Fish Oil Therapy for IgA Nephropathy: Efficacy and Interstudy Variability

JOHN J. DILLON
Section of Nephrology, The University of Chicago, Chicago, Illinois.

Abstract. Published reports examining the efficacy of fish oil for preserving renal function in immunoglobulin A (IgA) nephropathy have yielded conflicting results. This investigation was a meta-analysis conducted to determine whether the medical literature supports this therapy. In addition, the sources of variability among published findings were examined. Studies were combined using a random effects model. Five controlled studies were identified, two with positive results and three with negative results. Forty-four percent of the between-study variance could be attributed to differences in follow-up times and, less significantly, the number of renal function measurements; a weighting procedure was developed, eliminating this variance from the combined result. When all studies were combined, the mean effect, \( +0.25 \pm 0.23 \) SD (positive effects indicate that treatment was superior to control), was not statistically significant; however, the probability of at least a minor beneficial effect was 75%. Mixed-effects regression suggested that this therapy may be more effective among individuals with more proteinuria. The medical literature, therefore, does not prove the efficacy of fish oil therapy in IgA nephropathy, but suggests that an additional placebo-controlled trial is warranted. A sample-size calculation indicated that such a trial should be larger than those to date or should attempt to increase the treatment effect, perhaps by treating for more than 2 yr or enrolling more severely proteinuric individuals. (J Am Soc Nephrol 8: 1739–1744, 1997)

Materials and Methods

Controlled studies examining the effect of fish oil on the rate at which renal function declines in individuals with IgA nephropathy were identified by searching MEDLINE for the years 1966 through 1995 and from references in published articles. Individual study outcomes were based on different measures of renal function and were, therefore, converted from their original units to units of standard deviations. The effects of the treatment, relative to the control, in standard deviations, were referred to as “effect sizes.” Positive effect sizes indicated that the treatment was superior to the control. The effect sizes were combined using a random effects model (3). Effect size weights were calculated as the inverse of each study’s variance, adjusted for the length of follow-up and the number of measurements of renal function, as described below. Heterogeneity of treatment effects was measured using the Q statistic (3). Unexplained variance between studies was calculated as \( Q = (k-1) \), where \( k \) is the number of studies. The correlations between the treatment duration and the effect size, between the eicosapentanoic acid (EPA) dose and the effect size, between the docosahexanoic acid (DHA) dose and the effect size, between the natural logarithm of the mean, initial, serum creatinine concentration and the effect size, and between the natural logarithm of the mean, initial, daily, urinary protein excretion and the effect size were calculated using weighted regression under an assumption of mixed effects (4). The natural logarithm of protein excretion was used because protein excretion is not normally distributed. The statistical significance of outlier results was tested by contrasting individual study results with the mean of all other studies (5) under a null hypothesis of no difference among effect sizes, using the Dunn-Šidák method (6) to correct for \( k \) comparisons. One study (7) presented paired observations, which were analyzed as such. All tests were two-tailed. Values were expressed as the mean ± SEM. Calculations were made using Quattro Pro for Windows version 5.0 (Corel Corp., Orem, UT) and Minitab for Windows release 11 (Minitab, Inc., State College, PA). (Note that the calculated estimates of the SEM of regression coefficients must be hand-divided by the square root of the mean squared error when Minitab, and most other commercial programs, are used for meta-analysis (4).)

Adjusting for Length of Follow-Up and Number of Measurements of Renal Function

In longitudinal studies, measuring changes in renal function over time, the reliability with which individual outcomes are measured is,
in part, dependent on the follow-up time and the number of measurements made. In particular, the shorter the time between the first and the last measurement, the more the measured rate of change reflects random differences in renal function measurement and the less it reflects true changes in renal function. If a linear change over time is assumed, this may be expressed mathematically as:

\[
S^2 = S_1^2 + S_2^2 / \sum (t_i - \bar{t})^2,
\]

where \(S^2\) is the total variance of an individual rate of change, \(S_1^2\) is the variance attributable to intersubject variability, \(S_2^2\) is the variance attributable to intrasubject variability, \(t_i\) is the time of an individual's \(i^{th}\) measurement and \(\bar{t}\) is the mean time of all measurements for that individual (8).

