The Effect of Peer Review on the Quality of Data Graphs in *Annals of Emergency Medicine*

David L. Schriger, MD, MPH*; Brian Raffetto, MD; Claire Drolen, BA; Richelle J. Cooper, MD, MSHS

*Corresponding Author. E-mail: schriger@ucla.edu.

Study objective: We determine how peer review affects the quality of published data graphs and how the appointment of a graphics editor affects the quality of graphs in an academic medical journal.

Methods: We conducted an observational time-series analysis to quantify the qualities of data graphs in original manuscripts and published research articles in *Annals of Emergency Medicine* from 2006 to 2012. We retrospectively analyzed 3 distinct periods: before the use of a graphics editor, graph review after a manuscript's acceptance, and graph review just before the first request for revision. Raters blinded to study year scored the quality of original and published graphs using an 85-item instrument. Editorial comments about graphs were classified into 4 major and 16 minor categories.

Results: We studied 60 published articles and their corresponding original submissions during each period (2006, 2009, and 2012). The number of graphs increased 31%, their median data density increased 50%, and quality (completeness [+42%], visual clarity [+64%], and special features [+66%]) increased from submission to publication in all 3 periods. Although geometric mean (0.69, 0.86, and 1.2 pieces of information/cm^2) and median data density (0.44, 0.70, and 1.2 pieces of information/cm^2) were higher in the graphics editor phases, mean data density, completeness, visual clarity, and other markers of quality did not improve or decreased with dedicated graphics editing. The majority of published graphs were bar or pie graphs (49%, 53%, and 60% in 2006, 2009, and 2012, respectively) with low data density in all 3 years.

Conclusion: Peer review unquestionably improved graph quality. However, data densities of most graphs barely exceeded that of printed text, and many graphs failed to present the majority of available data and did not convey those data clearly; there remains much room for improvement. The timing of graphics editor involvement appears to affect the effect of the graph review process. [Ann Emerg Med. 2017;69:444-452.]

Please see page 445 for the Editor's Capsule Summary of this article.

A **feedback** survey is available with each research article published on the Web at www.annemergmed.com.
A **podcast** for this article is available at www.annemergmed.com.

0196-0644/$-see front matter

http://dx.doi.org/10.1016/j.annemergmed.2016.06.046

INTRODUCTION

Background and Importance

Most research articles undergo a peer review process that can involve adding information, clarifying key points, checking for completeness, acknowledging limitations, or ensuring the article is presented in the preferred style of the journal. Although peer review has been in place for centuries, its efficacy is unclear. A 2002 meta-analysis failed to demonstrate any overarching benefit from peer review,[1] but studies that examine specific components of the peer review process provide support for its utility. Wagner and Middleton[2] demonstrated that technical editing, defined as any process designed to "improve accuracy or clarity or impose a predefined style, "increase[d] the readability and quality of articles"; however, the investigation failed to specify the exact editorial processes involved. A 2002 study in this journal concluded that the creation of methodology reviewers resulted in more focused editorial comments,[3] and a second study showed that these comments resulted in modest improvements in article quality.[4]

Despite graphs and figures being commonplace in the biomedical literature, studies demonstrate that published graphs are often suboptimal and fail to use the potential of the format.[5-9] The peer review process has been shown to do little to correct these shortcomings.

Annals of Emergency Medicine has a comparatively sophisticated and well-studied peer review process relative to other journals.[3,4,7,10-13] Papers are initially assessed by one of approximately 40 decision editors, who send approximately 50% of submissions for further review. Since 1997, each article sent for peer review has also been reviewed by a dedicated methodology or statistical editor.

Editor's Capsule Summary

What is already known on this topic

Data graphs in medical science papers are common but vary greatly in their design. How they might be improved to convey more information has been little studied.

What question this study addressed

Was the quality of data graphs affected by increased resources at one journal compared with no special efforts when an editor reviewed graphics after acceptance of the paper and when the editor review occurred before the first request for revision?

What this study adds to our knowledge

The number of graphs, the amount of information conveyed, and measures of quality increased during implementation of the added measures. However, the majority remained simple pie or bar graphs with low data density despite the extra effort.

How this is relevant to clinical practice

Although the graphics editing yielded some improvement, there remained a great deal of room for more.

In addition, beginning in 2008, a senior "graphics" editor was assigned to review each paper. In 2008, this review took place at final acceptance and was the last step in the peer review process before publication. In 2011, the process had changed so that the graphics editor saw the paper before the first request for revision, and his comments were included along with those of the other peer reviewers. During both periods, the graphics editor had access to all comments made by other reviewers.

Goals of This Investigation

In this paper, we examine how the characteristics of graphs changed from submission to publication and from year to year, with a particular focus on whether the introduction of a graphics editor resulted in increased critique of graphs and improved graphic data presentation.

MATERIALS AND METHODS

Study Design

We conducted an observational time-series analysis of the quality of data graphs in *Annals* from 2006 to 2012. We created a sample of 180 randomly selected published research articles composed of 60 manuscripts initially submitted for publication in 2006, 60 submitted in 2009, and 60 submitted in 2012. This represents approximately 60% of original science articles published in these years. These 3 years represent 3 distinct periods in the development of the journal's editorial process: the period before formalized graph review (2006), graph review after the paper was accepted for publication (2009), and graph review at first "revise and resubmit" request (2012).

