Models predict change in plasma triglyceride concentrations and long-chain n-3 polyunsaturated fatty acid proportions in healthy participants after fish oil intervention

Tilly I.T. Potter[1], Graham W. Horgan[2], Anne J. Wanders[3], Elizabeth H. Zandstra[3,4], Peter L. Zock[4], Helena L. Fisk[5], Anne M. Minihane[6], Philip C. Calder[5,7], John C. Mathers[8], Baukje de Roos[1]

[1]Rowett Institute, University of Aberdeen, Foresterhill, Aberdeen, UK

[2]Biomathematics and Statistics Scotland, Aberdeen, UK

[3]Unilever Foods Innovation Centre Wageningen, Wageningen, The Netherlands

[4]Division of Human Nutrition & Health, Wageningen University & Research, Wageningen, The Netherlands

[5]Faculty of Medicine, University of Southampton, Southampton, UK

[6]Norwich Medical School, University of East Anglia, Norwich, UK

[7]NIHR Southampton Biomedical Research Centre, University Hospital Southampton NHS Foundation Trust and University of Southampton, Southampton, UK

[8]Human Nutrition Research Centre, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK*

Correspondence:
Baukje de Roos
b.deroos@abdn.ac.uk

Keywords: Precision nutrition[1], omega-3[2], fish oil[3], statistical modelling[4], secondary analysis[5], crossover study[6]

Abstract

Introduction: Substantial response heterogeneity is commonly seen in dietary intervention trials. In larger datasets, this variability can be exploited to identify predictors, for example genetic and/or phenotypic baseline characteristics, associated with response in an outcome of interest.

Objective: Using data from a placebo-controlled crossover study (the FINGEN study), supplementing with two doses of long chain n-3 polyunsaturated fatty acids (LC n-3 PUFAs), the primary goal of this analysis was to develop models to predict change in concentrations of plasma triglycerides (TG), and in the plasma phosphatidylcholine (PC) LC n-3 PUFAs eicosapentaenoic acid (EPA) + docosahexaenoic acid (DHA), after fish oil (FO) supplementation. A secondary goal was to

establish if clustering of data prior to FO supplementation would lead to identification of groups of participants who responded differentially.

Methods: To generate models for the outcomes of interest, variable selection methods (forward and backward stepwise selection, LASSO and the Boruta algorithm) were applied to identify suitable predictors. The final model was chosen based on the lowest validation set root mean squared error (RMSE) after applying each method across multiple imputed datasets. Unsupervised clustering of data prior to FO supplementation was implemented using k-medoids and hierarchical clustering, with cluster membership compared with changes in plasma TG and plasma PC EPA+DHA.

Results: Models for predicting response showed a greater TG-lowering after 1.8g/d EPA+DHA with lower pre-intervention levels of plasma insulin, LDL cholesterol, C20:3*n*-6 and saturated fat consumption, but higher pre-intervention levels of plasma TG, and serum IL-10 and VCAM-1. Models also showed greater increases in plasma PC EPA+DHA with age and female sex. There were no statistically significant differences in PC EPA+DHA and TG responses between baseline clusters.

Conclusion: Our models established new predictors of response in TG (plasma insulin, LDL cholesterol, C20:3n-6, saturated fat consumption, TG, IL-10 and VCAM-1) and in PC EPA+DHA (age and sex) upon intervention with fish oil. We demonstrate how application of statistical methods can provide new insights for precision nutrition, by predicting participants who are most likely to respond beneficially to nutritional interventions.

1 Introduction

There is often a large degree of variability in physiological outcomes within nutritional intervention studies (1–3). This means that some participants respond beneficially to an intervention, while others may respond poorly or not at all (4). Precision nutrition aims to identify the drivers of these differences, and predict who may respond beneficially (5). While determining response at the level of a single individual requires multiple measurements over time, e.g. through an N-of-1 study (6), predictors of response to outcomes at a group level may be identified through appropriate application of statistical methods in well-powered studies (7). Understanding associations between phenotype, genotype and physiological response could lead to greater understanding of the mechanisms responsible for differential response to interventions, and provide a rational basis for the tailoring of dietary interventions to subgroups of the population (8–10).

Response heterogeneity is seen for physiological markers that can have daily fluctuations, such as plasma triglyceride (TG) concentration (3), as well as those that can vary over longer time periods, such as plasma long-chain n-3 polyunsaturated fatty acids (LC n-3 PUFAs, also called omega-3 fatty acids) (9,11). Plasma concentration of TG and LC n-3 PUFAs are common outcomes of interest in LC n-3 PUFA supplementation trials. Fish oil (FO) is a good source of LC n-3 PUFAs, including eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), which have been shown to lower TG concentrations in many intervention trials (12). An increase in the omega-3 index (EPA + DHA as a percentage of total fatty acids in erythrocyte membranes) has been linked to lower risk of cardiovascular disease (13,14).

The FINGEN study was a double-blind, placebo-controlled crossover study investigating the effects of low (0.7 g EPA+DHA/d, 0.7FO) and medium (1.8 g EPA+DHA/d, 1.8FO) doses of fish oil for 8 weeks on cardiovascular disease risk biomarkers, including plasma TG concentration (15). The

75 FINGEN study revealed greater body weight-adjusted increases in plasma phosphatidylcholine (PC)
76 DHA in men compared with women, with lowering of TG concentration in response to 1.8FO being
77 3 times greater in males, and a trend towards reductions seen in apolipoprotein E4 (APOE4) carriers
78 (15). Significantly higher baseline TG concentrations were observed in APOE4 carriers compared
79 with E2 and E3 carriers (16). However, previous analyses only stratified by two factors (gender and
80 APOE genotype) but did not exploit the whole dataset to identify which of the many available
81 variables could best predict response to intervention, in terms of reductions in plasma TG and
82 increases in PC EPA+DHA after supplementation.

83 Using data from the FINGEN study, the primary goal of this analysis was to identify the predictors
84 that best explain the response heterogeneity of plasma TG and plasma PC EPA+DHA to LC n-3
85 PUFA supplementation, using variable selection methods and validation approaches. The second
86 goal was to determine whether unsupervised analysis of pre-intervention and baseline data could help
87 to identify groups that responded differentially to LC n-3 PUFA supplementation.

88 **2 Methods**

89 **2.1 FINGEN study design and participants**

90 Characteristics of the participants recruited to the FINGEN study, and the methods used, have been
91 reported in full elsewhere (15,16). The original study was approved by the ethics committee at each
92 of the four universities involved in the study (15). Briefly, 312 healthy participants who consumed
93 oily fish less than once a week, recruited at 4 centers in the UK, completed three 8-week intervention
94 periods. They consumed a control oil (an 80:20 blend of palm oil and soybean oil) containing no
95 EPA or DHA, 0.7FO and 1.8FO in a random order, separated by two 12-week washout periods. The
96 participant flow chart can be found in Supplementary Figure 1.