Relative weights, for studies with different follow-up times and measurement numbers, were expressed as:

\[
w_i = S_i^2 / S^2 = S_i^2 / (S_1^2 + S_2^2 / \sum (t_i - \bar{t})^2),
\]

where \(w_i\) is the relative weight, based on follow-up time and measurement number, but not study size, for the \(i^{th}\) study. Actual, rather than relative, weights were calculated as \(k w_i W_i\), where \(W_i\) is the weight, based on study size alone, of the \(i^{th}\) study and \(k\) is a constant calculated by applying the following constraint, for a quality-weighted variance, to the all-studies variance (9):

\[
\text{all-studies variance} = 1 / \sum k w_i W_i = (\sum w_i^2 W_i) / (\sum w_i W_i)^2
\]

so that:

\[
k = (\sum w_i W_i) / (\sum w_i^2 W_i).
\]

\(S_1^2\) and \(S_2^2\) were approximated as 0.24 and 12.14, respectively (with time measured in months), the values observed for GFR measurements in the Modification of Diet in Renal Disease Trial, Phase II, Study A (8). The \(t_i\) and \(\bar{t}\) were calculated using the mean follow-up and the mean number of measurements in each trial. Because the expression above is little affected by deviations from equal spacing of measurements, the measurements in each study were approximated as equally spaced. Moreover, the duration of follow-up affects this expression far more than does the number of measurements.

When a study’s outcome was failure, defined as a fixed increase in the serum creatinine concentration, \(w_i = 1\), because no longitudinal variable was calculated. This type of outcome is self-weighting regarding duration of follow-up because the study size depends upon the number of individuals reaching the end point, not upon the total number participating in the study.

Choice of Effect Size in Studies Measuring Renal Function by Multiple Methods

The various methods used to measure changes in renal function were ranked by an academic nephrologist who was blinded to the question under study. These rankings, from most reliable to least reliable, were: (1), inulin, \(51\text{Cr}-\text{EDTA}\) or \(125\text{I}-\text{o}l\text{thalmate clearances; (2) creatinine clearance rates; (3) serum creatinine concentrations or their reciprocal; and (4) the presence or absence of a doubling or other percentage change in serum creatinine concentrations (personal communication, Dr. J. Asplin, with permission). For studies in which the rate of change in renal function was measured by more than one method, the single method ranked most reliable was used to calculate the effect size.

Results

Five studies meeting the inclusion criteria were identified (1,7,10–12). No studies were excluded from the analysis. The total number of patients was 202. Table 1 lists the characteristics of the studies, and Table 2 lists the initial patient characteristics. The study outcomes used for this analysis, the sources of the effect sizes, and the weights of the individual studies are shown in Table 3. A detailed description of the effect size and weight sources is provided in the appendix. Whether the Hamazaki et al. study was randomized or blinded was not stated. The effect size for the study by Hamazaki et al. was calculated from the \(P\) value, which was presented as “< .01”; the possibly conservative value of .01 was used. The study by Cheng et al. compared a control period with a (later) treatment period and required progressive deterioration during the control period for inclusion in the study.

The results of the individual studies and the all-studies result, expressed in standard deviations, are displayed in Figure 1. The therapeutic effect of fish oil was statistically significant in the two positive studies, but the detrimental effect was not statistically significant in the three negative studies. The mean effect size for all studies was \(+0.25 \pm 0.23\) SD (\(P = 0.27\)).

Table 1. Study characteristics

<table>
<thead>
<tr>
<th>Author (Reference No.)</th>
<th>(n)</th>
<th>Randomized</th>
<th>Double-Blind</th>
<th>EPA (g/d)</th>
<th>DHA (g/d)</th>
<th>Placebo</th>
<th>Treatment Duration (yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamazaki et al. (1)</td>
<td>20</td>
<td>?</td>
<td>?</td>
<td>1.6</td>
<td>1.0</td>
<td>None</td>
<td>1</td>
</tr>
<tr>
<td>Bennett et al. (10)</td>
<td>37</td>
<td>Yes</td>
<td>No</td>
<td>1.2</td>
<td>1.8</td>
<td>None</td>
<td>2</td>
</tr>
<tr>
<td>Cheng et al. (7)</td>
<td>11\text{c}</td>
<td>Yes</td>
<td>No</td>
<td>1.8</td>
<td>1.8</td>
<td>None</td>
<td>0.8</td>
</tr>
<tr>
<td>Pettersson et al. (11)</td>
<td>32</td>
<td>Yes</td>
<td>Yes</td>
<td>3.3</td>
<td>1.8</td>
<td>Corn oil</td>
<td>0.5</td>
</tr>
<tr>
<td>Donadio et al. (12)</td>
<td>102</td>
<td>Yes</td>
<td>Yes</td>
<td>1.8</td>
<td>1.2</td>
<td>Olive oil</td>
<td>1.7\text{d}</td>
</tr>
</tbody>
</table>

