that very few (1.7%) of these GIM articles actually cited a statistical article in their reference list. Their study reported similar findings to the others [6–8]: numeric descriptive statistics, epidemiological methods, and basic hypothesis tests were commonplace and that novel and complex methods continue to appear in the medical research literature. Those familiar with more sophisticated methods are able to more fully appraise medical research literature than those without this knowledge.

Our study builds on the existing body of literature. This article used a computationally efficient text-mining design to process and extract information from a large number of medical research articles. The text-mining design allowed us to investigate the methods used in medical research over the past 2 decades: characterizing past trends and making predictions regarding methods for which there exists empirical evidence of increased uptake. This review is broadly useful to readers and users of medical research looking to understand which methods are, or are becoming, increasingly relevant at this moment in time.

2. Methods

2.1. Study approach and design

We obtained a sample of articles from the PubMed Central (PMC) Open-Access (OA) subset of research articles. We used a text-mining design that used efficient computational procedures to process and retrieve information from a large number of articles. We report precise point estimates regarding the proportion of articles using specific statistical and epidemiological methods in OA medical research.

2.2. Databases searched

We included all articles from the PMC OA archive (last updated May 9, 2015). The entire corpus consisted of 1,028,786 articles. The articles were stored as XML (extensible markup language) files and were freely available for download at the following URL: http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/.

2.3. Inclusion/exclusion criteria for sample of articles

We initially included all articles from the PMC OA subset uploaded as of May 9, 2015 (1,028,786 articles). We narrowed our focus to include only medical research articles. We used the Thompson Reuters Journal Citation Report [11] tool to identify articles published in one of 2,001 specific journals from 30 different medical subspecialty groups (250,771 articles). We excluded articles where the type was defined other than ‘‘Original Research’’ (e.g., Review, Commentary, Case Report, etc); the article had a duplicate title; or the article had a short body (less

than 500 words). After applying these exclusion criteria, our final analysis corpus consisted of 113,450 articles.

2.4. Processing free text data

The R software environment for statistical computing (http://www.r-project.org) [12] was used to process and extract the following information from each of the XML articles in the PMC OA subset:

- Title
- Year of Publication
- Journal Title
- Type of Article (Original Research, Commentary, Review, etc)
- Full Free Text Body

The XML package [13] from R was used for processing and extracting information from the raw PubMed XML files.

We used the command line tool GNU-Parallel [14] to process articles in parallel, as opposed to serially. This was essential, as the size of the PubMed OA corpus was challenging to fit into memory on standard computers.

We conducted our analysis using the SciNet General Purpose Cluster computing system (http://www.scinethpc.ca/) [15].

2.5. Statistical terms

We created a list of statistical methods using a modified Delphi style approach, as no comprehensive (gold standard) list of statistical/epidemiological methods exists [16]. A biostatistician (C.M.) drafted the first list of terms. This list was then circulated to another biostatistician (R.M.) and two health services researchers (M.A.O'B. and T.V.) who provided comments and feedback on how to condense/expand the list of terms. C.M. incorporated comments and circulated back to the group. No further changes were suggested. The final list consisted of 177 statistical terms that were thought to be relevant for clinical researchers. Collaborators agreed on headings under which specific statistical methods could be categorized (16 groups); this collapsing of information into a binary composite index permitted a more simplistic summary of broad trends. We were cautious when making inferences from these composite indices, and for transparency, we present results for each specific term comprising each of the composites [17].

In determining terms for inclusion in our list, we specifically focused on statistical and epidemiological methods relevant to medical research. We did not consider terms related to statistical philosophies, traditional and novel study designs, bioinformatics, genetics/genomic analysis, health economics and econometrics, qualitative research, systematic reviews, meta-analysis, and so forth. Future studies could easily build on our methodology to investigate the use of these techniques in medical research.

For transparency, precise details on the 282 terms searched (including acronyms and synonyms), their mapping to 177 specific statistical methods and further their mapping to 16 coarse methods groups are given in Appendix Table 1/Appendix A at www.jclinepi.com.