97 Before and after each intervention period, a fasting (12h-fast) blood sample was collected for the
98 measurement of plasma lipids, apolipoproteins, glucose and insulin concentrations (15). Plasma was
99 used for assessment of fatty acid proportions (15); PC is the most abundant phospholipid in plasma
100 (17) and plasma PC EPA+DHA has been shown to be a suitable biomarker of LC n-3 PUFA intake in
101 long-term studies (18). Plasma PC fatty acid composition was determined by gas chromatography.

102 For genotyping, the buffy layer was collected from an ethylenediaminetetraacetic acid (EDTA) tube
103 (BD Biosciences, San Diego, CA, USA) and genomic DNA was extracted using a DNA extraction
104 kit (Qiagen, Hildenberg, Germany), following the manufacturer's instructions. SNP Genotyping was
105 conducting using a commercial SNP genotyping service, TaqMan™ SNP Genotyping Assay, human,
106 Applied Biosystems.

107 **2.2 Data overview**

108 Data were received in Excel sheets and amalgamated to form a single dataset. The dataset included
109 descriptive and physiological variables, dietary intake data, information on single nucleotide
110 polymorphisms (SNPs) and plasma PC fatty acid data. All variables included in this analysis can be
111 found in Supplementary Table 1. Due to lack of variability, SNPs with ≥99% genotype similarity
112 between participants were removed. Data from two participants were removed due to >10% missing
113 data. The complete dataset was imported into R (version 4.1.0), which was used for all statistical
114 analyses. A copy of the (un-imputed) dataset was created, with numeric variables standardized for
115 comparing coefficients in the final models.

116 Prior to multiple imputation, all SNPs and sex (M/F) were coded as factor variables. SNP data was
117 coded 1-3, with 1 corresponding to two reference alleles and 2 and 3 corresponding to one and two
118 non-reference alleles, respectively. All other numeric variables were mean-centered to improve
119 interpretability of the final model coefficients (19). Using the mice package in R (20), collinear
120 variables were removed prior to multiple imputation, which replaced missing values with estimates
121 from the distribution of the remaining data (20). Missing data per variable was between 0-6%, with
122 total missing data just under 1%. Multiple imputation generated 5 complete imputed (independent)
123 datasets. 5 imputations were chosen and deemed acceptable due to the size of the dataset and low
124 amount of total missing data, meaning the variation between the imputed datasets was expected to be
125 low (20). To improve statistical power, SNPs were converted back to numeric variables after
126 imputation, aside from codes designating APOE variant (2 = E2/E2 + E2/E3, 3 = E3/E3, 4 = E3/E4 +
127 E4/E4; rs429358 and rs7412) and endothelial nitric oxide synthase (eNOS, rs1799983; 1 = GG, 2 =
128 GT, 3 = TT) due to their inclusion as basic characteristics in the original dataset. Details of all SNPs
129 and their reference IDs can be found in Supplementary Table 1.

130 Each imputed dataset was divided into a dataset containing all baseline variables and data collected
131 prior to the 0.7FO treatment arm (0.7FO dataset), and a dataset containing all baseline variables and
132 data collected prior to the 1.8FO treatment arm (1.8FO dataset), to examine the predictors of
133 response prior to each treatment arm separately. In total, each imputed dataset contained 98 variables
134 (including volunteer identifier and outcome variables) and 310 participants.

135 This study focused on two outcomes: change in plasma TG concentration, and change in plasma PC
136 EPA+DHA calculated from the difference in EPA+DHA proportion, as a percentage of total fatty
137 acids, pre- and post- fish oil supplementation. For the purpose of this report, these outcomes are
138 referred to as change scores. Outcomes were used on a continuous scale rather than as a dichotomous
139 classification (e.g. response/non-response) to maximize use of information and statistical power
140 (21,22). To examine if there were significant differences in the outcomes of interest between
141 treatment arms, ANOVA tests with Huynh-Feldt correction were conducted (23). To determine
142 whether supervised analysis for both outcomes was appropriate after each FO intervention, the
143 standard deviation (SD) of the change scores after 0.7FO or 1.8FO were compared with the change
144 scores after control oil for each outcome. A greater change score SD after either 0.7FO or 1.8FO
145 compared with control oil was indicative of response heterogeneity (24). However, if the control oil
146 change score SD was larger than either of the FO change score SDs, no further supervised analysis
147 was undertaken, as differences between participants after FO could be explained by random
148 variability alone (24).

149 **2.3 Data analysis strategy**

150 **2.3.1 Clustering of pre-intervention data.**

151 Figure 1 provides an overview of the procedures for data analysis. After imputation, unsupervised
152 cluster analysis was conducted with all non-outcome variables, in the 0.7FO and 1.8FO datasets
153 respectively. For each imputed dataset, a dissimilarity matrix was constructed using the “daisy”
154 command within the R cluster package. Each value in the matrix referred to the distance between
155 participants, with higher values corresponding to greater dissimilarity (25).

156 Two different clustering methods were conducted, in order to determine which method led to clearest
157 cluster segregation upon visual inspection. These methods were PAM (Partitioning Around Medoids)
158 also known as k-medoids clustering, where k, the number of clusters, must be stipulated (26); and
159 hierarchical clustering (27), calculating the distance between participants and merging them via

160 application of linkage methods (28). The highest average silhouette value was used to determine the
161 optimal number of clusters after PAM clustering, while the cluster dendrogram informed the number
162 of clusters after hierarchical clustering, with clusters separated using the cutree function. The optimal
163 linkage method for computing the cluster dendrograms was selected by comparing the agglomerative
164 coefficient of four methods (average, single and complete linkage, and Ward's minimum variance),
165 with the highest value determining the method chosen. These procedures were performed using the
166 cluster and stats R packages. Final cluster membership was defined as the cluster most frequently
167 assigned to each participant across the 0.7FO and 1.8FO imputed datasets, respectively (≥3/5 of the
168 imputed datasets).

169 Dimension reduction, via principal components analysis (PCA), was undertaken using the stats R
170 package, with results visualized using the *ggbiplot* package. The variables with the greatest loadings
171 on each component were examined.

172 **2.3.2 Supervised analysis methods**

173 Several variable selection techniques were chosen to generate models with relevant predictors for
174 each outcome of interest. Results across the 5 imputed datasets were aggregated to form final models
175 and to compare methods. Figure 1 presents a general overview of the analysis procedure.

176 Using the leaps package in R, forward stepwise selection was used to add predictors sequentially that
177 maximally improved the fit of the model to the given outcome. Then, backwards selection was used,
178 starting with a model containing all predictors and sequentially removing predictors that added least
179 to the fit. Both methods were appropriate for the FINGEN dataset since the number of participants
180 was greater than the number of predictors (29).