\text{EPA, eicosapentanoic acid; DHA, docosahexanoic acid.}
\text{b Treatment durations and follow-up times were equal.}
\text{c Paired observations.}
\text{d The nominal treatment duration was 2 yr; however, some patients reached earlier end points. The mean treatment duration, calculated from a survival curve, was used.}
A single study that randomized patients between fish oil therapy and a placebo would require 250 patients per group to have an 80% probability of detecting a difference between the two groups equal to the mean effect size, 0.25 SD, at the 0.05 level of significance. If the difference between the two groups were actually +0.70 SD, the upper boundary of the 95% confidence interval, 34 patients per group would be required. (The number of patients required is approximately proportional to the effect size^2.)

Sources of Variability Among Study Outcomes
Weighting for follow-up time and measurement number eliminated 44% of the unexplained variance from the all-studies result, reducing this unexplained variance to a level that was not statistically significant. (When weighted for study size alone, Q = 11.6, P = 0.02 for the null hypothesis of homogeneous treatment effects. After additionally weighting for follow-up time and measurement number, Q = 8.3, P = 0.08.) Of the variance remaining in the final result, 52% was unexplained and 48% was expected from the variances of the individual studies. The greatest outlier was the Hamazaki et al. study, but its deviation from the mean of the other studies was not statistically significant (P = 0.15). Increased proteinuria correlated with increased treatment effect (r = 0.99, P = 0.04; only the three trials (10–12) in which daily urinary protein excretion was measured were used for this calculation). The inverse of the initial serum creatinine concentration (r = 0.0, P = 0.99), the treatment duration (r = 0.1, P = 0.82), the EPA dose (r = −0.55, P = 0.21), and the DHA dose (r = −0.65, P = 0.12) did not correlate with the treatment effect.

Sample-Size Calculation
A single study that randomized patients between fish oil therapy and a placebo would require 251 patients per group to have an 80% probability of detecting a difference between the two groups equal to the mean effect size, 0.25 SD, at the 0.05 level of significance. If the difference between the two groups were actually +0.70 SD, the upper boundary of the 95% confidence interval, 34 patients per group would be required. (The number of patients required is, approximately, proportional to the effect size^2.)

Sensitivity Analysis
No statistically significant all-studies effect was obtained by excluding any single study from the analysis. Likewise, a statistically significant all-studies effect was not obtained by limiting the analysis to the three randomized trials (10–12) (all-studies effect size = +0.17 ± 0.28 SD, P = 0.55). Finally, a statistically significant effect size was not obtained by lim-
Evaluating the analysis to the two randomized trials with the longest treatment durations (10,12) (all-studies effect size = +0.28 ± 0.30 SD, P = 0.36).

The primary outcome in the largest study, that of Donadio et al. (12), an increase of 50% or more in the serum creatinine concentration, was more positive than the outcome dictated by the methods of this analysis, the change in the creatinine clearance rate (effect size = +1.79 ± 0.58 for a 50% serum creatinine concentration increase, +0.53 ± 0.20 for the change in the creatinine clearance rate). Using the larger effect size in place of the smaller did not produce a statistically significant all-studies effect (all-studies effect size = +0.48 ± 0.39, P = 0.22). Furthermore, the study by Donadio et al., using the more positive outcome, was a statistical outlier when contrasted with the mean of the other studies (P = 0.02). Using the actual change in the serum creatinine concentration (effect size = +0.67 ± 0.20) as the outcome for the Donadio et al. study did not produce a statistically significant all-studies effect (all-studies effect size = +0.29 ± 0.26, P = 0.26) and did not produce an outlier result (P = 0.47 for the contrast of the Donadio et al. study with the mean of all other studies).

No choice of P value for the Hamazaki et al. study, over the range between 0.01 and 0.001, produced a statistically significant all-studies effect. This study was a statistical outlier when P values less than 0.002 were used.

Failing to account for follow-up time and measurement number would not have led to a statistically significant all-studies effect (all-studies effect size = +0.21 ± 0.24, P = 0.38).

Discussion

A meta-analysis is no better than its underlying assumptions and clinical trials. Inevitably, the studies analyzed differ in their study populations, methods, and outcomes; dealing with such heterogeneity is one of the key challenges in any meta-analysis. Heterogeneity is both a limitation and an opportunity: a limitation because it may obscure truth, an opportunity because it may suggest explanations.

The five studies analyzed here were performed in different parts of the world and enrolled patients with differing degrees of renal insufficiency and proteinuria. The patients received different doses and were treated for different durations. Changes in renal function were measured in different ways and analyzed using different statistical methods.