We separated each figure, including its caption, from the paper and selected a maximum of 5 graphs per paper. If a paper included more than 5 graphs, we randomly selected 5. In this report, we use "figure" to designate an object that is listed as a figure in the paper and "graph" to indicate a depiction of study data in a nontext, nontabular format. As such, figures such as flow diagrams of patient enrollment, theoretical models, or photographs are not graphs. Regardless of whether graphs were labeled as a single figure (eg, Figure 5*A*, *B*, and *C*) or were treated separately (eg, Figures 5, 6, and 7), we scored them as a single graph if they shared at least 1 identical axis and were thematically connected and as separate graphs if they were not.

Data Collection and Processing

We conducted the 3 parts of our data collection process independently: measuring the size of each graph, scoring the quality of each graph, and cataloguing the peer review comments about each graph. For the first part, an assistant, blinded to the purpose of the study and the year of each graph, calculated the area occupied by graph and caption by measuring the height and width in centimeters, using the measurement tool in Adobe Acrobat X Pro 10.1.16 (Adobe Systems, San Jose, CA).

For quality scoring, we trained 5 raters on a standardized set of graphs, and, after achieving adequate interrater reliability on a 20-graph sample (percentage agreement >90% on all 85 items), they scored the content and quality of study graphs presented in isolation, separate from their paper. Each rater scored a balanced number of submitted and published graphs from each of the years. Raters, who were never asked to score both the submitted and published versions of the same graph, were blinded to graph year, but not to whether each graph was from an unpublished or a published article because of differences in formatting that could not be blinded.

Raters scored each figure with a standardized form that captured both the graph's type (eg, bar graph, scatter plot, parallel line plot) and 84 attributes, which included items used to calculate the numerator of the data density index (how many pieces of information are presented per square centimeter of graph), the dimensionality, and items about special features (eg, the depiction of clustering), graph completeness (eg, the presence of appropriate titles and labels), visual clarity (eg, the absence of excessive overlapping), and a gestalt assessment of whether the graph was self-explanatory (Appendix E1, available online at http://www.annemergmed.com). After completing the data form, raters were allowed to refer to the whole paper to confirm that they had correctly understood any abbreviations and had interpreted the purpose of the graph correctly. The data collection form (Excel 2010; Microsoft Corp, Redmond, WA) contained explanations on how to score each item and pull-down menus showing scoring options (Appendix E1, available online at http://www.annemergmed.com). Raters were encouraged to discuss with the authors any scoring decisions for which they were uncertain. During the comparison of published and unpublished versions of a graph, one author (C.D.) had the opportunity to review the scoring of both graphs side by side and correct (in consultation with D.L.S. or R.C. when necessary) any errors.

We created the portmanteau variables “completeness” and “clarity” to indicate whether all elements of these constructs were present and “special features” to indicate whether any such feature was present (Appendices E1 and E2, available online at http://www.annemergmed.com). To be deemed “complete,” a graph had to include 10 essential features, listed and described in full in the scoring form (Appendix E1, available online at http://www.annemergmed.com). A graph with visual “clarity” did not display any of 13 common “visual problems,” which are also listed and defined in the scoring form.

We calculated the data density index, the number of pieces of information per square centimeter, using a modified version of the formula proposed by Tufte.[5,14] In brief, the data density index denominator is the area of the graph and its caption in the journal. Unpublished graphs were assigned the area of the published ones. The numerator was the number of pieces of discrete, unique information contained in a graph. A properly identified bar in a bar graph received 2 points, 1 for the height of the bar and 1 for the bar’s identifier. Each point in a scatter plot received 3 points, 1 for each axis value and 1 for the linkage of those 2 values. In all graphs, points were awarded for additional unique information such as the quantity depicted by an axis, labels, annotations, regression lines, and statistics (see Schriger et al[5] and Tufte[14] for details). For example, the data density index numerator for Figure 1, a bar graph, was calculated as 18 bars plus 18 bar identifiers plus 1 for the labeled, distance-dependent x axis, or 37 points in total. The data density index for Figure 1 is 0.42 (37/86 cm^2).

Determining the numerator of data density index for Figure 2, a series of univariate plots, is slightly more complicated. The calculation can be understood as follows: the approximately 540 visible points in the 6 rightmost sections (the “All years” plots are redundant and are not counted) are multiplied by 3 (1 for the axis value, 1 for the color, and 1 for the shape); to this we add 8 column identifiers, 32 median lines, 24 medians depicted by words or letters, 6 shape and 3 color identifiers, and an axis title, for a total of 1,694 pieces of information and a data density index of 7.5. Figures 3 and 4 have data density indexes of 1.07 (102/97 cm^2) and 0.76 (100/132 cm^2), respectively, and Figure 5 has a data density index of 0.95 (263/277).

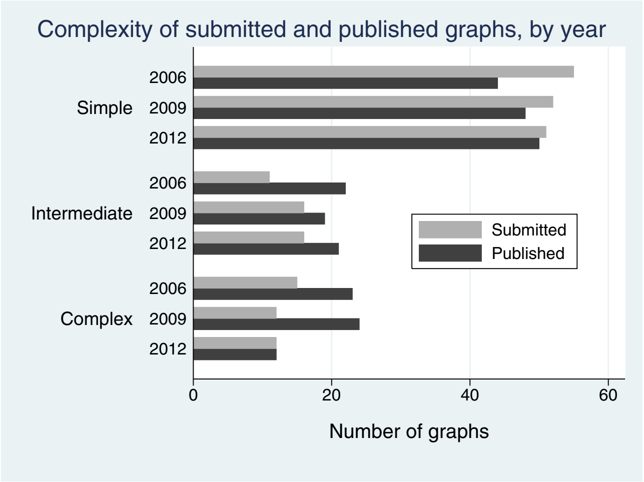

Figure 1. The complexity of the data graphs, stratified by year and submitted manuscript versus published paper. There are less simple graphs and more intermediate and complex graphs in published papers compared with submitted manuscripts. The number of complex graphs submitted was highest in 2006, which may reflect that papers in later years had categorical outcomes or less complex data structures. This graph has a data density index of 0.42 (37/86 cm^2).