181 Next, the shrinkage method LASSO (Least Angle Selection and Shrinkage Operator) was applied
182 using the *glmnet* package in R (30). Briefly, the method applies a parameter, lambda (λ), which
183 shrinks the model coefficients to zero as it increases. Non-zero coefficients therefore represent the
184 most useful predictors. These can be any combination of variables, unlike stepwise selection where
185 predictors are added or subtracted iteratively (29). Finally, a variable selection technique that makes
186 use of a non-linear method, Random Forest regression, was applied – the Boruta algorithm, using the
187 Boruta package in R. The algorithm works by comparing the importance of each variable in the
188 dataset to a set of randomly shuffled values, known as shadow features. Variables are confirmed as
189 important or rejected after a series of iterations (31).

190 **2.3.3 Model selection and method comparison**

191 For each analysis method, and for each imputed dataset, 10-fold cross-validation or separate training
192 and validation sets were used to select and validate models. For the stepwise selection techniques,
193 10-fold cross-validation was used to identify the optimal model size that led to the lowest validation
194 set root mean squared error (RMSE) - the amount of error using the remainder of the data not used in
195 model development. Participants were split into 10 random folds using the set.seed function in R. For
196 each possible model size (from 1:n, constrained by the number of participants per fold), 9 folds were
197 used as the training set, while 1 fold was used as a test of the model, providing the validation RMSE.
198 This was repeated for each fold, with the average validation RMSE taken across all folds for each
199 model size. To maximize power, the selected model size was run using all data to identify the
200 relevant predictors. For example, if a model containing 3 predictors had the lowest validation RMSE
201 after 10-fold cross-validation, the 3-variable model using the full dataset was examined to identify
202 the resulting variables and coefficients.

203 The *glmnet* package for LASSO automatically performs 10-fold cross-validation and provides a
204 range of plausible λ values. To determine the optimal λ value and resulting model, validation was
205 performed using a random 2/3 of the data as the training set with the other 1/3 as the validation set.
206 The λ value associated with the lowest validation set RMSE was used to select the corresponding full
207 model. Similarly, for the Boruta algorithm, a random 2/3 of the data was retained in the training set,
208 to maximize shuffling of the shadow features and to improve variable selection. Random Forest
209 regression using the selected variables only was then run with the training data, and used to predict
210 the outcome using the test data, with RMSE calculated.

211 For stepwise methods, a variable was included in a final pooled linear model if it was included in at
212 least 3 out of 5 of the imputed dataset models. The pooled regression was conducted on all imputed
213 datasets simultaneously using the with function in R and pool function within the mice package (20).
214 Non-zero coefficients that remained across ≥3/5 of the LASSO models were averaged and retained as
215 important predictors. Variables identified as important across ≥3/5 Boruta models were considered
216 the most relevant for the given outcome.

217 The method that led to models with the lowest average validation set RMSE across the 5 imputed
218 datasets was considered the best fit for a given outcome, i.e., the model gave the best predictions for
219 change in plasma TG or plasma PC EPA+DHA after intervention. Final models, with the lowest
220 validation set RMSE, are presented in two forms: with numeric coefficients mean-centered but
221 unstandardized, for model interpretability; and with standardized numeric coefficients, for the
222 relative importance of predictors to be compared. For stepwise selection methods, the adjusted R2
223 value quantified the goodness of fit of the models.

224 Due to anticipated high correlation between change score and pre-intervention value (e.g. TG change
225 vs pre-intervention TG levels), Oldham's transformation was performed to determine whether the
226 relationship could be explained by regression to the mean (32). The transformation compares the
227 mean of baseline and final values of an outcome against the change score. If the relationship between
228 change score and pre-intervention value was due to regression to the mean, no significant relationship
229 would remain after the transformation.

230 3 Results

231 3.1 Outcome change scores

232 Table 1 shows the average changes in plasma TG and PC EPA+DHA after each intervention arm of
233 the study. A repeated measures ANOVA with Huynh-Feldt correction showed that mean plasma TG
234 change differed significantly between intervention arms [$F(1.936, 598.2) = 10.19$, $p<0.001$], as has
235 been previously reported (15). Pairwise comparisons using Bonferroni correction revealed that there
236 was a significant reduction in TG concentrations after 0.7FO and 1.8FO compared with control oil,
237 but the difference in TG change between 0.7FO and 1.8FO was not significant (Table 2). For plasma
238 TG change, the change score SD was greater after 1.8FO than after the control oil, but was greater
239 after control oil compared with 0.7FO. This meant that subsequent supervised analyses of TG change
240 after 1.8FO only could be conducted.

241 Repeated measures ANOVA with Huynh-Feldt correction showed that mean PC EPA+DHA change
242 differed significantly between intervention arms [$F(1.895, 585.5) = 636.1$, $p<0.001$]. Pairwise
243 comparisons with Bonferroni correction revealed that there were significant differences in PC
244 EPA+DHA change between all intervention arms (Table 2), with mean plasma PC EPA+DHA as a
245 proportion of total fatty acids increasing by 3.05% and 4.65% after 0.7FO and 1.8FO, respectively

246 (Table 1). The change score SD was greater after both 0.7FO and 1.8FO compared with control oil,
247 meaning subsequent supervised analyses could be conducted after both fish oil interventions (Table
248 1).

249 **3.2 Clustering analysis**

250 **3.2.1 0.7FO dataset**

251 Hierarchical clustering using Ward's method led to clearest discrimination of clusters, resulting in
252 two clusters with 161 and 149 participants in clusters 1 and 2, respectively (Figure 2a). PCA revealed
253 a degree of separation of the two clusters across the first two principal components (PCs) (Figure 2b).
254 There was no significant difference in plasma TG change after 0.7FO between the two clusters. Mean
255 change in plasma PC EPA+DHA for participants in cluster 1 (3.22%) was not significantly greater
256 than EPA+DHA change for participants in cluster 2 (2.86%), p=0.058 (Figure 2c).

257 **3.2.2 1.8FO dataset**

258 Hierarchical clustering using Ward's method was also found to lead to the clearest discrimination of
259 clusters with the 1.8FO dataset, with four clusters found to be optimal (1, n = 82; 2, n = 51; 3, n =
260 112; 4, n = 65) (Figure 3a). Clusters did not segregate clearly upon application of PCA. Due to
261 differences in imputed values between datasets for plasma TG change, a significant difference in TG
262 change between clusters was observed in one of the imputed datasets only [F(3,206)=2.67, p<0.05],
263 with participants in cluster 3 having a mean reduction in plasma TG of -0.247 mmol/L, significantly
264 greater than a mean reduction of -0.052 mmol/L for participants in cluster 1 (p<0.05, Bonferroni
265 corrected) (Figure 3b). The difference in EPA+DHA change between clusters was not significantly
266 different (p=0.073).

