International Journal of Technology Assessment in Health Care, **20:3** (2004), 296–299.

Retrieval of missing data for meta-analysis: A practical example

George A. Kelley
Kristi S. Kelley
West Virginia University

Zung Vu Tran
National Jewish Medical & Research Center

Objectives: To examine the feasibility of retrieving missing outcome data for summary meta-analyses using an example dealing with the effects of aerobic exercise on lipids and lipoproteins in adults.
Methods: Missing lipid and/or lipoprotein data from a currently developed meta-analytic data base were requested by means of electronic mail from 39 of 174 (22.4 percent) eligible studies. Binary logistic regression was used to examine whether year of publication and country were significant predictors for whether data would be provided.
Results: Of the thirty-nine studies from which data were requested, usable data were received for thirteen (33.3 percent) of the studies. The addition of these previously missing data decreased the percentage of eligible studies that would have had to be excluded by 33.5 percent (from 22.4 percent to 14.9 percent). Neither year of publication nor country in which the study was conducted (United States versus other) were significant predictors of whether missing data would be provided or not ($p > .05$).
Conclusions: Moderate success was achieved in the acquisition of missing outcome data dealing with the effects of aerobic exercise on lipids and lipoproteins in adults. However, whether this level of response is true in other areas of research needs to be determined by additional research.

Keywords: Meta-analysis, Systematic review, Missing data

The use of meta-analysis to review the scientific literature is now common across most fields of research. For example, a recent search of the Medline database for the year 2002 using "meta-analysis" as the keyword resulted in a total of 1,835 available citations (GAK, June 28, 2003). One of the primary goals of meta-analysis is to identify all studies, within the smallest margin of search error possible, that meet the meta-analyst's inclusion criteria on the topic of interest. However, a problem encountered by all meta-analysts is the absence of adequate data from eligible studies for the outcome of interest. For example, a recent meta-analysis on the effects of exercise on resting blood pressure in children and adolescents found that 34 percent of eligible studies in which resting systolic and diastolic blood pressure was assessed had necessary data that was not reported (i.e., missing). For example, the standard deviation of the outcome measure was neither reported nor could it be obtained (1). Because this may introduce a possible bias in the results (data retrieval bias), retrieving as much of this missing data as possible is important. While different statistical procedures for handling missing outcome data exist (replacement with the mean, multiple regression to predict missing values, etc.) all have weaknesses (5). Consequently, it would seem more appropriate to first try and retrieve the actual data from as

This study was supported by grant R01 HL 69802 from the National Institutes of Health, National Heart, Lung and Blood Institute (G.A. Kelley, Principal Investigator).

many eligible studies as possible. Unfortunately, we are not aware of any research that has focused on both the methodology and success of the retrieval of outcome data for a summary meta-analysis. Given the importance of retrieving summary outcome data from study authors and the lack of research documenting methodology and success, the purpose of this study was to examine the feasibility of retrieving missing outcome data using a specific example of a meta-analysis dealing with the effects of aerobic exercise on lipids and lipoproteins in adults.

METHODS

Data Sources

Studies were identified using (i) computerized searches (Medline, Embase, Current Contents, Sport Discus, Dissertation Abstracts International), (ii) cross-referencing from review and original articles, (iii) hand searching selected journals, and (iv) having an expert review our reference list for thoroughness and completeness (Dr. William Haskell, personal communication). All literature searches were performed by the first two authors, independent of each other.

Inclusion Criteria

We attempted to retrieve missing data from studies that met the following inclusion criteria: (i) randomized and non-randomized controlled clinical trials; (ii) chronic aerobic exercise of at least 8 weeks as an intervention (no diet intervention); (iii) adult humans ages 18 years and older; (iv) studies published in journal, dissertation, or masters thesis format; (v) studies published in the English-language; (vi) studies published between January 1955 and December 2002; and (vii) reported assessment of one or more of the following lipids and lipoproteins outcomes: total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, ratio of total cholesterol to high-density lipoprotein cholesterol, ratio of low-density lipoprotein cholesterol to high-density lipoprotein cholesterol, and triglycerides. We did not include studies from non-English language sources because it was beyond the scope of this study. Multiple publication bias was avoided by examining each study for duplicate data on the same subjects and only including the one study that provided the most recent and/or relevant data. We chose 1955 as the starting point for searches because this was the first year that a study on the effects of aerobic exercise on lipids and lipoproteins was conducted (3). Screening to determine whether studies met our inclusion criteria was performed by the first two authors, independent of each other.