We calculated dimensionality by determining the number of characteristics of each data point depicted in the graph. For example, a simple scatter plot would have a dimensionality of 2 because both the x and y axes provide a unique characteristic. However, if the points were color coded to define a particular attribute, the dimensionality would be 3, and, if a symbol or shape were used to convey another attribute, dimensionality would be 4 (Appendix E3, available online at http://www.annemergmed.com). We considered bar, line, and pie charts to be “simple” graphs; box plots, histograms, and receiver operating characteristic curves to be “intermediate” graphs; and scatter plots, survival curves, parallel line plots, parallel coordinate plots, maps, and hybrid graphs to be “complex” graphs.

We analyzed the effect of editorial comments on the quality of graphic presentation by collecting all editorial feedback to authors during the peer review process from the

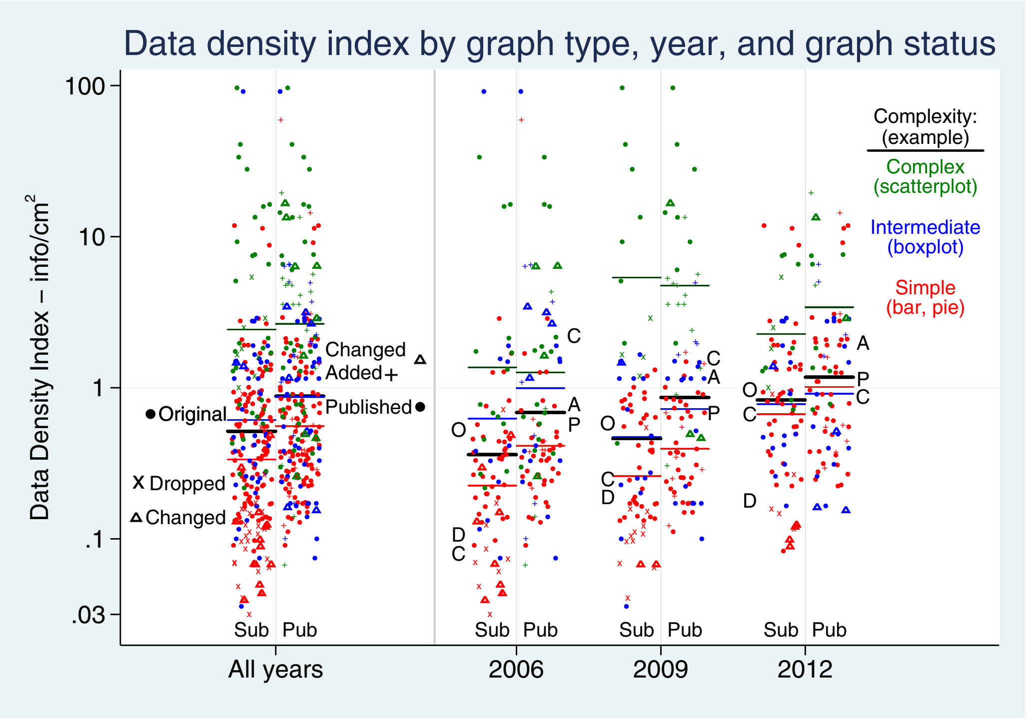

Figure 2. The pattern of data density graphed on a log scale for each study year comparing graphs in submitted manuscripts and published papers. Each point represents 1 graph, with the color indicating the complexity of the graph (simple, intermediate, or complex; see text for definitions) and the shape indicating whether the graph was deleted from the manuscript (x), was added during the revision process (+), changed format (eg, changed from pie chart to scatter plot) (Δ), or was unchanged or underwent minor modifications during the review process (•). The location along the *y* axis of these bolded words in the "All years" panel indicates the geometric mean data density index for each category; the locations of their first letters (O, D, C, A and P) do the same for the by-year data. The thick black line represents the geometric mean data density index for each stratum (submitted or published). The thin colored lines represent subgroup means based on graph complexity. The mean data density index for published graphs was approximately 1 piece of information/cm^2. However, individual graphs had data density indexes ranging from 0.03 to 97 piece of information/cm^2. Several patterns emerge: (1) mean geometric mean data density index increased each year for both submitted manuscripts and published papers, although mean data density index was lower in 2012 because of the absence of very-high-density graphs; (2) data density index and geometric mean data density index increased from submission to publication; (3) simple graphs with low data density indexes tended to be deleted or changed to a different format and added graphs tended to have higher complexity and data density index; (4) graph complexity was highly correlated with data density index. This graph has a data density index of 7.5 piece of information/cm^2.

journal's database (Editorial Manager; Aries Systems, North Andover, MA). This correspondence included comments from the decision editor, the regular content reviewers, the methodology and statistics reviewer, and, depending on the study phase, the graphics editor. One author (C.D.) used the "find" function of Microsoft Word (version 2011; Microsoft Corp) to highlight graphic and statistical terms, such as "graph," "figure," "plot," "bar," "scatter," "box," "axes," and "labels." Using the highlighted terms as a guide, she carefully scanned all text for comments about the manuscript's graphs. She then attributed each comment to 1 or more of the paper's graphs, assigned the topic of the comment to one of 16 categories in 4 major classes (Appendix E4, available online at http://www.annemergmed.com), and noted the role of the person who made the comment (regular reviewer, methodology/statistical reviewer, paper editor, or graphics editor). When a sentence in a review contained concepts that could be attributed to several categories, she parsed that sentence into multiple distinct comments. The fidelity of this process was confirmed by independent coding of a sample of reviews by a senior author (D.L.S.).