A basic, often-used model for meta-analysis is fixed effects. A fixed effects model calculates a weighted, average effect, assuming that outcome differences resulted from chance alone. Fixed effects would have been a poor assumption here, given the differences among trial populations and methods and the finding of statistically significant heterogeneity among treatment effects. Random effects, the model used here, assumes that outcome differences result both from chance and from other, between-study differences. A random effects model separates the variance of the combined outcome into two components, the first derived from the variances of the individual studies and the second derived from the degree to which the study outcomes differ. Random effects differs from fixed effects in that the second variance component is absent under fixed effects. Obtaining statistical significance is more difficult with a random effects model, but the results are likely to be more broadly applicable.

Specific explanations for outcome differences are preferable to accounting for them as random effects. A mixed effects model, used here for regression, is one method for explaining outcome differences. This is a random effects model to which an additional, explanatory, variable has been added. Outcome differences are assumed to result from differences in the explanatory variable, from chance, and from other between-study differences.

An important source of unexplained variation among study outcomes was differing lengths of patient follow-up. This analysis presented a method for weighting individual results on the basis of follow-up time, as well as measurement number, and eliminated these sources of variability from the combined result. The remaining between-study variance, although numerically greater than expected from chance, was not statistically significant. An alternative explanation for this variability would be a change in the treatment effect with more prolonged therapy. These data did not support this alternative; the duration of therapy was uncorrelated with the treatment effect. This lack of correlation should be viewed cautiously; if the treatment effect required more than 2 yr to develop, this analysis could not detect it. In fact, the decades-long course of this disease and the relative lack of side effects from this therapy argue for studying longer treatment durations. I view short treatment times as the most important limitation of this study.

Heavy proteinuria has been associated with more rapid disease progression among individuals with IgA nephropathy (13). This meta-analysis suggests that fish oil may be more effective among individuals with more proteinuria. Although statistically significant, this correlation is tentative because it was based on only three studies (10–12); was based on summary, rather than primary, data; and was based on small between-study differences in proteinuria. Other differences in initial patient characteristics, study characteristics, or omega-3 fatty acid doses did not appear to explain the remaining differences among study results. None of the trials examined histologic changes or differences. These may have been more sensitive indicators of disease progression, especially among individuals with early disease, than functional measurements.

Changes in renal function were approximated as linear. There is little experimental justification for assuming a more complicated function (8). The linear approximation is a study limitation that becomes especially worrisome if effects are projected beyond the 0.5- to 2.0-yr time frames of the trials. In general, nonlinearity is a greater problem with shorter follow-up times; it is one plausible source for the heterogeneity reduced by the weighting procedure used here.

The design used in the study by Cheng et al., requiring progressive deterioration during a control period for later inclusion in a treatment period, was vulnerable to bias in favor of treatment from regression to the mean (14). Conversely, the mean initial serum creatinine concentration in the Cheng study was higher than in the others; the possibility that advanced
disease may diminish the treatment effect has been suggested (15). These data did not support this hypothesis; however, with only five studies, such an effect could have been missed. Eliminating the Cheng et al. study from the analysis did not materially alter the results. The Hamazaki et al. study, published as a letter, provided few details about methods. In particular, the methods used to obtain the P value, from which the effect size was derived, were not provided. Eliminating the Hamazaki et al. study from the analysis did not materially alter the results, nor did limiting the analysis to the three randomized, controlled trials by eliminating both the Cheng et al. and the Hamazaki et al. studies.