267 **3.3 Supervised analysis**

268 **3.3.1 Predicting plasma TG change after 1.8FO**

269 Table 3 presents the average RMSEs from supervised analysis of the 5 imputed datasets. For
270 predicting plasma TG change, the lowest average RMSE across all 5 imputed datasets corresponded
271 to models generated by LASSO. Table 4 presents the mean-centered and standardized shrunk
272 coefficients, averaged across all imputed datasets. In total, 18 predictors were selected across 3 or
273 more imputed datasets. The highest positive coefficient corresponded to baseline plasma insulin
274 concentration, while the highest negative coefficient corresponded to pre-intervention TG
275 concentration. These two variables were also selected by the other supervised analysis methods. For
276 the other numeric predictors, the standardized coefficients were all less than ±0.1, with the next
277 largest coefficients corresponding to baseline LDL and the fatty acid C20:3n-6, both positively
278 associated with TG change; and baseline IL-10 levels, negatively associated with TG change. For the
279 categorical variables, carriers of the T allele for rs1800588, a polymorphism of the LIPC gene, was
280 also positively associated with TG change. Figure 4a shows the relationship between predicted
281 plasma TG change using the LASSO model, and actual plasma TG change, with an R2 upon
282 application to the original (un-imputed) dataset of 0.47. Upon applying Oldham's transformation,
283 Figure 4b shows a significant negative correlation (R = -0.19, p<0.001) between the average of (log-
284 transformed) pre- and post-intervention TG values against observed TG change, indicating that
285 participants with higher pre-intervention plasma TG show greater reduction after 1.8FO, after
286 adjusting for regression to the mean.

287 ### 3.3.2 Predicting plasma PC EPA+DHA change after 0.7FO

288 The lowest average RMSE for predicting plasma PC EPA+DHA change after 0.7FO corresponded to
289 models generated by forward stepwise selection (Table 3). Table 5 shows both the mean-centered
290 coefficients, pooled from the 5 imputed datasets, and standardized coefficients calculated from
291 running the model against the standardized non-imputed dataset, with an adjusted R2 value of 0.32.
292 The final model contained 6 predictors with positive coefficients for age, sex, a SNP in the tumor
293 necrosis factor alpha (TNFα) gene (rs1800629) and pre-intervention PC docosapentaenoic acid
294 (DPA) proportion, and negative coefficients for pre-intervention proportions of EPA and DHA.
295 Figure 5a shows the relationship between predicted and actual EPA+DHA change using the forward
296 stepwise model, with an R2 of 0.33 after application to the un-imputed dataset. After application of
297 Oldham's transformation, Figure 5b shows no relationship between the average of pre- and post-
298 intervention EPA+DHA with observed EPA+DHA change, indicating that the relationship between
299 pre-intervention EPA+DHA and subsequent EPA+DHA change after 0.7FO can be explained by
300 regression to the mean.

301 ### 3.3.3 Predicting plasma PC EPA+DHA change after 1.8FO

302 The lowest average RMSE for predicting plasma PC EPA+DHA change after 1.8FO corresponded to
303 models generated by backward stepwise selection (Table 3). The final model contained 11 predictors
304 with positive coefficients for age, sex and a SNP in the Fatty Acid Binding Protein 1 (FABP1) gene
305 (rs2241883), and negative coefficients for body mass index (BMI) and a number of pre-intervention
306 PC fatty acids, as shown in Table 6. Figure 6a shows the relationship between predicted and actual
307 EPA+DHA change using the backward stepwise model, with an R2 of 0.38 after application to the
308 un-imputed dataset. After application of Oldham's transformation, Figure 6b shows a significant
309 positive correlation ($R = 0.23$, $p<0.001$) between the average of pre- and post-intervention PC
310 EPA+DHA and observed PC EPA+DHA change, meaning that after accounting for regression to the
311 mean, there was a greater change in PC EPA+DHA for participants with higher pre- and post-
312 intervention average PC EPA+DHA proportions.

313 To examine the different results after Oldham's transformation with 0.7FO and 1.8FO more closely,
314 the relationship between pre- and post-intervention PC EPA+DHA with PC EPA+DHA change were
315 examined separately (supplementary Figure 2). For both fish oil doses, there was a negative
316 association between pre-intervention plasma PC EPA+DHA and subsequent PC EPA+DHA change,
317 of a similar magnitude for both fish oil doses (supplementary Figure 1a-1b). However, when
318 comparing post-intervention PC EPA+DHA proportion with PC EPA+DHA change, there was a
319 higher positive correlation after 1.8FO ($R=0.68$, supplementary Figure 1d) than after 0.7FO ($R=0.46$,
320 supplementary Figure 1c), with PC EPA+DHA increase more uniform after 1.8FO than after 0.7FO.

321 ## 4 Discussion

322 Nutrition studies typically reveal substantial heterogeneity in physiological response after an
323 intervention. Studies that collect data on a large array of predictors of response, in a sufficient
324 number of participants, can be utilized to identify potential predictors of this response variability.
325 This is of interest in the growing fields of precision and personalized nutrition, where elucidation of
326 predictors of response may help to identify the characteristics of people most and least likely to
327 respond beneficially. The results of this analysis revealed that the application of variable selection
328 techniques, in particular, can identify new and clinically important predictors that explain between a
329 third to a half of the variability in change in plasma TG and PC EPA+DHA, after an intervention
330 with fish oil. Our predictive models showed greater TG-lowering with lower pre-intervention levels

331 of plasma insulin, LDL cholesterol and C20:3n-6 levels, along with C carriers (compared with T
332 carriers) of the SNP rs1800588; and greater TG-lowering in those with higher pre-intervention levels
333 of plasma TG (additional to regression to the mean) and serum IL-10. For predicting change in
334 plasma PC EPA+DHA, greater increases were observed with higher age and female sex, along with
335 lower levels of baseline plasma C20:5n¬-3 (EPA) and C22:6n-3 (DHA), for both doses of fish oil.
336 However, the relationship between baseline EPA+DHA levels and degree of change differed between
337 the 0.7FO and 1.8FO fish oil interventions, with the relationship for 0.7FO explained by regression to
338 the mean, while increases in EPA+DHA after 1.8FO were more uniform. This means that greater
339 increases in EPA+DHA than expected were observed in those with higher baseline EPA+DHA
340 levels.

341 Change in plasma TG and plasma PC EPA+DHA were the outcomes of interest in this study and
342 were used on a continuous scale rather than being dichotomized into "responders" or "non-
343 responders" to the intervention to maximize statistical power (33,34). Findings from this study
344 identify important physiological predictors of response heterogeneity at a group level for the given
345 outcomes of interest. The final models were generated through application of different variable
346 selection methods – with forward and backward stepwise selection, and LASSO, generating the
347 models with the lowest RMSE for predicting change in plasma TG after 1.8FO and in PC EPA+DHA
348 after 0.7FO and 1.8FO. Stepwise selection methods such as forward and backward stepwise selection
349 have been criticized (35,36) as they are often overfit to training data and undergo lack of validation,
350 or are used as the sole model-building approach. In this study, we mitigated these limitations by
351 using cross-validation to select the final model size, repeating the process across 5 imputed datasets
352 to determine the most appropriate predictors to retain in the final model, and comparing the
353 validation set RMSEs with models generated by other variable selection methods. While cross-
354 validation helps to prevent model overfitting, it will be important to validate these models using
355 external, independent datasets to ascertain whether findings from the FINGEN study are
356 generalizable to other populations (37).