Primary Data Analysis

We sought to describe differences in a variety of measures of graph quality between original and published versions of articles and among study years. We programmed Stata (version 14; StataCorp, College Station, TX) to compare each element in the original and published versions of each graph in the context of this inventory of figure-related comments. The difference in medians and its CI were calculated with the cid command in Stata (version 14). For graphs that appeared

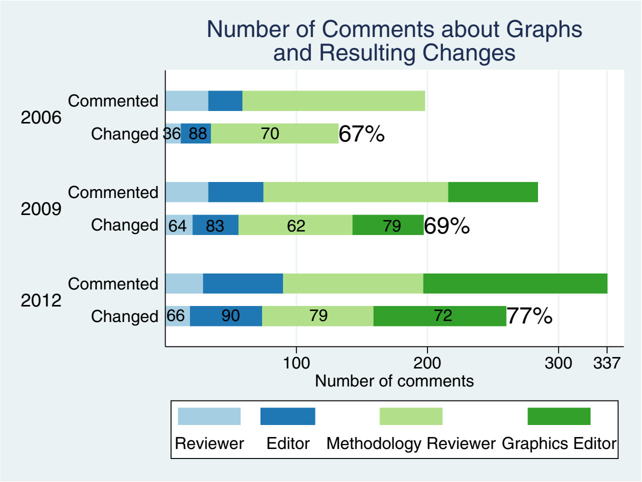

Figure 3. The "Commented" bar indicates the number of comments made about the data graphics by each type of reviewer for each year. The "Changed" bar indicates the number of comments that led to a change. The numbers within each bar represent the percentage of comments that resulted in change for each reviewer type, with the value at the end of each bar reflecting the overall percent changed for that year. Over time there were more comments, largely because of the addition of the graphics editor, and a larger percentage of comments resulted in changes. The data density index of this graph is 1.07 (102/97 cm^2).

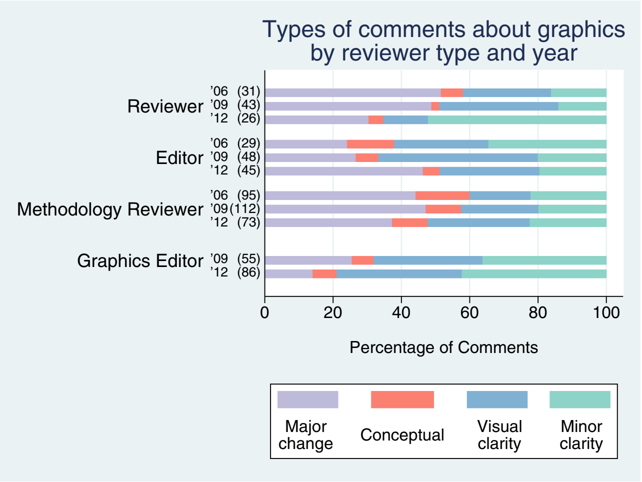

Figure 4. Bars represent the types of comments made by the different types of reviewers in each year. The number in parentheses indicates the total number of comments depicted by the bar. "Major changes" include deleting or adding a graph or changing its format (eg, bar to box plot). "Conceptual" comments refer to issues about the organization of the graph, including failure to depict clustering, failure to depict the full dimensionality of the data, or failure to show the number of subjects represented by a symbol. "Visual clarity" refers to issues related to the appearance of the graph such as overlapping symbols and improperly labeled or scaled axes. "Minor clarity" refers to issues such as ambiguous or unnecessary abbreviations and failure to follow standard conventions for histograms and box plots (Appendix E4, available online at http://www.annemergmed.com). The graphics editor, who received the paper at a later stage in the review process and knows what the other reviewers have written, tends to make more comments about clarity issues with the graph. The methodology reviewers make the most comments about conceptual issues because their percentages are highest and they make the most comments overall. The data density index of this graph is 0.76 (100/132 cm^2).

in only the manuscript or published article, we examined whether comments were responsible for the elimination or addition of the graph. Through this process, we determined how often changes in graphic presentation occurred in response to editorial feedback and how often editorial feedback resulted in a change.

RESULTS

We analyzed 60 original manuscripts and their corresponding published versions during each of the 3 periods (2006, 2009, and 2012) (Figure 5). Between 53% and 60% of the articles in each year contained data graphs, averaging 1.6 (576/360) graphs per all articles and 2.9 (576/201) for articles with at least 1 data graph.