2.6. Statistical analysis

Inferences regarding trends in the use of specific statistical and epidemiological methods were derived from counting the number of articles that cited a given method (yes/no; ignoring the actual word occurrence counts within a given document). The specific statistical and epidemiological methods were also collapsed into 16 different composite indices using a maximum-value approach [17]. That is, the composite indicator for any given document is evaluated to 1 (yes, TRUE, etc) if any binary member of the composite is 1; else the composite indicator is evaluated to 0 (no, FALSE, etc). We reported counts/proportions of the number of articles in the corpus referencing a specific method in the body of the article (denominator is 113,450). We used a word cloud to visualize which specific terms (of the 177 queried) appeared most frequently in medical research [18]. We present results for the entire corpus and also stratified into three blocks of time (1995−2005; 2006−2010; 2011−2015). The use of two 5-year blocks and one 11-year block was used because the earliest time block consists of the fewest number of OA articles. This is expected as OA articles are being published exponentially more often in recent years [19]. We estimated the relative and absolute percent change in the use of statistical and epidemiological terms for the following time points:

- 2006−2010 vs. 1995−2005
- 2011−2015 vs. 1995−2005
- 2011−2015 vs. 2006−2010

3. Results

3.1. Describing the corpus

As of May 9, 2015 the PMC OA subset contained 1,028,786 XML articles. A total of 1,13,450 articles were original research medical publications, which fell into 1 of 30 Thompson Reuters JCR-defined medical subspecialty classifications. Overall, 987 of 2001 (49.3%) journals specified in the inclusion criteria provided at least one article to the corpus. The remaining journal titles do not contribute to the PMC OA corpus. Table 1 describes the proportional representation of each medical subspecialty group in our corpus, both overall and stratified into three time blocks. Overall, the corpus was comprised of a large proportion of articles from oncology (17.1% articles), GIM (12.5% articles), and immunology (9.6% articles) journal groups. In terms of representation by each specific medical specialty group identified by Thompson Reuters JCR tool, the corpus changed slightly over time. The largest percent decreases in appearance of journal groups

Table 1. Percentage of articles characterized into 30 categories defined by the Thompson Reuters Journal Citation Reports (JCRs) tool, overall and by time period[a,b]

JCR methodological group	Number of journals in JCR category	Percentage of journals in JCR category	Percentage overall (N = 130,469)	Percentage 1995–2005 (N = 9,206)	Percentage 2006–2010 (N = 34,538)	Percentage 2011–2015 (N = 86,725)
Allergy	19	0.8	0.2	0.0	0.1	0.3
Anesthesiology	28	1.1	0.4	0.3	0.2	0.4
Cardiac & Cardiovascular Systems	125	5.1	3.5	1.2	3.0	3.9
Clinical Neurology	190	7.7	3.5	0.9	2.9	3.9
Critical Care Medicine	27	1.1	2.0	3.7	2.7	1.5
Dermatology	59	2.4	0.6	0.0	0.1	0.8
Emergency Medicine	20	0.8	0.4	0.0	0.4	0.5
Endocrinology Metabolism	118	4.8	6.0	2.1	6.8	6.1
Gastroenterology & Hepatology	73	3.0	2.4	1.2	2.0	2.7
Geriatrics & Gerontology	50	2.0	1.7	0.5	1.1	2.1
Hematology	69	2.8	0.7	0.0	0.6	0.9
Immunology	145	5.9	9.6	37.3	8.4	7.1
Infectious Diseases	72	2.9	6.8	10.3	8.4	5.8
Medicine: General Internal	162	6.6	12.5	7.0	12.0	13.2
Obstetrics & Gynecology	79	3.2	1.7	0.9	1.3	1.9
Oncology	200	8.1	17.1	12.2	16.5	17.8
Ophthalmology	58	2.4	2.5	0.4	3.8	2.2
Orthopedics	66	2.7	3.2	1.5	4.1	3.0
Otorhinolaryngology	45	1.8	0.4	0.0	0.3	0.5
Pediatrics	119	4.8	1.9	0.9	1.6	2.1
Peripheral Vascular Disease	65	2.6	0.3	0.0	0.2	0.3
Primary Health Care	18	0.7	0.9	1.2	1.0	0.9
Psychiatry	130	5.3	3.2	1.6	2.6	3.6
Radiology, Nuclear Medicine & Medical Imaging	119	4.8	3.0	2.5	3.4	2.9
Respiratory System	53	2.2	2.2	2.4	2.9	2.0
Rheumatology	29	1.2	3.4	4.3	4.7	2.8
Surgery	201	8.2	3.4	5.0	2.9	3.4
Transplantation	26	1.1	0.1	0.0	0.1	0.2
Tropical Medicine	21	0.9	4.9	1.5	5.0	5.2
Urology & Nephrology	76	3.1	1.4	0.8	0.9	1.6