357 The variables selected by LASSO for predicting plasma TG change after 1.8FO (Table 4) included
358 baseline BMI, plasma insulin concentration and saturated fat intake, and pre-intervention LDL-
359 cholesterol concentration, all of which had positive (shrunk) coefficients, meaning that higher values
360 of these predictors were associated with less TG-lowering. Each of these predictors is known to be
361 associated with higher TG concentrations, with obesity and insulin resistance being features of the
362 metabolic syndrome (38). Conversely, other predictors had negative coefficients, including APOE4
363 carriers, meaning this variant was associated with greater plasma TG-lowering than other APOE
364 genotypes. This supports the previous findings from the FINGEN cohort for a non-significant trend
365 in greater TG reductions in APOE4 carriers, with the greatest TG reductions in men carrying APOE4
366 (15). Baseline concentration of plasma interleukin 10 (IL-10) and self-reported fruit consumption
367 were also among the predictors with negative coefficients; higher values of both are associated with
368 better health status, and these participants were more likely to show falls in plasma TG in response to
369 the intervention. Apart from the association of higher pre-intervention plasma TG concentration with
370 greater TG-lowering, the variables selected by LASSO suggest that participants with a profile
371 indicative of lower heart disease risk are more likely to have greater plasma TG-lowering after 1.8
372 g/d EPA+DHA.

373 Participants who were older and female tended to have the greatest increases in plasma PC
374 EPA+DHA (Table 5, 6), confirming findings from a previous study (39). For change after 1.8FO
375 only, higher BMI was associated with a lower increase in PC EPA+DHA, in line with previous
376 findings (39). For predicting PC EPA+DHA change after 1.8FO, higher pre-intervention levels of the

377 saturated fatty acids palmitic (C16:0) and stearic acid (C18:0), the trans fatty acid vaccenic acid
378 (C18:1n-7) and the unsaturated fatty acids linoleic acid (C18:2n-6) and arachidonic acid (C20:4n-6)
379 were associated with a lesser increase in PC EPA+DHA (Table 6), which has, to the best of our
380 knowledge, not been reported before. On the other hand, for the model predicting PC EPA+DHA
381 change after 0.7FO, a higher proportion of DPA in plasma PC was associated with greater increases
382 in PC EPA+DHA in response to supplementation. As desaturation of DPA leads to the formation of
383 DHA, DHA levels are likely to increase if more DPA is available (40), and DPA supplementation has
384 been shown to increase DHA levels in plasma TG (41). As plasma PC fatty acid proportions were
385 included in this analysis, this suggests that lower levels of other fatty acids will enable EPA+DHA to
386 form a greater proportion of total plasma PC fatty acids. Unsurprisingly, higher pre-intervention
387 concentrations of EPA (C20:5n-3) and DHA (C22:6n-3) were associated with a smaller increase in
388 PC EPA+DHA after both fish oil interventions, as has been observed previously (39). The
389 standardized coefficients for pre-intervention EPA were approximately twice as large as the
390 coefficients for DHA (Table 5, 6), suggesting that EPA status was a more important predictor of
391 incorporation of EPA+DHA into PC. This makes sense given that DHA is a downstream metabolite
392 of EPA (40). Interestingly, different results were observed upon applying Oldham's transformation to
393 EPA+DHA change after each fish oil intervention. As the relationship between the average of pre-
394 and post-intervention EPA+DHA with EPA+DHA change was not significant for 0.7FO, this
395 suggests the relationship can be explained by regression to the mean. However, the significant
396 positive association that remained after 1.8FO suggests that a greater increase in EPA+DHA occurred
397 than would be expected in those with higher pre-intervention EPA+DHA. This finding supports a
398 lack of a "ceiling effect", meaning higher pre-intervention plasma PC EPA+DHA levels do not limit
399 further increases in EPA+DHA in response to supplementation. The findings of the JELIS trial lend
400 support to this claim, where Japanese participants had a reduction in coronary events after EPA
401 supplementation, despite high habitual consumption of fish and thus high pre-intervention plasma LC
402 n-3 PUFA status (42).

403 A strength of this analysis approach was the use of a large dataset with many variables, with the
404 potential to uncover new variables associated with change in plasma TG and PC EPA+DHA levels.
405 Furthermore, the crossover design enabled analyses to be performed on the same volunteers, enabling
406 better comparisons to be made between the results for EPA+DHA change after both 0.7FO and
407 1.8FO. However, the analysis may have been limited by the statistical power of the dataset, with a
408 large number of predictors considered in relation to the number of participants. Despite this, the
409 supervised analysis methods applied in this paper were suitable for use on high-dimensional datasets,
410 where the power is even smaller due to the number of predictors being greater than the number of
411 volunteers (27). These types of dataset are increasingly common in an era of precision medicine,
412 where information on an array of markers including genotype, metabolomics and microbiome are
413 increasingly collected (1,43). While limiting the number of variables considered in this analysis
414 would have improved statistical power, this would not have made full use of the dataset, nor enabled
415 potential discovery of new predictors of response to the outcomes of interest. Using validation
416 approaches such as cross-validation to determine the size of models selected, and performing
417 analyses across 5 imputed datasets, also increased the likelihood that models contained relevant
418 variables, as final models considered variables that were only in common across at least 3 of the 5
419 imputed datasets.

420 In conclusion, the application of supervised analysis approaches, particularly variable selection
421 methods, led to the identification of new variables for predicting change in plasma TG and plasma
422 PC EPA+DHA after fish oil supplementation. This means that females and those who are older are
423 more likely to benefit from fish oil supplements in terms of increasing the omega-3 index. In

424 addition, those with higher levels of plasma TG and certain inflammatory markers, together with
425 lower levels of plasma insulin, LDL cholesterol, C20:3n-6, and saturated fat consumption, are more
426 likely to benefit from fish oil supplements in terms of TG lowering, based on the results of this study.
427 A similar analysis approach applied to data from other large fish oil supplementation studies could
428 provide an external validation of our models, or help to identify additional markers of response. Our
429 study highlights how application of appropriate statistical methods to rich datasets can develop our
430 knowledge of the factors underpinning physiological response heterogeneity to interventions, and
431 hence provide a useful tool for precision nutrition and in the future tailoring of dietary
432 recommendations.

433

434 **Abbreviations:**

435 0.7FO - 0.7 g/d EPA+DHA from fish oil
436 1.8FO - 1.8 g/d EPA+DHA from fish oil
437 APOE(4) - apolipoprotein E(4)
438 DPA - docosapentaenoic acid
439 FABP1 - Fatty Acid Binding Protein 1
440 FO – fish oil
441 LASSO - Least Angle Selection and Shrinkage Operator
442 LC n-3 PUFAs - long-chain n-3 polyunsaturated fatty acids
443 PC - plasma phosphatidylcholine
444 PCA - principal components analysis
445 RMSE - root mean squared error
446 TG – triglycerides

447 **Acknowledgments**

448 Authors would like to thank Mintu Nath for additional statistical support.

449 **Author Contributions**

450 TITP, AJW, EHZ, PLZ and BdR conceptualized and designed the research; TITP conducted the
451 research, analyzed the data and wrote the paper; AMM, PCC, JCM designed and conducted the
452 original FINGEN study. All authors reviewed and approved the final manuscript.