In each scoring period, there were more graphs, on average, in the published manuscripts (1.8; 318/180) than in the original submissions (1.4; 260/180), although the median was 1 in both. Roughly 90% of graphs added between initial submission and publication were prompted by a request from a reviewer or editor (Figure 5). The majority of graphs in both submitted and published articles were bar charts. Box plots, histograms, and scatter plots were the next most common formats (Figure 5). Although simple graphs predominated (Figure 1) in all periods, the peer review process produced a decline in the percentage of them (eg, bar graphs) and a concomitant increase in intermediate and complex graphs in all 3 periods. It also increased the amount of data presented, as demonstrated by higher data density in published data figures than original submissions in each study year (Figure 2). The median (0.46 versus 0.70 pieces of information/cm^2; Δ=0.18; 95% confidence interval [CI] 0.09 to 0.29), geometric mean (0.52 versus 0.88 pieces of information/cm^2; Δ=0.59; 95% CI 0.46 to 0.75), and mean (2.3 versus 3.2 pieces of information/cm^2; Δ=0.85; 95% CI –0.85 to 2.5) data density index increased from submission to publication (the Δ for the geometric mean is multiplicative [submitted is 59% as large as published).

Compared with graphs in initial submissions, published ones were more likely to be self-explanatory (86% versus 93%; Δ=7%; 95% CI 2% to 12%), have visual clarity (27% versus 44%; Δ=17%;95% CI 9% to 25%), be complete (32% versus 45%; Δ=13%; 95% CI 5% to 22%),

	2006		2009		2012			
Graph Editor intervention	None		After Acceptance		Before First Revision		All years	
	Sub.	Pub.	Sub.	Pub.	Sub.	Pub.	Sub.	Pub.
A. Characteristics of study data graphs								
Papers sampled	60	60	60	60	60	60	180	180
Papers with graphs	32	36	33	35	32	33	97	104
Total graphs	91	99	90	134	79	83	260	316
Papers with>5 graphs	2	3	4	5	0	0	10	13
Mean graphs in papers with graphs	2.8	2.8	2.7	3.8	2.5	2.5	2.7	3.0
Mean graphs in all papers	1.5	1.7	1.5	2.2	1.3	1.4	1.4	1.8
Total graphs scored*	81	89	80	91	79	83	240	263
B. Major requests/changes made to graphs								
"Delete from manuscript" (% deleted)	14 (71)		21 (71)		20 (45)		55 (62)	
"Add graph to manuscript" (% added)	15 (73)		23 (48)		6 (38)		44 (55)	
"Change format of graph" (% changed)	17 (65)		18 (67)		13 (54)		48 (63)	
C. Types of graphs, %	Sub.	Pub.	Sub.	Pub.	Sub.	Pub.	Sub.	Pub.
Bar and pie	68	49	65	53	65	60	66	54
Box plots	7	10	14	7	9	12	10	10
Histograms	4	12	5	13	6	7	5	11
Receiver operating characteristic curves	2	2	1	1	5	6	3	3
Parallel axis and parallel line plots	0	2	1	1	1	2	1	2
Survival curves	6	6	4	5	4	0	5	4
Scatter plots	6	10	8	18	3	6	5	11
Maps and hybrid graphs	6	8	3	2	8	6	5	5
D. Efficiency and complexity								
Mean data density index (points/cm²)	2.4	3.4	2.9	3.6	1.6	2.4	2.3	3.2
Geometric mean data density index (points/cm²)	.36	.69	.46	.86	.83	1.2	.52	.88
Median data density index (points/cm²)	.33	.44	.40	.70	.83	1.2	.46	.70
Dimensionality (mean)	2.7	2.8	2.5	2.8	2.7	2.8	2.6	2.8
Percent with any special feature (eg. pairing), %	23%	40%	20%	43%	29%	36%	24%	43%
E. Clarity and completeness (%)								
Graph is self-explanatory	85	90	80	90	92	99	86	93
Complete (elements defined, labeled axes, etc)	32	54	33	42	30	39	32	45
Clear (no redundancies, distortion, etc)	28	52	25	45	28	35	27	44

*This is the denominator for sections C, D, and E

Darker shading denotes improvement, lighter shading worsening between submitted (Sub.) and published (Pub.) manuscript.

The DDI for this table is 1.1 (263/212).

Figure 5. Characteristics of data graphs by study year.

and have special features associated with graphic excellence (24% versus 43%; Δ=19%; 95% CI 11% to 27%).

The number of comments related to data graphs increased from 193 to 286 to 337 during the 3 periods, largely because of comments from the graphics editor (Figure 3). Authors were most likely to act on comments made by the editor, followed by the graphics editor and methodology reviewers, and frequently ignored comments made by the regular reviewers. Overall, authors acted on comments more frequently over time, increasing from a frequency of 67% in 2006 to 69% in 2009 to 77% in 2012. However, this does not hold true in regard to comments requesting a major change. Authors were less likely to comply with requests to delete or add a graph or completely change the format (eg, box plot to scatter plot) in later years (Figure 5). For all reviewers, and especially for the graphics editor, comments about "visual clarity" and "minor clarity" (eg, the appearance of the legend, the overlapping of errors bars) were more common than those requesting a major revision (Figure 4).

The involvement of a graphics editor in 2009 and 2012 was not associated with a consistent improvement in the complexity and quality of published graphs. The complexity of graphs did not increase (Figure 1) and dimensionality remained constant (Figure 5). Although geometric mean data density index (0.69 to 0.86 to 1.2) and median data density index (0.44 to 0.70 to 1.2) increased markedly during the study, mean data density index did not (3.4 to 3.6 to 2.4). The decline in mean data density index (Figures 2 and 5) and the lack of complex graphs (Figure 1) in 2012 may well have been due to an absence of papers requiring complex, high-density graphics in that year (note the absence of green symbols high on the *y* axis in Figure 2 in the submitted articles that year).