[a] Very little OA research existed before the year 2000. As such, we created two 5-year blocks and one 11-year block. An analysis based on more year-balanced time blocks from 1995 to 2000 and 2001–2005 would yield very small sample sizes in the earlier time block.

[b] The overall sample size of articles (N = 130,469) exceeds the number of unique articles in the corpus (N = 113,450). Thompson Reuters JCR tool does not necessarily assign a given journal (and hence article) to a single group. A journal can be classified into multiple groups (e.g., Annals of Family Medicine: both Primary Health Care and General Internal Medicine).

came from immunology and infectious diseases, whereas the largest percent increase in appearance of journal groups was observed for GIM, oncology, endocrinology, and metabolism. Figure 1 depicts the 25 journals publishing the greatest proportion of open-access original research medical content housed in PMC. These 25 journals accounted for 46.7% of all articles in our final corpus. 16/30 JCR medical subspecialty categories are represented in this top-25 list, whereas, 14 of 30 are not. The top-25 list also acts to illustrate that some small-sized/midsized JCR medical subspecialty categories are overrepresented with many articles coming from a single (or few) journal titles; for example: critical care medicine, infectious diseases, primary health care, rheumatology, and tropical medicine. Other mid-/large-sized JCR medical subspecialty categories are underrepresented with no journal titles appearing on the top-25 list; for example: cardiac and cardiovascular systems, clinical neurology, and surgery.

3.2. *Estimating the frequency of use of statistical methods*

Overall, numeric descriptive statistics are the most commonly encountered method group and appeared in almost all medical research articles (96.4% articles). Other frequently encountered methods in medical research included: tools for statistical inference (52.9% articles); epidemiological measures of association (53.5% articles); and epidemiological tools for diagnostic accuracy, discrimination, and classification (40.1% articles). Statistical hypothesis testing (28.8% articles), regression (22.6% articles), and ANOVA (23.2% of articles) were also common. Estimates regarding the use of statistical methods groups in our corpus are presented in Table 2, and estimates for the 177 specific statistical terms are given in Appendix Table 2/Appendix B at www.jclinepi.com. In terms of specific statistical and epidemiological methods, the reader of medical research will frequently encounter numeric measures of location/spread