Data Retrieval

After identifying those studies that met our inclusion criteria and determining whether additional data were needed, a request was sent by means of electronic mail to the corresponding author of each study. We chose electronic mail as our mode of communication because of its cost-effectiveness and because it is a common form of business communication in most countries today. If no response was received from the corresponding author after our initial request, a second request was sent after approximately 2 weeks by means of electronic mail. If the corresponding author did not respond to our second request within approximately 2 weeks, requests were then sent by means of electronic mail to the other authors listed on the study in the order that each was listed. Each correspondence described the reason for the request, listed the citation from which data were being requested, and described the exact type of data needed. To maximize our response rate, we asked each author to respond at their earliest convenience but left the method (flat file database, relational database, word processing document, embedded in text form in electronic mail) and mode (electronic mail, facsimile, postal mail) up to the discretion of the author. In addition, we limited our request for data to the lipid outcomes from the studies, that is, total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, very low-density lipoprotein cholesterol, ratio of total cholesterol to high-density lipoprotein cholesterol, ratio of low-density lipoprotein cholesterol to high-density lipoprotein cholesterol, and triglycerides. The most common type of missing data we requested for each of these variables was the standard deviations for the above-mentioned outcomes. We did not request other outcome data for variables such as body weight, body fat, etc. (even though they are important covariates), because we believed that it might reduce the amount of lipid data we received. Before the start of this study, approval was obtained by the Institutional Review Board at West Virginia University.

Statistical Analysis

Descriptive statistics (frequencies, percentages, ranges, means and standard deviations, modes) were used to report overall results related to the retrieval of summary data. Binary logistic regression was used to examine potential predictors for whether or not missing summary data were provided. Based on our previous research dealing with the retrieval of individual patient data (2), predictors in our model included country in which the study was conducted (United States versus other), and year of publication. The Hosmer and Lemeshow test was used to identify whether the model adequately fit the data while the Nagelkerke R-squared statistic was used to identify the amount of variance accounted for by the predictor variables. The Nagelkerke R-squared statistic is an adjusted version of the Cox and Snell R-squared. This adjustment was necessary because the Cox and Snell R-squared statistic has a value less than 1 even for a perfect model. Significance of regression coefficients for individual predictor variables was examined using the Wald statistic. In addition, odds ratios and 95 percent confidence intervals were also used to examine the significance of individual

predictor variables. If the confidence intervals overlapped the value of 1.00, they were considered to not be statistically significant (4;6).

RESULTS

Data Retrieval

Of the 174 studies that met our inclusion criteria (references provided upon request), missing lipid outcome data were needed from 39 (22.4 percent). Thirty-four (87.2 percent) of the requests were from studies published in journals, whereas the remaining five (12.8 percent) were from non-journal sources (three dissertations and one masters thesis). The year of publication ranged from 1963 to 2002 (mode year = 1995). We were unable to locate any type of professional address from the authors of two (5.1 percent) studies. Thus requests for missing lipid and lipoprotein data were sent to the authors of thirty-seven of thirty-nine studies (94.9 percent). Of these thirty-seven requests, twenty-nine (78.4 percent) responded. The number of days it took for authors to respond to our initial request ranged from zero to ninety-six (mean ± standard deviation, 15 ± 26 days). Of the thirty-seven studies in which data were requested, acceptable data were received from thirteen (35.1 percent). The number of days from our initial request to the receipt of data ranged from 0 (same day response) to 431 (mean ± standard deviation, 57 ± 121 days). Nine of thirteen studies (69.2 percent) provided us with data embedded in electronic mail, three (23.1 percent) by means of facsimile, and one (7.7 percent) by means of postal mail. Data supplied by the author from one study was excluded because it was not the data that we requested. Thus, of the original 39 studies in which data were needed, acceptable data were retrieved from 33.3 percent of the studies. This decreased our overall percentage of missing data that met our inclusion criteria by 7.5 percent, from 22.4 percent to 14.9 percent. The reasons given by authors who responded to our request but did not provide data are shown in Table 1. As can be seen, 38 percent said they would provide us with data but never did, 23 percent indicated that the data were not available, 15 percent indicated they could not find the data, and 5 percent each reported that the data were either destroyed per Institutional Review Board regulations, not available because of assessment problems, or would take too much time and effort.

Table 1. Responses for Those Studies From Which Data Were Not Supplied

N	Responses
5	Said they would try and locate but data never supplied
3	Data no longer available
2	Couldn't find data
1	Data destroyed per IRB regulations
1	Data not available because of assessment problems
1	Too much time and effort

N, number of responses; IRB, Institutional Review Board.

Logistic Regression Analysis

The results of our logistic regression analysis are shown in Table 2. The Hosmer and Lemeshow test demonstrated that the model adequately fit the data ($\chi^2 = 4.56$, $p = .71$). However, as can be seen, neither country in which the study was conducted nor year of publication were statistically significant predictors of whether or not data were provided ($p > .05$). In addition, the Nagelkerke R-squared statistic showed that year of publication and the country in which the study was conducted predicted only 6.1 percent of the variance for whether or not data were received.