By any standard, the data density index was low in all study years; simple text can convey up to 3 pieces of information per square centimeter, and data graphics in journals such as *Nature* and *Science* have mean data density indexes greater than 25.[13] Medical journals have been reported to have data density indexes as high as 13,[15] but in a recent study of 342 graphs in 20 top medical journals the median journal data density index was 1.08 pieces of information/cm^2 (range 0.57 to 1.96), although 10% of graphs had data density indexes greater than 10.[16]

In addition to higher geometric mean and median data density index, graphs in 2009 and 2012 were more often "self-explanatory" at publication (Figure 5). Other markers of quality such as "completeness," "clarity," and "special features" decreased over time in published graphs (Figure 5).

LIMITATIONS

A time-series design is subject to confounding by secular trends. In our case, the types of papers being submitted to the journal may not have been constant over time. For example, as shown in Figure 2, there were some very-high-density scatter plots in 2009, which were not observed in the other periods. We cannot completely discern whether there were more papers that required such graphs in 2009 or whether there were papers of this kind in the other years, but the authors failed to produce such figures. However, median and mean data density index values follow divergent patterns over time, suggesting that papers were not similar among the phases. This dissimilarity confounds our ability to estimate the effect of the graphics editor.

Manuscript and published graphs are formatted differently, and blinding reviewers to their source was all but impossible. However, by restricting reviewer judgments to a series of very discrete items (85 small items instead of a smaller number of general questions), we believe we avoided or at least largely avoided this bias. The denominator of the data density index is the area of the graph, which is determined primarily by the compositor, not the authors. As such, the data density index might vary as much as 4-fold. Given that our graphs showed variations of 1,000-fold (densities of 0.1 to 100) and, in general, that professional compositors will choose an appropriate size for the graph's content, we do not believe that this limitation biased our comparisons in an important way. The rules for calculating the numerator of the data density index are arbitrary and not validated but are certainly a rough proxy for the density of the graph.

In both 2009 and 2012, the graphics editor received each paper after the other reviewers had made their comments. As such, the graphics editor would comment only on novel issues or issues for which he disagreed with the existing comments. Therefore, we cannot tell what suggestions he did not make because they had already been made by another reviewer. Our study is limited to a single journal with a sophisticated peer review process. Results may not apply to other journals. Last, our study focused exclusively on published manuscripts, preventing any analysis of the influence of graphic quality on the accept-reject decision.

DISCUSSION

Our study demonstrates the positive influence of the peer review process on the quality of data graphs from submission to publication; all quantitative measurements suggest that published graphs were more comprehensible, descriptive, and meaningfully complex than their original versions. Methodology and statistical reviewers and graphics editors were the most effective at encouraging these clarifying changes, suggesting the advantage of specialization in the peer review process, particularly in nuanced areas such as graphics (Figure 4). General reviewers made relatively few salient comments about graph quality and conceptual aspects of data portrayal, and their comments frequently went unheeded. Our findings suggest that dedicated reviewers promote necessary and important changes that otherwise may go overlooked.

Peer review is widely accepted and used, but the evidence in its support is surprisingly underwhelming. Our paper adds specific evidence demonstrating that at least 1 aspect of the peer review process unquestionably improves the product. We believe that this would hold for other areas of peer review, but our inability to quantify quality limits research. Although reporting guidelines provide a laundry list of items to be included in a paper, they do not provide guidance about how to rate the quality of the reporting of each item or its importance or how to judge the overall presentation. Without such benchmarks, comparing a pre–peer review and post–peer review methods or results section is difficult. The existence of a strong theoretical basis for rating graphs[14,17-20] and a body of work that operationalized that theory[6-9,16,21] facilitated our research.

Despite clear evidence that peer review improves graphs, the overall quality of the published graphs remained suboptimal. In each year, simple displays such as bar and point graphs vastly outnumbered more complex and informative formats such as box plots, histograms, and bivariate plots. Although the geometric mean and the median data density index increased with the creation of the graphics editor role, the complexity of graphs showed no such increase during the 7-year study period, nor did other measures of graph quality. The absence of clear improvement in graph complexity over time could be due to our study's failure to control for differences in papers' inherent study design—largely whether study outcomes were categorical or continuous—which affects the need for high-complexity plots.

The other possibility is that the timing of the graphics editor's involvement with each manuscript adversely affected performance in 2012. In 2009, the graphics editor received the paper at the end of the editorial process, after the paper was accepted. However, concerns that post-acceptance review was too late to effect substantial improvements motivated the journal to move the graphics editor's involvement to earlier in the peer review process. Although this sends the graphics editor's requests to the author when the other reviews are sent, there may be 2 negative consequences. First, the graphics editor's requests may be diluted by the many requests of the content and methodology/statistical peer reviews and the decision editor and may therefore be ignored. Second, and unlike the situation in 2009, the author's reply went to the decision editor, not the graphics editor, and this person may have failed to recognize inadequate responses from the authors. This may also explain the decrease in clarity, completeness, and special features from 2009 to 2012 because graphs did not receive a final check. To avoid this problem, manuscripts submitted since 2013 have been examined by the graphics editor at initial request for revision and before acceptance.

