Extrapolation of Correlation Between 2 Variables in 4 General Medical Journals

Yen-Hong Kuo, ScM, MS

ADDING A FITTED REGRESSION line to a scatterplot is helpful for describing an estimated relationship between 2 continuous variables. However, this estimation is valid only within the range of data. Therefore, without knowing information beyond the observed characteristics, it is very risky to extrapolate the fitted line. In spite of this concern, several articles have been published in which the fitted line not only exceeded the range of the data but also reached an undesirable value.

This study was undertaken to assess how prevalent the extrapolation problem is and how this issue was managed in 4 general medical journals.

METHODS

All of the articles published from January through June 2000 in 4 weekly general medical journals (BMJ, JAMA, The Lancet, and The New England Journal of Medicine (NEJM)) were manually reviewed. The main outcome measure was the proportion of articles that involved data extrapolation problems. Extrapolation was defined as an instance in which the fitted line exceeded the observed data range of explanatory variables in the regression model (as shown in the FIGURE). The first step was to identify articles in which at least 1 scatter plot was presented. Subsequently, to assess extrapolation, only articles showing a scatterplot of raw data and a corresponding fitted regression line were included in the analysis.

RESULTS

A total of 178 articles with at least 1 scatter plot were identified. Among them, 37 articles (21%) with scatterplots presenting raw data and a corresponding fitted regression line were included in this study: 5 from BMJ, 7 from JAMA, 23 from The Lancet, and 2 from NEJM. Twenty-two articles (59% [95% confidence interval, 42%-75%]) from all 4 journals involved extrapolation. None changed the line type to indicate extrapolation. Four articles (11%) contained a plot in which the fitted line reached unreasonable or meaningless values. Three articles (8%) stated explicit conclusions about values outside the range of the observed data.

Conclusions A high proportion of the articles analyzed from all 4 weekly general medical journals involved extrapolation without indication. Researchers, reviewers, and editors should be aware of this problem and work to eliminate it.
The hypothetical response scores are on a scale of 0 to 40. The regression line is incorrectly extrapolated to an unreasonable weight (negative value) and meaningless score (>40).

Mathematically, a fitted regression line can be drawn by plugging in almost any real numbers to the estimated equation. However, in clinical applications, this line should not be presented to exceed the range of data. Otherwise, extrapolation can result in reaching an undesirable value. For example, in one study, the fitted line reached a negative value for time from stroke symptom onset to emergency department arrival. This suggests that patients arrived at the emergency department before the onset of stroke symptoms. Another article presented 2 regression lines that unreasonably arrived at a negative level of proteinuria.

Sometimes the error of reaching an undesirable value is not obvious and is difficult to discern unless readers can carefully identify the reasonable range of data in the study (which is not always available). For example, in one article, the fitted line crossed a meaningless area of the Townsend score. The range of scores was not described in the legend of the graph but, rather, in the text. Another study showed a fitted line reaching a value of undetectable cytomegalovirus viral load. A nominal value of negative result was described in the text but not shown in the graph. As a consequence, using an undetectable amount to make a prediction is apparently meaningless.

Limitations of or errors generated by computer programs could be a possible cause of extrapolating or reaching an undesirable value. For example, in the 4 articles that involved extrapolating an undesirable value, all of the lines reached the edges of the graph. However, 9 of 22 articles had extrapolation problems in which the fitted lines did not reach the margins of the graph. Authors are responsible for ensuring that the estimation and presentation of their data are clinically meaningful and should carefully check all data and graphs generated by a computer program.

Problems with extrapolation also involved stating explicit conclusions about values outside the range of the observed data. In some cases, the reader is required to perform some calculations to become aware of the extrapolation problem. For example, in one study, 2 doses used for demonstrating the effect of inhaled corticosteroids on bone mineral density were both outside the range of observed data; one was apparently higher, the other one, after computation, was lower. In another study, the expected time to reach undetectable levels of thymus-dependent T-cell antigen-receptor epitopes not only exceeded the maximum number of years after transplantation but also did not correspond to the fitted line in the graph as well as the regression equation.

With raw data, it is very easy for a reader to be aware of an extrapolation problem from a graph. However, if only a fitted line is presented without the original data points, or even an integrated plot, the extrapolation problem becomes harder to identify. For example, in one article, some of the areas in the constructed contour plot exceeded the range of data. However, this problem could not be identified readily by the reader without performing calculations based on information from the text.

Extrapolation is very dangerous for medical decision making and can result in damaging outcomes. Describing or presenting the estimated correlation within the range of data can prevent potential problems. However, if making an extrapolation is necessary for a researcher, such analysis must be handled with caution. That is, it needs to be described explicitly in the text or indicated in the plot by use of differentiating line types, especially when the range of data is not provided. In addition, during the peer review process, editors and reviewers need to be aware of and identify such extrapolation issues to prevent and eliminate potential problems.

**Author Contribution:** Mr Kuo contributed the conception and design; data acquisition, analysis, and interpretation; manuscript drafting and critical revision for important intellectual content; statistical expertise; and administrative, technical, and material support for this article.

**REFERENCES**


**EXTRAPOLATION OF CORRELATION BETWEEN 2 VARIABLES**

**Figure. Illustration of Extrapolation Problems**

![Illustration of Extrapolation](http://example.com/figure.png)
How Statistical Expertise Is Used in Medical Research

Douglas G. Altman, DSc
Steven N. Goodman, MD, PhD
Sara Schroter, PhD

Statistical content and complexity of medical research has increased steadily over recent decades. Although there is considerable evidence that methodological errors are common in articles in medical journals, much published research does not have substantive contribution from a statistician. Anecdote suggests that many physicians have difficulty getting expert advice or involvement in their research and statisticians are often brought in only at the analysis stage or later.

In 1949, Luykx wrote, “It is now almost inconceivable that a study of any dimensions, in medical science, can be planned without the advice of a statistician.” He clearly appreciated that the most important contribution of the statistician to medical research is at the design stage. We are unaware of any study of the degree to which persons with quantitative expertise are involved in medical research. We report a survey of authors submitting to 2 major medical journals to investigate the nature and frequency of such involvement in their research and statistical involvement in published research does not have substantial methodological errors are common in articles

Methods

Authors of original research articles who submitted to BMJ and Annals of Internal Medicine from May through August 2001 were sent a short questionnaire at the time of manuscript submission. Authors were asked if they received assistance from a statistician with statistical expertise, the nature of any such contribution, and reasons why, if no statistical input was received.

Results

The response rate was 75% (704/943); methodological input was reported for 514 (73%) of these papers. In 435 papers (85%), such input was provided by biostatisticians or epidemiologists and, if deemed significant, was typically associated with authorship. A total of 33 of 122 methodologists (27%) whose main contribution started at the analysis stage received neither acknowledgment nor authorship. Research without methodological assistance was more likely to be rejected without review (71% vs 57%; χ² = 10.6; P = .001) and possibly less likely to be accepted for publication (7% vs 11%; χ² = 2.37; P = .12).

Conclusions

Statistical input to medical research is widely recommended but inconsistently obtained. Individuals providing such expertise are often not involved until the analysis of data and many go unrecognized by either authorship or acknowledgment.

©2002 American Medical Association. All rights reserved.