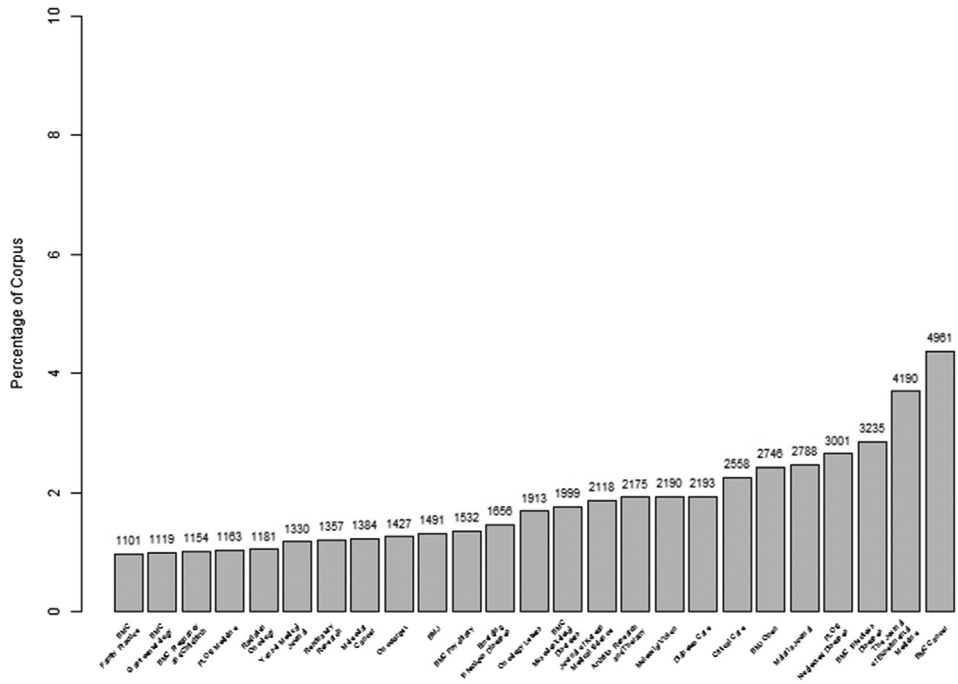

Fig. 1. A bar chart illustrating the 25 journals contributing the greatest proportion of original medical research articles to the PubMed Central Open-Access corpus. The number of articles contributed by each specific journal title is reported above the respective bar in the bar chart.

for continuous data (e.g., mean, median, standard deviation, variance, and interquartile range) counts/percentages for categorical data, regression methods (logistic regression, linear regression and Cox regression), ANOVA, classical hypothesis tests (e.g., *t*-test, Fisher exact test), inferential statistical machinery (e.g., confidence intervals and *P*-values), concepts about missing data, epidemiological measures of disease burden (e.g., incidence and prevalence), measures of association between exposures and outcomes (e.g., odds ratio, risk ratio, hazard ratio), and concepts about classification and diagnostic accuracy (e.g., sensitivity and specificity). The word cloud in Figure 2 graphically depicts specific statistical and epidemiological methods that were encountered frequently in the PubMed corpus.

Fig. 2. A word cloud which uses font size to visualize the most frequently occurring statistical and epidemiological methods observed in the PubMed Central Open-Access corpus.

Certain statistical methods are ubiquitous across medical research (e.g., numeric descriptive statistics, classical epidemiological measures of disease occurrence and association, and statistical inferential concepts); however, it is worth noting that the use statistical and epidemiological methods can vary by medical subspecialty. Appendix Table 3/Appendix C at www.jclinepi.com describes the use of specific statistical methods across the 30 JCR-defined medical subspecialty categories. Appendix Table 4/Appendix D at www.jclinepi.com describes the use of coarse methods groups across the 30 JCR-defined medical subspecialty categories. Some interesting findings emerge when we consider these stratified analyses, for example: time to event regression models are most likely to be observed in oncology research; multilevel models and other regression models for correlated data are most likely to be observed in primary health care research; structural equation modeling, factor analysis, and principal components analysis are most likely observed in psychiatric research; methods in diagnostic medicine (e.g., sensitivity, specificity, positive predictive value, negative predictive value, and receiver operator characteristics curves) are common in radiological research; and time series models (e.g., autoregressive integrated moving average [ARIMA]) are most likely observed in medical specialties related to infectious disease and tropical medicine.

Table 2. Percentage of articles in corpus (N = 113,450) citing the following 16 Delphi panel–derived statistical methods groups