Previous studies have demonstrated the utility of the peer review process in improving specific aspects of the quality of published journal articles.[2-4] Our analysis establishes that the use of specialized editors—in this case, one focused on data graphs—resulted in detailed comments leading to changes that increase data graphs' informational capacity. Nevertheless, with mean data densities that barely exceed that of printed text and large numbers of graphs that fail to maximize the information conveyed or fail to provide that information clearly and completely, there is much room for improvement. This must occur through the education of authors and peer reviewers about the importance of data presentation,[22] greater dissemination of existing materials on how to design and make quality data graphs,[23-27] and the development of better tools for making such graphs.

The authors acknowledge Daniel Savage, MD, and Marco Fossati-Bellini, MD for assistance in quality scoring the graphs; Tracy Napper, BA, managing editor of Annals of Emergency Medicine, *for help with Editorial Manager; and Samantha Schriger for measuring the size of all graphs.*

Supervising editor: Michael L. Callaham, MD

Author affiliations: From the Emergency Medicine Center (Schriger, Drolen, Cooper) and the UCLA School of Medicine (Schriger, Raffetto, Cooper), University of California, Los Angeles, CA; the Los Angeles County/USC Medical Center & Keck School of Medicine at University of Southern California, Los Angeles, CA (Raffetto); and Amherst College, Amherst, MA (Drolen).

Author contributions: DLS conceived the study and supervised study conduct and data collection. DLS and RJC designed the study. DLS, BF, and CD were responsible for data management, descriptive statistics, and graphics. All authors participated in the drafting and editing of the article. DLS takes responsibility for the paper as a whole.

Funding and support: By *Annals* policy, all authors are required to disclose any and all commercial, financial, and other relationships in any way related to the subject of this article as per ICMJE conflict of interest guidelines (see www.icmje.org). The authors have stated that no such relationships exist and provided the following details: Dr. Schriger's work on this project was supported in part by an unrestricted grant from the Korein Foundation. Dr. Schriger is a deputy editor, graphics editor, and methodology/statistical editor and Dr. Cooper a methodology/statistical editor at *Annals of Emergency Medicine*. Both report receiving compensation for their work at *Annals of Emergency Medicine*.

Publication dates: Received for publication March 15, 2016. Revisions received June 10, 2016, and June 20, 2016. Accepted for publication June 22, 2016. Available online September 7, 2016.

Presented as a poster at the Seventh International Congress on Peer Review and Biomedical Publication, September 2013, Chicago, IL.

REFERENCES

1. Jefferson T, Alderson P, Wager E, et al. Effects of editorial peer review: a systemic review. *JAMA*. 2002;287:2784-2786.
2. Wagner E, Middleton P. Effects of technical editing in biomedical journals: a systemic review. *JAMA*. 2002;287:2821-2824.
3. Day FC, Schriger DL, Todd C, et al. The use of dedicated methodology and statistical reviewers for peer review: a content analysis of comments to authors made by methodology and regular reviewers. *Ann Emerg Med*. 2002;40:329-333.
4. Schriger DL, Cooper RJ, Wears RL, et al. The effect of dedicated methodology and statistical review on published manuscript quality. *Ann Emerg Med*. 2002;40:334-337.

5. Schriger DL, Sinha R, Schroter S, et al. From submission to publication: a retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the *British Medical Journal*. *Ann Emerg Med*. 2006;48:750-756.
6. Cooper RJ, Schriger DL, Close RJ. Graphical literacy: the quality of graphs in a large-circulation journal. *Ann Emerg Med*. 2002;40:317-322.
7. Cooper RJ, Schriger DL, Tashman D. An evaluation of the graphical literacy of the *Annals of Emergency Medicine*. *Ann Emerg Med*. 2001;37:13-19.
8. Schriger DL, Altman DG, Vetter JA, et al. Forest plots in reports of systematic reviews: a cross-sectional study reviewing current practice. *Int J Epidemiol*. 2010;39:421-429.
9. Pocock SJ, Travison TG, Wruck LM. Figures in clinical trial reports: current practice and scope for improvement. *Trials*. 2007;8:36.
10. Callaham ML, Schriger DL. Effect of structured workshop training on subsequent performance of journal peer reviewers. *Ann Emerg Med*. 2002;40:323-328.
11. Callaham ML, Schriger DS, Cooper RJ. An instructional guide for peer reviewers of biomedical manuscripts. Available at: http://www.annemergmed.com/pb/assets/raw/Health%20Advance/journals/ymem/index.html. Accessed July 26, 2015.
12. Green SM, Callaham ML. Implementation of a journal peer reviewer stratification system based on quality and reliability. *Ann Emerg Med*. 2011;57:149-152.
13. Callaham M, McCulloch C. Longitudinal trends in the performance of scientific peer reviewers. *Ann Emerg Med*. 2011;57:141-148.
14. Tufte ER. Data density and small multiples. In: *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press; 1983: 161-175.
15. Tufte ER. Data density and small multiples. In: *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press; 1983:167.
16. Chen JC, Cooper RJ, McMullen ME, et al. Graph quality in top medical journals. *Ann Emerg Med*. 2016; http://10.1016/j.annemergmed.2016.08.463.
17. Tukey JW. *The Collected Works of John Tukey, Vol. V. Graphics: 1965-1985*. Monterey, CA: Wadsworth Advanced Books and Software; 1984.
18. Cleveland WS. *Visualizing Data*. Summit, NJ: Hobart Press; 1993.
19. Cleveland WS. *Elements of Graphing Data*. Monterey, CA: Wadsworth Advanced Books and Software; 1985:229-294.
20. Tufte ER. *Envisioning Information*. Cheshire, CT: Graphic Press; 1990.
21. Cooper RJ, Schriger DL, Wallace RC, et al. The quantity and quality of scientific graphs in pharmaceutical advertisements. *J Gen Intern Med*. 2003;18:294-297.
22. Schriger DL, Savage DF, Altman DG. The presentation of continuous outcomes in randomized trials: an observational study. *BMJ*. 2012;345:e8486.
23. Schriger DL, Cooper RJ. Achieving graphical excellence: suggestions and methods for creating high quality visual displays of experimental data. *Ann Emerg Med*. 2001;37:75-87.
24. Schriger DL. Guidelines for presenting tables and figures in scientific manuscripts. In: Moher D, Altman DG, Schulz K, et al, eds. *Reporting Health Research: A User's Manual*. London, England: J Wiley & Sons; 2014:275-285.
25. Gelman A, Pasarica C, Dodhia R. Let's practice what we preach: turning tables into graphs. *Am Stat*. 2002;56:121-130.
26. MacGregor AL. *Graphics Simplified. How to Plan and Prepare Effective Charts, Graphs, Illustrations and Other Visual Aids*. Toronto, Ontario, Canada: University of Toronto Press; 1979.
27. Wainer H. How to display data badly. *Am Stat*. 1984;38:137-147.