DISCUSSION

The purpose of this study was to examine the feasibility of retrieving missing relevant outcome data for a meta-analysis using an example dealing with the effects of aerobic exercise on lipids and lipoproteins in adults. As a result of this retrieval process, we experienced what we consider to be "moderate" success in obtaining such data, with useable summary data being obtained from approximately one third of the studies in which data were needed. These response rates are similar to previous research in which individual patient data were obtained from 38.2 percent of eligible studies (2).

The acquisition of additional data increased the number of eligible studies we were able to include in our meta-analysis using the original metric. Based on our findings, we believe it is a worthwhile effort to retrieve summary outcome data. We also recommend that the results of such efforts be reported in future meta-analytic studies so that the reader will have more complete information by which to judge the validity of results. Such reporting would be similar to authors

Table 2. Results of Logistic Regression Analysis (n = 37)

Variable	B	SE	df	Wald	Significance	Exp(B)	Exp(B) (95% CI)
Constant	135.35	108.24	1	1.56	0.21	0.00	—
Country	−0.73	0.92	1	0.63	0.43	0.48	0.08–2.94
Year	−0.07	0.05	1	1.57	0.21	0.94	0.84–1.04

B, beta for the regression coefficients of the logistic regression; SE, standard error of the regression coefficients; df, degrees of freedom; Wald statistic calculated as the ratio of B to the SE and then squaring the result; Exp(B), odds ratio; 95% CI for Exp(B), 95% confidence interval for the odds ratio.

reporting data on the percentage loss of subjects in clinical trials.

While we were encouraged that some authors were willing to share their previously unreported data with us, it was disappointing that more than two thirds were not able to provide the information we requested. We were particularly concerned about the response from the author of one study who indicated that the retrieval of such data would take too much time and effort. Because cooperation and trust are the basic foundations of science, we believe that a response such as this is unacceptable.

That neither year of publication nor country in which the study was conducted were predictors of whether or not data were provided is in contrast to previous work dealing with the retrieval of *individual patient data* (2). One possible explanation for the differing results may have to do with the fact that we were requesting *summary* instead of individual patient data. Given these discrepant findings, we would suggest that future research is needed to identify those factors associated with the retrieval of missing outcome data.

The results of this study may contribute to the future planning of meta-analyses so that the retrieval of missing summary outcome data will be part of the research plan. For example, because we were able to obtain electronic mail addresses for approximately 95 percent of studies in which data were needed as well as the finding that 78 percent of those responded to our electronic request, the use of electronic mail appears to be an appropriate and low-cost method for requesting such data. In addition, for authors who responded and supplied data, we found that the vast majority (more than two-thirds) supplied data by means of electronic mail with the remainder supplying data by means of facsimile and postal mail. Consequently, it appears that requests by means of electronic mail will result in data being provided by means of electronic mail. Furthermore, for those authors who responded, we found that it took anywhere from 0 to 96 days to receive any response and 0 to 431 days to obtain data. That it took as long as 431 days to obtain some data is probably unacceptable for most meta-analyses. Therefore, we suggest that future meta-analyses set some type of deadline for the receipt of data.

Given the lack of research in this area as well as the possibility that our results may not be representative of what happens across different types of meta-analyses on different topics, we are unable to generalize our results beyond our current example at this time. Consequently, it is appropriate to suggest that future research is needed to study the retrieval of summary data for meta-analyses dealing with different topics. This information will enable us to better understand the feasibility as well as the most optimal and practical methods for maximizing the amount of outcome data obtained.

In conclusion, moderate success was achieved in the acquisition of missing outcome data for the meta-analytic example provided. However, additional research on other meta-analytic topics is needed.

REFERENCES

1. Kelley GA, Kelley KS. Exercise and resting blood pressure in children and adolescents: A meta-analysis. *Pediatr Exerc Sci*. 2003;15:83-97.
2. Kelley GA, Kelley KS, Tran ZV. Retrieval of individual patient data for an exercise meta-analysis. *Am J Med Sports*. 2002;4:350-354.
3. Mann GV, Teel K, Hayes O, McNally A, Bruno D. Exercise in the disposition of dietary calories: regulation of serum lipoprotein and cholesterol levels in human subjects. *N Engl J Med*. 1955;253:349-355.
4. Petitti DB. *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine*. 2nd ed. New York: Oxford University Press; 2000.
5. Pigott TD. Handling missing data in research synthesis. In: Cooper H, Hedges LV, editors. *The handbook of research synthesis*. New York: Russell Sage Foundation; 1994:163-175.
6. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods of meta-analysis in medical research*. West Sussex, England: Wiley; 2000.