Using survival analysis, with a 50% increase in the serum creatinine concentration as the primary outcome, produced an outlier result for the study by Donadio et al. This outcome has at least two limitations. First, information is lost. Substituting survival analysis for changes in a continuous variable substitutes failure times for changes in the continuous variable; the information content of these failure times is similar to the information content of the changes in the continuous variable. However, failure times exist only for those individuals who fail. The statistical power and information content of a survival analysis (the inverse of the variance of the effect size, its weight, in the language of this meta-analysis) is proportional to the number of failures, not to the total number of subjects (16,17). In the study by Donadio et al., only 17 of 106 individuals failed. Numerically, the weight was 4.2 for the survival analysis, 27.2 for the change in the creatinine clearance rate, and 27.5 for the change in the serum creatinine concentration. If following-up the subjects in this study until all had failed had been possible, the weight would have approximated the continuous-variable weights. The second limitation is analogous to the loss of reliability inherent to short follow-up times when measuring longitudinal changes in renal function. The reliability of a survival analysis, defining failure as a fixed elevation in the serum creatinine concentration, decreases as the elevation required decreases. For example, a survival curve based on 10% increases would be less reliable than a curve based upon 200% increases. The smaller the failure-defining change, the more the end point reflects random changes in the creatinine concentration and the less it reflects true changes in renal function. Studies based upon small changes could be down-weighted, but there is no data analogous to that from the MDRD study to estimate the weighting constants. Survival analysis, analyzed nonparametrically as in the study by Donadio et al., has the advantages of avoiding both the assumption of linear changes in renal function and the assumption, inherent to many statistical tests, of normally distributed data. In the Donadio et al. study, a survival curve calculated using either death or end-stage renal disease (ESRD) as the definition of failure produced less positive results, but only modestly so (effect size = +1.34 ± 0.52 SD; when tested as an outlier, P = 0.05). The extent to which the highly positive effect observed with a 50% creatinine concentration increase resulted from the limitations of this outcome, from poor results in the control group (15), from more proteinuria among the subjects enrolled in this study, or from other reasons (for example, today's outlier result may be reproduced by later studies) is uncertain.

Converting treatment effects to standard deviations permitted different measures of renal function to be combined but obscured the clinical meaning of the outcome. Rekola et al. (13) found that the standard deviation of the GFR change among 153 individuals with IgA nephropathy was 6.5 ml/min per yr per 1.73 m². Therefore, the all-studies treatment effect of 0.25 SD may be approximated as 1.6 ml/min per yr per 1.73 m². Although modest and not statistically significant, this effect, if present, would probably be worthwhile clinically for this relatively-benign treatment in this disease without an effective alternative therapy. The probability that fish-oil therapy has at least a small beneficial effect, 0.1 SD or more, may be calculated from the all-studies effect size, its variance and the one-tailed standard normal distribution and was 75%.

This analysis suggests that additional, randomized studies of fish oil therapy for IgA nephropathy are worthwhile but will require large numbers of subjects to detect the modest effect observed. Alternatively, one could plan for a larger treatment effect, attempting to obtain this by treating for more than 2 yr or enrolling more proteinuric participants. A multicenter, randomized, placebo-controlled trial, evaluating 2 yr of therapy with either fish oil or steroids, followed by 3 yr of additional follow-up, is in progress (18); however, the number of subjects, 150 divided among three groups, will be adequate only if the true effect approaches the upper limit of the 95% confidence interval calculated here.

Appendix

A detailed description of the effect size sources is provided, by study, below. Weights were calculated, unless otherwise specified, as N1N2/(N1 + N2), where N1 and N2 are the numbers of subjects in the treatment and control groups. These weights were then adjusted for the length of follow-up and the number of measurements of measurements of renal function, as described in the Methods section.

Hamazaki et al. (1): The effect size was calculated from the P value and the weight using Hedges’s g (19). A P value of 0.01 was used in place of the stated P value of < 0.01.

Bennett et al. (10): The effect size was calculated from the individual creatinine clearance rate changes, presented in Figure 1 of the Bennett et al. article, using Hedges’s g.

Cheng et al. (7): The effect size was calculated from the individual reciprocal serum creatinine concentrations, presented in Table 1 of the Cheng et al. article, using Hedges’s g. Because this was a paired analysis, the weight was equal to the number of subjects.

Pettersson et al. (11): The effect size was calculated from the mean rates of GFR decline in each group during treatment (6.9 ml/min per yr for the fish-oil-treated group, 1.8 ml/min per yr for the corn-oil-treated group) and their standard deviations (12 ml/min per yr for each group) using Hedges’s g.

Donadio et al. (12): The creatinine clearance rate effect size was calculated from the P value of 0.009 and the weight by
using Hedges's g. The 50% serum creatinine increase effect size was calculated as the negative natural logarithm of the hazard ratio (16). The hazard ratio was defined as the hazard in the treatment group divided by the hazard in the control group. Each hazard was calculated as the number of failures divided by the total follow-up time for that group. The total follow-up time for each group was taken from Figure 1 of the Donadio et al. article, the midpoint of each time interval being used for failure or censoring occurring during that interval. The 50% serum creatinine increase weight was calculated as \( \left( \frac{m'}{\text{effect size}} \right)^2 \), where \( m' \) was the value of the two-tailed, standard normal distribution, evaluated at \( P = 0.002 \). Note that the 50% serum creatinine increase effect size and weight may also be calculated nonparametrically, using the log rank test (20). Doing so produced similar results and changed none of the conclusions of this study.

References