APPENDIX E1

Excel Data Collection Form

APPENDIX E2

Definitions for completeness, visual clarity, and special features

Completeness:

- Title
- *x* Axis, *y* axis titles
- *x* Axis, *y* axis labels (ie, for tick marks or categories; label and units clear)
- Number of subjects is discernable for each graph element
- All data elements defined
- Error bar meaning defined (eg, standard error, 95% CI)
- Box plot elements defined (eg, convention used, whiskers represent 1.5×IQR or whiskers represent 5th and 95th percentile)
- Absence of non–American Medical Association (AMA) abbreviations in graph
- If there are non-AMA abbreviations, abbreviations are defined in graph or caption
- Figure (including content of the caption and legend) is self-explanatory

Visual Clarity (Absence of the Following):
Distortion:

- Improperly scaled axes (nonuniform distance)
- Improperly ranged axes (eg, data are a percentage and thus can range from 0% to 100% and axes display only from 10% to 15%)
- Improper 3D effects
- Improperly connected points (points should be connected only when they are purportedly making measurements on the same population)
- Nonstandard box plots
- Nonstandard histograms (eg, space between bins suggesting zero subjects that is not true)

Chart Junk:

- Moiré/cross-hatching patterns
- Dark/thick/unnecessary grid lines
- Numeric redundancy (eg, clearly labeled *y* axis and number on the top of the bin/category)
- Textual redundancy
- Axis/title/tick label redundancy

Readability:

- Errors bars too cluttered
- Text labels in nonhorizontal orientation
- Graph would be better rotated 90 degrees
- Superimposition of data elements
- Display too small to see symbols
- Problems with labels (eg, labels too small to read or unclear to which graph item label refers)

Special Features:

- Error bars
- Stacking (bar graphs)
- Small multiples
- Number at risk (survival curves)
- Illustration of pairing
- Symbolic dimensionality
- Depicts clustering (on MD, site)
- Extra summary measure(s), ie, mean on box plot

APPENDIX E3

Dimensionality Definitions and Scoring

We calculated a score for dimensionality, using the summation of the following characteristics in each analyzed graph.

Axes in which distance has meaning: Author created axes in which movement along axis has quantitative significance. A scatter plot of height vs weight has 2 such axes.

Axes in which order is important: Author organized elements along a particular axis in an order that makes it easy for the reader to look for patterns within another variable. A bar graph of ordered categories (eg, poor, fair, good, great) has 1 such axis.

Repetition (small multiples): Data figure contains similar graphs showing the data for 2 or more strata. For example, a graph format repeated *x* times for different age groups would receive a score of 1 for repetition. If there were 6 graphs depicting men and women in 3 age strata, there is repetition over age and sex and the figure gets 2 points.

Stratification with labeling: If words are used to identify additional strata. For example, a bar graph with labels distinguishing 2 dimensions (eg, sex and race) would receive a score of 2 (but would not receive any axis points because the axis is categorical).

Shape: Author added shape to signify another variable, such as a scatter plot in which *x*'s represent men and *o*'s represent women.

Shading: Author added shading to data points to signify another variable; would receive a score of 1.

Pattern: Author used pattern to fill data points to signify another variable; would receive a score of 1.

Color: Author used color to introduce another variable that could not be indicated through grayscale shading.

Color intensity (saturation): Author added color intensity or saturation to demonstrate a spectrum. This should be applied only when there are more than 2 colors; otherwise, shading is more appropriate.

Examples:

a) Scatter plot with hollow and solid triangles and circles:
Receives 4 points (2 axes, shape, shading)

b) Bar graph of mean height of hormone and nonhormone recipients, with separate bars for sex and age:
Receives 3 points (*y* axis, 2 labels)

APPENDIX E4

Categories used to classify reviewer and editor comments about graphics

- Major Change
  - New graph
  - Change format of graph (eg, figure to table)
  - Change type of figure (eg, bar graph to scatter plot)
- Conceptual
  - Clustering
  - Dimensionality
  - Depict number behind an element
  - Depict summary statistics on continuous plot (eg, adding a mean to a histogram)
- Visual Clarity
  - Legend
  - Visual problems
  - Error bars
  - Axes
  - Title
- Minor Clarity Issues
  - Abbreviations
  - Problems with histogram design
  - Labels
  - Gridlines