Method groups	Percentage of articles citing specific method group overall and by strata				Relative percent change in method group over time[a]			Examples of commonly occurring terms in each statistical method group
	Overall[b]	95-05[c]	06-10[d]	11-15[e]	06-10 vs. 95-05[f]	11-15 vs. 06-10[g]	11-15 vs. 95-05[h]	
Numeric summary measures	96.4	92.8	96.4	96.8	3.9	0.4	4.3	Average, mode, percentage, standard deviation
Epidemiological measures of risk/effect	53.5	38.0	52.3	55.7	37.5	6.5	46.5	Prevalence, incidence, odds ratio, hazard ratio
Statistical inference concepts	52.9	36.4	52.1	55.0	43.2	5.7	51.4	*P*-value, confidence interval, multiple comparisons
Epidemiological concepts of classification	40.1	43.7	39.8	39.9	−9.0	0.2	−8.8	Sensitivity, specificity, ROC curve
Specific hypothesis test	28.8	23.8	29.0	29.2	21.8	0.7	22.7	*t*-test, Fisher exact test, chi-square test
ANOVA	23.2	14.9	22.2	24.4	49.0	9.8	63.6	ANOVA, ANCOVA, RMANOVA
Regression	22.6	11.9	21.6	24.1	82.1	11.5	103.0	Linear, logistic, poisson regression
Graphics	8.8	8.5	8.7	8.8	2.7	1.3	4.0	Histogram, scatter plot, box plot
Survival analysis	6.8	3.0	6.6	7.3	123.2	11.0	147.6	Cox regression, Kaplan–Meier
Missing data	6.8	2.4	5.9	7.6	148.4	28.0	217.9	Missing data, multiple imputation, LOCF
Computationally intensive algorithms	6.3	3.8	6.2	6.5	63.5	4.5	70.9	Simulation, bootstrap, Monte Carlo, MCMC
Multivariate methods	5.9	3.1	5.5	6.4	73.6	16.6	102.5	Cronbach α, factor analysis, PCA, cluster analysis
Correlated data analysis	4.1	1.6	3.7	4.6	137.3	23.1	192.2	GEE, LMM, GLMM, multilevel model
Machine learning	3.1	1.8	3.0	3.3	66.0	9.9	82.5	Lasso, wavelet, neural network
Time series	1.4	0.9	1.3	1.5	48.8	12.0	66.7	ARIMA, forecasting, spectral analysis
Causal inference observational studies	1.0	0.2	0.8	1.2	268.3	57.8	481.2	Propensity score, instrumental variable

Abbreviations: ROC, receiver operating characteristic; ANCOVA, analysis of covariance; RANOVA, repeated-measures ANOVA; LOCF, last observation carried forward; MCMC, Markov chain Monte Carlo; PCA, principal components analysis; GEE, generalized estimating equations; LMM, linear mixed models; GLMM, generalized linear mixed models.

Percentages quoted for all years combined (1995–2015) and for each stratified time block (1995–2005; 2006–2010; 2011–2015). Relative changes in the percent of articles citing specific methods classes are also reported.

[a] Absolute measures of effect are not displayed; note measures of effect/change are presented on a relative scale.

[b] Overall: N = 113,450.

[c] 1995–2005 strata: N = 8,077.

[d] 2006–2010 strata: N = 29,925.

[e] 2011–2015 strata: N = 75,448.

[f] Relative percent change in statistical method group between 2006 and 2010 vs. 1995–2005.

[g] Relative percent change in statistical method group between 2011 and 2015 vs. 2006–2010.

[h] Relative percent change in statistical method group between 2011 and 2015 vs. 1995–2005.

3.3. Trends in the use of statistical methods over time

The more frequently encountered statistical and epidemiological methods mentioned previously showed stable percent use in each of the three time blocks (e.g., numeric descriptive statistics). Use of regression methods increased appreciably over time (103.0% relative increase from the earliest to the most recent block; 12.2% absolute increase). Other lesser prevalent methods groups that increased in use over time include survival analysis (147.6% relative increase; 4.3% absolute increase), procedures for handling missing data (217.9% relative increase; 5.1% absolute increase), correlated data analysis (192.2% relative increase; 3.0% absolute increase), and methods for causal inference from observational studies (481.2% relative increase; 1.0% absolute increase). A more thorough presentation of percent change in the use of statistical methods across time blocks is given in Table 2.

4. Discussion

Medical statistics and epidemiology involve the collection, analysis, and reporting of data [1]. Quantitative methods play an integral role in all empirical disciplines, including medicine. As such, those reading and using medical research findings should be well versed in statistical and epidemiological methods in order that they can critically appraise this body of literature [2–4].