453 **Funding**

454 The FINGEN study was funded by the UK Food Standards Agency (grant no. RRD7/N02/A). TP is
455 sponsored by Biotechnology and Biological Sciences Research Council (BBSRC) and Unilever
456 Foods Innovation Centre, Wageningen, The Netherlands (Collaborative Training Partnership [CTP]).
457 The research of BdR is funded by the Scottish Government's Rural and Environment Science and
458 Analytical Services Division (RESAS).

459 **Conflict of Interest**

460 AJW and EHZ are employees of Unilever Foods Innovation Centre Wageningen, which markets food
461 products.

462 **Data Availability Statement**

463 Analytic code can be made available upon request to the first author pending application and
464 approval.

465

466 **References**

467 1. Berry SE, Valdes AM, Drew DA, Asnicar F, Mazidi M, Wolf J, et al. Human postprandial
468 responses to food and potential for precision nutrition. Nat Med. 2020;26:964-73.
469 2. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized
470 Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079–94.
471 3. Minihane AM. Fatty acid–genotype interactions and cardiovascular risk. Prostaglandins Leukotr
472 Essent Fatty Acids. 2010;82(4):259–64.
473 4. de Roos B, Brennan L, de Roos B, Brennan L. Personalised Interventions—A Precision
474 Approach for the Next Generation of Dietary Intervention Studies. Nutrients. 2017;9(8):847.
475 5. Matusheski NV, Caffrey A, Christensen L, Mezgec S, Surendran S, Hjorth MF, et al. Diets,
476 nutrients, genes and the microbiome: recent advances in personalised nutrition. Br J Nutr.
477 2021;126:1489-97.
478 6. Potter T, Vieira R, de Roos B. Perspective: Application of N-of-1 Methods in Personalized
479 Nutrition Research. Adv Nutr. 2021;12(3):579-89.
480 7. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction:
481 advantages, problems, and suggested solutions. Cancer Treat Rep. 1985;69:1071–7.
482 8. Garcia-Perez I, Posma JM, Chambers ES, Mathers JC, Draper J, Beckmann M, et al. Dietary
483 metabotype modelling predicts individual responses to dietary interventions. Nat Food.
484 2020;1(6):355–64.
485 9. Madden J, Williams CM, Calder PC, Lietz G, Miles EA, Cordell H, et al. The Impact of
486 Common Gene Variants on the Response of Biomarkers of Cardiovascular Disease (CVD) Risk
487 to Increased Fish Oil Fatty Acids Intakes. Annu Rev Nutr. 2011;31:203–34.
488 10. Celis-Morales C, Livingstone KM, Marsaux CFM, Forster H, O'Donovan CB, Woolhead C, et
489 al. Design and baseline characteristics of the Food4Me study: a web-based randomised
490 controlled trial of personalised nutrition in seven European countries. Genes Nutr.
491 2015;10(1):450.
492 11. Walker RE, Jackson KH, Tintle NL, Shearer GC, Bernasconi A, Masson S, et al. Predicting the
493 effects of supplemental EPA and DHA on the omega-3 index. Am J Clin Nutr.
494 2019;110(4):1034–40.
495 12. Balk EM, Lichtenstein AH, Chung M, Kupelnick B, Chew P, Lau J. Effects of omega-3 fatty
496 acids on serum markers of cardiovascular disease risk: A systematic review. Atherosclerosis.
497 2006;189(1):19–30.
498 13. Harris WS, von Schacky C. The Omega-3 Index: a new risk factor for death from coronary heart
499 disease? Prev Med. 2004;39(1):212–20.
500 14. Harris WS, Del Gobbo L, Tintle NL. The Omega-3 Index and relative risk for coronary heart
501 disease mortality: Estimation from 10 cohort studies. Atherosclerosis. 2017;262:51–4.
502 15. Caslake MJ, Miles EA, Kofler BM, Lietz G, Curtis P, Armah CK, et al. Effect of sex and
503 genotype on cardiovascular biomarker response to fish oils: the FINGEN Study. Am J Clin Nutr.
504 2008;88(3):618–29.

505 16. Kofler BM, Miles EA, Curtis P, Armah CK, Tricon S, Grew J, et al. Apolipoprotein E genotype
506 and the cardiovascular disease risk phenotype: Impact of sex and adiposity (the FINGEN study).
507 Atherosclerosis. 2012;221(2):467–70.
508 17. van der Veen JN, Kennelly JP, Wan S, Vance JE, Vance DE, Jacobs RL. The critical role of
509 phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease. Biochim
510 Biophys Acta Biomembr. 2017;1859(9B):1558–72.
511 18. Browning LM, Walker CG, Mander AP, West AL, Madden J, Gambell JM, et al. Incorporation
512 of eicosapentaenoic and docosahexaenoic acids into lipid pools when given as supplements
513 providing doses equivalent to typical intakes of oily fish. Am J Clin Nutr. 2012;96(4):748–58.
514 19. Enders CK, Tofighi D. Centering predictor variables in cross-sectional multilevel models: A new
515 look at an old issue. Psychol Methods. 2007;12(2):121-138.
516 20. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in
517 R. J Stat Softw 2011;45(3):1–67.
518 21. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332:1080.
519 22. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple
520 regression: a bad idea. Stat Med 2006;25(1):127–41.
521 23. Haverkamp N, Beauducel A. Violation of the Sphericity Assumption and Its Effect on Type-I
522 Error Rates in Repeated Measures ANOVA and Multi-Level Linear Models (MLM). Front
523 Psychol. 2017;8:1841.
524 24. Swinton PA, Hemingway BS, Saunders B, Gualano B, Dolan E. A Statistical Framework to
525 Interpret Individual Response to Intervention: Paving the Way for Personalized Nutrition and
526 Exercise Prescription. Front Nutr. 2018;5:41.
527 25. Kaufman L, Rousseeuw PJ. Partitioning Around Medoids (Program PAM). In: Finding Groups
528 in Data: An Introduction to Cluster Analysis. Hoboken: John Wiley & Sons, Inc; 1990. p. 68-
529 125.
530 26. Laan MV der, Pollard K, Bryan J. A new partitioning around medoids algorithm. J Stat Comput
531 Simul. 2003;73(8):575–84.
532 27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining,
533 Inference, and Prediction. 2nd ed. New York: Springer; 2009.
534 28. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdiscip
535 Rev Data Min Knowl Discov. 2012;2(1):86–97.
536 29. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with
537 Applications in R. New York: Springer; 2013.
538 30. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol.
539 1996;58(1):267–88.
540 31. Kursa M, Rudnicki W. Feature Selection with Boruta Package. J Stat Softw 2010;36:1–13.
541 32. Gill JS, Beevers DG, Zezulka AV, Davies P. Relation between initial blood pressure and its fall
542 with treatment. Lancet 1985;325(8428):567–9.
543 33. Atkinson G, Williamson P, Batterham AM. Issues in the determination of 'responders' and 'non-
544 responders' in physiological research. Exp Physiol. 2019;104(8):1215-25.
545 34. Senn S. Statistical pitfalls of personalized medicine. Nature. 2018;563:619–21.
546 35. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise
547 modelling in ecology and behaviour? J Anim Ecol 2006;75(5):1182–9.
548 36. Heinze G, Wallisch C, Dunkler D. Variable selection – A review and recommendations for the
549 practicing statistician. Biom J. 2018;60(3):431–49.
550 37. Ivanescu AE, Li P, George B, Brown AW, Keith SW, Raju D, et al. The importance of
551 prediction model validation and assessment in obesity and nutrition research. Int J Obes.
552 2016;40(6):887–94.
553 38. Eckel RH, Grundy SM, Zimmet PZ. The metabolic syndrome. Lancet 2005;365(9468):1415–28.