Previous studies have reviewed and characterized the use of statistical and epidemiological methods in medicine.

Three smaller studies focused on articles published exclusively in NEJM between 1983 and 2005 [6–8]. These studies have demonstrated that the sophistication of statistical methods is increasing (in a specific journal). Previous studies derived similar conclusions, namely, that the more familiar the reader is with statistical methods, the more accessible the corpus of medical research is in terms of their ability to critically appraisal the published literature [6,7]. In terms of the frequency of methods encountered in medical research, our study agrees with previous studies conducted in this field [6–10], namely, that traditional methods related to numeric summaries of data are pervasive. In addition, quantitative epidemiological concepts and inferential tools are also common. Readers must also be familiar with concepts related to hypothesis testing, regression, and ANOVA. However, not all methods are uniformly useful across all medical subspecialties, as such, the use of specific methods can vary according to the medical subspecialty that a researcher/clinician/student is associated with. Appendix Tables 3 and 4/Appendices C and D at www.jclinepi.com act to highlight this discipline-specific variability in the use of statistical/epidemiological methods. The subanalysis produces a wealth of information, and the findings tend to have reasonable face validity. Individuals interested in their discipline-specific methods should consult Appendices Tables 3 and 4/Appendix C and D at www.jclinepi.com for further information.

Altman and Goodman [8] and Nietert et al. [10] designed studies that enabled them to compare the use of specific statistical methodologies across time periods. Most methods increased in use, a trend that we observed as well, which may reflect the increasing empiricism of medical research [20]. Nietert et al. [10] found increasing numbers of citations over time for methods related to meta-analysis (and its extensions), missing data analysis, and epidemiological methods for classification and diagnostic accuracy studies. Altman and Goodman [8] made specific predictions as to methods that they thought may become more prominent in medical research, such as the bootstrap (for estimation of standard errors, model selection, etc), Gibbs sampling procedures (for Bayesian estimation of complex models), generalized additive models, generalized estimating equations, random effect models, and classification methods (CART, neural networks). Our results (Appendix Table 1/Appendix A at www.jclinepi.com) demonstrate the accuracy of some of these predictions. Altman and Goodman [8] were correct in that all methods they listed increased in their frequency of use; however, many of the methods listed are still novel in medical research. The bootstrap appeared in 2.1% of articles in our review (1.4% 1995–2005; 2.0% 2006–2010; 2.1% 2011–2015), Gibbs sampling appeared in 1.1% of articles (0.5% 1995–2005; 1.1% 2006–2010; 1.2% 2011–2015), generalized additive models appeared in 0.5% of articles (0.1% 1995–2005; 0.6% 2006–2010; 0.6% 2011–2015), generalized estimating equations appeared in 0.9% of articles (0.4% 1995–2005; 0.9% 2006–2010; 0.9% 2011–2015), multilevel models or random effects models appeared in 0.8% of articles (0.3% 1995–2005; 0.8% 2006–2010; 0.9% 2011–2015), CART methods appeared in 0.4% of articles (0.2% 1995–2005; 0.3% 2006–2010; 0.4% 2011–2015), and neural networks appeared in 0.5% articles (0.2% 1995–2005; 0.5% 2006–2010; 0.5% 2011–2015). We speculate that many of the methods mentioned by Altman and Goodman will continue to increase in use in upcoming years. Furthermore, we hypothesize that methods for imputation of missing data (1.5% of articles; 0.2% 1995–2005; 1.0% 2006–2010; 1.8% 2011–2015) and propensity score (0.5% of articles; 0.1% 1995–2005; 0.4% 2006–2010; 0.6% 2011–2015) will increase in use going forward. Trends toward the use of increasingly complex methods in medical research may reflect increased training among medical researchers and increased collaboration between clinicians/students and medical statisticians [21]. Alternatively, the trends may be arising from increasingly rigorous demands being placed on medical researchers by journal reviewers and editors who recommend the use of complex/emergent statistical methods to investigate medical research problems [22].

A strength of this study pertains to the computationally efficient text-mining approach used to process, extract, and summarize all the original medical research articles in the PubMed OA corpus. As the amount of medical research literature continues to increase, text mining provides a transparent, auditable, reproducible, and fast method to answer questions pertaining to the content of large medical literature databases. The R code required to conduct and expand on this analysis can be made available by sending an e-mail request to the corresponding author (C.M.). In this review, we processed all articles from the PMC OA subset (over 1 million articles, as of May 2015, and growing). By processing such a large amount of information, we obtain precise estimates of statistical quantities of interest. That said, whether the estimated proportion of articles using specific methods is unbiased relative to all medical research, and not just PMC OA research, is debatable.

This study is not without its limitations. A chief drawback to our methods is that we only consider articles from the OA section of PubMed. It is plausible that methods used in OA research differ from those used in non-OA research. At the current time, mining all medical research poses ethical/legal challenges that restrict using all medical research in this type of text-mining study (or any type of computationally driven review). Computationally, the programs/software used to analyze this set of data would easily scale to a corpus consisting of all original medical research articles (on existing computational architecture) were that pool of information accessible to researchers [23].

Another potential limitation relates to our selection of specific statistical and epidemiological terms. We created a list using a modified Delphi panel approach. Other terms/tokens could be searched, and by no means is our approach

exhaustive or perfect. However, as no comprehensive and agreed on list of statistical and epidemiological terms exist, we created our own, drawing on expertise from biostatisticians and health services researchers in a modified Delphi panel. In the interest of transparency and reproducibility, we have made our list of statistical methods available (see Appendix Table 1/Appendix A at www.jclinepi.com).

A final limitation relates to the text-mining approach itself. Although we searched for and counted the occurrence of statistical and epidemiological terms in a body of literature, the English language is complex and this poses certain problems for text mining and information retrieval. Issues related to synonymy (different words have the same meaning) and polysemy (a single word has multiple meanings) create challenges for the text-mining approach we used. As an example of synonymy, consider attempting to enumerate the occurrence of the rank sum test across our corpus (e.g., rank sum test, Wilcoxon test, Mann−Whitney−Wilcoxon test, etc). As an example of polysemy, consider words like sensitivity (food sensitivity vs. sensitivity of a classifier) or power (power struggle vs. power of a statistical test). For words with multiple synonymous meanings, we run the risk of underestimating their occurrence if we do not include all relevant terms. Conversely, for words with multiple distinct meanings, or short single token words, we run the risk of overestimating their use when we incorrectly attribute a statistical meaning to the word occurrence when really the authors used the word in a different context. Most statistical methods in this study were described by long multiword phrases (e.g., Cox regression, linear discriminant analysis, negative binomial regression, etc) and as such the risk of gross estimation errors in minimal. To further investigate this hypothesis, we conducted a small validation study where one author (T.V.) manually enumerated the occurrence of each specific statistical term across a random sample of 20 articles. For each of the 282 statistical/epidemiological terms under consideration and for each of the 20 randomly selected articles ($20 \times 282 = 5{,}640$ comparisons), we counted the number of times the manual extractor and the text-mining procedure was in agreement. The raw agreement/concordance between the two approaches was 97.6%. To improve on the text-mining approach, more sophisticated preprocessing or retrieval methods could be used to estimate the proportion of articles citing a given statistical or epidemiological method; this is a future area of work for our team [24,25].

5. Conclusions

Over the past 2 decades, the use of statistical and epidemiological methods in medical research continues to increase. We confirm that numeric descriptive statistics, epidemiological measures of association, procedures for inference, hypothesis testing, ANOVA, and regression are commonplace in medical research. However, modern methods are continuously being developed in the statistical and epidemiological research community and are being incorporated into medical research. An understanding of both commonly used and novel/emerging methods is important for researchers, clinicians, and students whose understanding and use of the medical research depends on a comprehensive understanding of the statistical/epidemiological methods used in those medical research articles.

Acknowledgments

The authors would like to acknowledge that the computations performed in this study took place on the GPC supercomputer operated by the SciNet HPC Consortium. The authors would like to thank Ramses van Zon for consultations surrounding our analyses on the SciNet HPC infrastructure.