554 39. Flock MR, Skulas-Ray AC, Harris WS, Etherton TD, Fleming JA, Kris-Etherton PM.
555 Determinants of Erythrocyte Omega-3 Fatty Acid Content in Response to Fish Oil
556 Supplementation: A Dose–Response Randomized Controlled Trial. J Am Heart Assoc.
557 2013;2(6):e000513.
558 40. de Roos B, Mavrommatis Y, Brouwer IA. Long-chain n-3 polyunsaturated fatty acids: new
559 insights into mechanisms relating to inflammation and coronary heart disease. Br J Pharmacol.
560 2009;158:413–28.
561 41. Miller E, Kaur G, Larsen A, Loh SP, Linderborg K, Weisinger HS, et al. A short-term n-3 DPA
562 supplementation study in humans. Eur J Nutr. 2013;52:895–904.
563 42. Yokoyama M, Origasa H, Matsuzaki M, Matsuzawa Y, Saito Y, Ishikawa Y, et al. Effects of
564 eicosapentaenoic acid on major coronary events in hypercholesterolaemic patients (JELIS): a
565 randomised open-label, blinded endpoint analysis. Lancet. 2007;369(9567):1090–8.
566 43. Eetemadi A, Rai N, Pereira BMP, Kim M, Schmitz H, Tagkopoulos I. The Computational Diet:
567 A Review of Computational Methods Across Diet, Microbiome, and Health. Front Microbiol.
568 2020;11:393.
569

570 **Supplementary material**

571 The Supplementary Material for this article, including a participant flow chart (Supplementary
572 figure) and a list of all SNPs included in the dataset (Supplementary table) can be found online at:
573 https://www.frontiersin.org/articles/xxx

574

575 **Table 1. Mean change (SD) in plasma TG and plasma PC EPA+DHA in response to fish oil**
576 **supplementation.**

Outcome	Treatment arm	Mean change (SD)
Change in plasma TG (mmol/l) between start and end of 8-week intervention	0.7 g/day EPA+DHA	-0.083 (0.428)
	1.8 g/day EPA+DHA	-0.152 (0.499)
	control oil	0.011 (0.460)
Change in plasma PC EPA+DHA (% of total fatty acids) between start and end of 8-week intervention	0.7 g/day EPA+DHA	3.05 (1.70)
	1.8 g/day EPA+DHA	4.65 (2.28)
	control oil	-0.089 (1.40)

577

578 **Table 2.** Bonferroni-adjusted pairwise comparisons after repeated measures ANOVA for differences
579 in plasma TG change and plasma PC EPA+DHA change between intervention groups.

Outcome/test	**Mean difference**	**Test statistic**	**Bonferroni-adjusted *p*-value**
Change in plasma TG between start and end of 8-week intervention (mmol/L)			
0.7 g/d EPA+DHA – control oil	-0.095	-2.594	0.0298
1.8 g/d EPA+DHA – control oil	-0.163	-4.162	0.0001
1.8 g/d EPA+DHA - 0.7 g/d EPA+DHA	-0.069	-2.082	0.1144
Change in plasma PC EPA+DHA between start and end of 8-week intervention (% of total fatty acids)			
0.7 g/d EPA+DHA – control oil	3.139	25.44	<0.0001
1.8 g/d EPA+DHA – control oil	4.740	31.45	<0.0001
1.8 g/d EPA+DHA - 0.7 g/d EPA+DHA	1.601	12.32	<0.0001

580

581 **Table 3**. Model RMSEs after application of supervised analysis methods to the outcomes plasma TG
582 change after 1.8 g/d EPA+DHA, plasma PC EPA+DHA change after 0.7 g/d EPA+DHA, and plasma
583 PC EPA+DHA change after 1.8 g/d EPA+DHA. Lowest RMSEs for each outcome are given in bold.

Outcome	**Plasma TG change after 1.8 g/d EPA+DHA**	**Plasma PC EPA+DHA change after 0.7 g/d EPA+DHA**	**Plasma PC EPA+DHA change after 1.8 g/d EPA+DHA**
Method	**Mean RMSE (SD), 5 imputed datasets**		
Forward stepwise	0.396 (0.006)	**1.470** (0.024)	1.982 (0.032)
Backward stepwise	0.400 (0.010)	1.488 (0.015)	**1.966** (0.013)
LASSO	**0.353** (0.058)	1.521 (0.051)	2.059 (0.170)
Boruta – test set RMSE	0.452 (0.064)	1.610 (0.127)	2.177 (0.106)

584

585 **Table 4.** Shrunk coefficients after LASSO analysis for predicting plasma TG change after 1.8 g/d
586 EPA+DHA.

Variable name	**Mean-centered coefficient (SD)**	**Standardized coefficient**
Intercept	-0.330 (0.103)	0
APOE – APOE4 variant	-0.010 (0.006)	
Baseline BMI (kg/m^2)	0.002 (0.001)	0.017
Baseline CRP	0.005 (0.002)	0.030
Baseline plasma insulin (mmol/L)	0.014 (0.003)	0.118
Baseline IL-10	-0.007 (0.002)	-0.045
Baseline VCAM-1	<-0.001	-0.030
Pre-intervention plasma TG (mmol/L)	-0.442 (0.048)	-0.577
Pre-intervention LDL-cholesterol (mmol/L)	0.035 (0.006)	0.066
Fruit consumption (g)	<0.001	-0.011
Saturated fat consumption (g)	0.001 (0.001)	0.040
rs320 (G>T)	-0.015 (0.004)	
rs2250656 (C>T)	-0.017 (0.009)	
rs1800588 (T>C)	0.058 (0.031)	
rs1800795 (C>G)	0.024 (0.012)	
rs1800896 (C>T)	0.015 (0.009)	
rs5370 (T>G)	0.054 (0.030)	
C20:3*n*-6	0.027 (0.012)	0.049
C20:4*n*-6	0.006 (0.002)	0.024

587 Variables listed were selected by 3 or more of the 5 imputed datasets, and depict the mean (SD) of their shrunk
588 coefficients across all imputed datasets for which they were selected. Both mean-centered (left) and standardized (right,
589 variables on continuous numeric scale only) shrunk coefficients are presented. APOE/APOE4 - apolipoprotein E3/E4 or
590 E4/E4, CRP – C-reactive protein, IL-10 – interleukin 10, LDL – low-density lipoprotein, TG – triglyceride, VCAM-1 –
591 vascular cell adhesion protein 1

592 **Table 5.** Model output after performing forward stepwise regression for predicting plasma PC EPA+DHA change after 0.7 g/d EPA+DHA.