Authors' contributions: C.M. conceived the study, obtained access to the data, conducted the analyses, organized and participated in the Delphi panel, and wrote the first draft of the article. R.M., T.V., and M.A.O'B. contributed to the design of the study, participated in the Delphi panel, and provided critical feedback regarding interpretation of results and article content. T.V. contributed to the validation study. P.K. and F.S. contributed to the design of the study and provided critical feedback regarding interpretation of results and article content. All authors have read and approved the final version of this article.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2015.10.020.

References

[1] Davidian M, Louis T. Why statistics? Science 2012;336:12.
[2] Brieger K, Hardin J. Medicine, statistics and education: the inextricable link. Chance 2012;25:31−4.
[3] Livingston E. Introducing the JAMA guide to statistics and methods. J Am Med Assoc 2014;312:35.
[4] Altman D, Bland M. Improving doctors understanding of statistics. J R Stat Soc Ser A 1991;154:223−67.
[5] Windush D, Huot S, Green M. Medicine residents understanding of the biostatistics and results in the medical literature. J Am Med Assoc 2007;298:1010−22.
[6] Emerson J, Colditz G. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983;309:709−13.
[7] Horton N. Statistical methods in the journal. N Engl J Med 2005;353:1977−9.
[8] Altman D, Goodman S. Transfer of technology from statistical journals to biomedical literature: past trends and future predictions. J Am Med Assoc 1994;272:129−32.
[9] Taback N, Krzyzanowska M. A survey of abstract of high impact clinical journals indicated most statistical methods presented are summary statistics. J Clin Epidemiol 2008;61:277−81.

[10] Nietert P, Wahlquist A, Herbert T. Characteristics of recent biostatistical methods adopted by researchers publishing in the general internal medicine journals. Stat Med 2013;32:1−10.
[11] Thompson Reuters. *Journal Citation Reports.* Available at http://thomsonreuters.com/en/products-services/scholarly-scientific-research/research-management-and-evaluation/journal-citation-reports.html. Accessed February 1, 2015.
[12] R Core Team. R: A language and environment for statistical computing 2015:Vienna, Austria. Available at http://www.R-project.org Accessed February 1, 2015.
[13] Temple-Lang D, Nolan D. XML and Web Technologies for Data Science with R. New York, NY: Springer; 2014.
[14] Tange O. GNU Parallel: the command line power tool. Berkeley, CA: USENix Magazine; 2011:42−7.
[15] *SciNet General Purpose Computing System.* Available at http://www.scinethpc.ca/. Accessed February 1, 2015.
[16] Van de Van A, Delbecq A. The effectiveness of Delphi and interacting group decision making processes. Acad Manage J 1974;17:605−21.
[17] Tomlinson G, Detsky A. Composite Endpoints in Randomized Trials: there is No free Lunch. J Am Med Assoc 2010;303:267−8.
[18] Fellows, I. *"wordcloud" package for R.* Available at http://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf. Accessed February 1, 2015.
[19] Laakso M, Welling P, Bukvova H, Nyman L, Bjork B, Hedlund T. The development of open access journal publishing from 1993 to 2009. PLoS One 2011;6:e20961.
[20] Guyatt G, Carins J, Churchill D, Cook D, Haynes B, Hirsh J. Evidence based medicine: a new approach to Teaching the Practice of medicine. J Am Med Assoc 1992;268:2420−5.
[21] Altman D, Goodman G, Schroter S. How statistical expertise is used in medical research. J Am Med Assoc 2002;287:2817−20.
[22] Altman D. Statistical reviewing for medical journals. Stat Med 1998;17:2661−74.
[23] *Text mining promises huge benefits but copyright laws can limit its use*. Available at http://www.researchinformation.info/news/news_story.php?news_id=908. Accessed May 1, 2015.
[24] Jurafsky D, Martin J. Natural Language Processing. Englewood Cliffs, NJ: Pearson; 2008.
[25] Manning C. Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press; 1999.