Pooled mean centered regression coefficients					**Standardized regression coefficients, un-imputed dataset**				
Term	**Estimate**	**Std. error**	**Test statistic**	***p***	**Term**	**Estimate**	**Std. error**	**Test statistic**	***p***
Intercept	2.536	0.129	19.61	<0.001	Intercept	2.686	0.119	22.49	<0.001
Age	0.021	0.006	3.280	0.001	Age	0.281	0.085	3.300	0.001
Sex – Female	0.681	0.165	4.139	<0.001	Sex – Female	0.694	0.170	4.094	<0.001
rs1800629 – G/A	0.400	0.178	2.243	0.026	*rs1800629*	0.230	0.082	2.806	0.005
rs1800629 – A/A	0.649	0.337	1.926	0.055	(G>A)				
C20:5*n*-3	-0.859	0.118	-7.285	<0.001	C20:5*n*-3	-0.727	0.102	-7.119	<0.001
C22:5*n*-3	1.514	0.346	4.371	<0.001	C20:5*n*-3	0.376	0.091	4.124	<0.001
C22:6*n*-3	-0.247	0.077	-3.206	0.001	C20:5*n*-3	-0.325	0.101	-3.218	0.001

593 Data showing mean-centered regression coefficients pooled across all imputed datasets (left), and upon applying the model to the standardized un-imputed dataset (right,
594 continuous numeric scale variables standardized only).

595

596 **Table 6** Model output after performing backward stepwise regression for predicting plasma PC EPA+DHA change after 1.8 g/d EPA+DHA.

Pooled mean centered regression coefficients					**Standardized regression coefficients, un-imputed dataset**				
Term	**Estimate**	**Std. error**	**Test statistic**	***p***	**Term**	**Estimate**	**Std. error**	**Test statistic**	***p***
Intercept	3.915	0.193	20.27	0	Intercept	4.235	0.157	26.95	<0.001
Age	0.043	0.009	4.777	<0.001	Age	0.563	0.115	4.897	<0.001
Sex – Female	0.799	0.224	3.572	<0.001	Sex – Female	0.774	0.224	3.451	0.001
BMI	-0.088	0.035	-2.537	0.012	BMI	-0.320	0.118	-2.716	0.007
rs2241883 – T/C	0.564	0.229	2.462	0.014	*rs2241883*	0.323	0.107	3.012	0.003
rs2241883 – C/C	0.806	0.343	2.350	0.019	(T>C)				
C16:0	-0.429	0.109	-3.922	<0.001	C16:0	-0.844	0.219	-3.852	<0.001
C18:0	-0.281	0.109	-2.572	0.011	C18:0	-0.496	0.197	-2.515	0.012
C18:1*n*-7	-0.350	0.120	-2.904	0.004	C18:1*n*-7	-0.488	0.173	-2.813	0.005
C18:2*n*-6	-0.454	0.091	-5.009	<0.001	C18:2*n*-6	-1.304	0.263	-4.966	<0.001
C20:4*n*-6	-0.491	0.111	-4.408	<0.001	C20:4*n*-6	-0.903	0.211	-4.287	<0.001
C20:5*n*-3	-1.670	0.202	-8.275	<0.001	C20:5*n*-3	-1.337	0.163	-8.217	<0.001
C22:6*n*-3	-0.548	0.112	-4.882	<0.001	C22:6*n*-3	-0.702	0.142	-4.935	<0.001

597 Data showing mean-centered regression coefficients pooled across all imputed datasets (left), and upon applying the model to the standardized un-imputed dataset (right,
598 continuous numeric scale variables standardized only).

Figure legends

Figure 1 Overview of analysis pipeline

Figure 2 Cluster plots of datasets containing baseline variables and data collected prior to intervention with 0.7 g/d EPA+DHA. Each participant is displayed as one data point, by visualizing the clusters using the first of the imputed datasets. a visualization of hierarchical clusters, cluster 1 ○ (black, n = 161), cluster 2 Δ (gray, n = 149); b PCA plot of pre-0.7 g/d data visualizing clusters across the first two principal components (clusters as described in a); c clustering as shown in a with gradation of shading relating to change in plasma PC EPA+DHA (as % of total fatty acids) after 0.7 g/d EPA+DHA intervention, with darker shading corresponding to greatest increases in EPA+DHA. Legend in top right shows range of EPA+DHA change. PC - plasma phosphatidylcholine; PCA – principal components analysis.

Figure 3 Cluster plots of datasets containing baseline variables and data collected prior to intervention with 1.8 g/d EPA+DHA. Each participant is displayed as one data point a visualization of hierarchical clusters using the first imputed dataset, cluster 1 ○ (white, n = 82), cluster 2 Δ (black, n = 51), cluster 3 □ (light gray, n = 112); cluster 4 + (dark gray, n = 65); b visualization of hierarchical clusters using the fourth imputed dataset, with gradation of shading relating to change in plasma TG concentration (mmol/L) after 1.8 g/d EPA+DHA intervention, with lightest shading corresponding to greatest reductions in plasma TG concentration. Legend in top right shows range of plasma TG change. TG – triglyceride.

Figure 4 Graphs depicting results from supervised analysis with plasma TG change after 1.8 g/d EPA+DHA as intervention. a scatter plot comparing actual TG change against predicted TG change using the LASSO model, averaged across all imputed datasets; b scatter plot depicting the correlation between the average of logged plasma TG values pre- and post-1.8g/d EPA+DHA intervention with observed TG change. Dashed line represents no change. LASSO - Least Angle Selection and Shrinkage Operator; TG – triglyceride.

Figure 5 Graphs depicting results from supervised analysis with plasma PC EPA+DHA change after 0.7 g/d EPA+DHA intervention. a scatter plot comparing actual PC EPA+DHA change against predicted change using the final forward stepwise model; b scatter plot depicting the correlation between the average of pre- and post-intervention plasma PC EPA+DHA proportion against observed change in EPA+DHA proportions. Dashed line represents no change. PC - plasma phosphatidylcholine.

Figure 6 Graphs depicting results from supervised analysis with plasma PC EPA+DHA change after 1.8 g/d EPA+DHA intervention. a scatter plot comparing actual PC EPA+DHA change against predicted change using the final backward stepwise model; b scatter plot depicting the correlation between the average of pre- and post-intervention plasma PC EPA+DHA proportion against observed change in EPA+DHA proportions. Dashed line represents no change. PC - plasma phosphatidylcholine.