Predicting fatty acid profiles in blood based on food intake and the FADS1 rs174546 SNP

Article

Accepted Version

Hallmann, J., Kolossa, S., Celis-Morales, C., Forster, H., O'Donovan, C. B., Woolhead, C. , Macready, A. L., Fallaize, R., Marsaux, C. F. M., Tsirigoti, L., Efstathopoulou, E., Moschonis, G., Navas-Carretero, S., San-Cristobal, R., Godlewska, M., Surwiłło, A., Mathers, J. C., Gibney, E. R., Brennan, L., Walsh, M. C., Lovegrove, J. A., Saris, W. H. M., Manios, Y., Martinez, J. A., Traczyk, I., Gibney, M. J. and Daniel, H. (2015) Predicting fatty acid profiles in blood based on food intake and the FADS1 rs174546 SNP. Molecular Nutrition & Food Research, 59 (12). pp. 2565-2573. ISSN 1613-4125 doi: https://doi.org/10.1002/mnfr.201500414

Predicting fatty acid profiles in blood based on food intake and the FADS1 rs174546 SNP

AUTHOR NAMES

Jacqueline Hallmann[1*], Silvia Kolossa[1*], Kurt Gedrich[1], Carlos Celis-Morales[2], Hannah Forster[3], Clare B O'Donovan[3], Clara Woolhead[3], Anna L Macready[4], Rosalind Fallaize[4], Cyril F M Marsaux[5], Christina-Paulina Lambrinou[6], Christina Mavrogianni[6], George Moschonis[6], Santiago Navas-Carretero[7], Rodrigo San-Cristobal[7], Magdalena Godlewska[8], Agnieszka Surwiłło[8], John C Mathers[2], Eileen R Gibney[3], Lorraine Brennan[3], Marianne C Walsh[3], Julie A Lovegrove[4], Wim H M Saris[5], Yannis Manios[6], J Alfredo Martinez[7], Iwona Traczyk[8], Michael J Gibney[3], and Hannelore Daniel[1], on behalf of the Food4Me Study.

AUTHOR AFFILIATIONS

[1] *ZIEL Research Center of Nutrition and Food Sciences, Biochemistry Unit, Technische Universität München, Germany.*

[2] *Human Nutrition Research Centre, Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK.*

[3] *UCD Institute of Food and Health, University College Dublin, Belfield, Dublin 4, Republic of Ireland.*

[4] *Hugh Sinclair Unit of Human Nutrition and Institute for Cardiovascular and Metabolic Research, University of Reading, Reading, UK.*

[5] *Department of Human Biology, NUTRIM, School for Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre, Maastricht, The Netherlands*

[6] *Department of Nutrition and Dietetics, Harokopio University, Athens, Greece.*

[7] *Department of Nutrition, Food Science and Physiology, University of Navarra; CIBER Fisiopatología Obesidad y Nutrición (CIBERobn), Instituto de Salud Carlos III, Spain (SN-C & JAM)*

[8] *National Food & Nutrition Institute (IZZ), Poland.*

* *JH and SK are shared first author*

Received: 28-May-2015; Revised: 30-Jul-2015; Accepted: 20-Aug-2015

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mnfr.201500414.

CORRESPONDING AUTHOR

M.Sc. Silvia Kolossa

ZIEL Research Center of Nutrition and Food Sciences

Biochemistry Unit

Technische Universität München

Gregor-Mendel-Str. 2

85354 Freising

silvia.kolossa@tum.de

Tel: +49 8161 71 2329

Fax: +49 8161 71 3999

KEYWORDS

FADS1, fatty acids, n-6 FA, n-3 FA, blood marker prediction

ABBREVIATIONS

AA, arachidonic acid; BMI, Body Mass Index; bolasso, bootstrapped LASSO; DGLA, dihomo-γ-linolenic acid; DHA, docosahexaenoic acid; DPA, docosapentaenoic acid; DBS, Dried Blood Spots; EER, estimated energy requirement; EPA, eicosapentaenoic acid; ETA, eicosatetraenoic acid; FA, fatty acid; FADS1, fatty acid desaturase 1; FAME, fatty acid methyl esters; FFQ, Food Frequency Questionnaire; LASSO, Least Absolute Shrinkage and Selection Operator; LC-PUFA, long-chain polyunsaturated fatty acids; MPE, mean prediction error; MSPE, mean squared prediction error; MHT, multiple hypothesis testing; PAL, physical activity level; PUFA, polyunsaturated fatty acids; SNP, single nucleotide polymorphism; SBER, squared base error rate

Abstract

Scope

A high intake of n-3 PUFA provides health benefits via changes in the n-6/n-3 ratio in blood. In addition to such dietary PUFAs, variants in the fatty acid desaturase 1 (FADS1) gene are also associated with altered PUFA profiles.

Methods and results

We used mathematical modelling to predict levels of PUFA in whole blood, based on MHT and bolasso selected food items, anthropometric and lifestyle factors, and the rs174546 genotypes in *FADS1* from 1,607 participants (Food4Me Study). The models were developed using data from the first reported time point (training set) and their predictive power was evaluated using data from the last reported time point (test set). Amongst other food items, fish, pizza, chicken and cereals were identified as being associated with the PUFA profiles. Using these food items and the rs174546 genotypes as predictors, models explained 26% to 43% of the variability in PUFA concentrations in the training set and 22% to 33% in the test set.

Conclusions

Selecting food items using MHT is a valuable contribution to determine predictors, as our models' predictive power is higher compared to analogue studies. As unique feature, we additionally confirmed our models' power based on a test set.

1. Introduction

An adequate dietary intake of long chain polyunsaturated fatty acids (LC-PUFAs) is recommended to protect from a variety of diseases mainly via the antagonism of the series 3 and series 2 products of the cyclooxygenase and lipoxygenase pathways [1][2]. The pattern of LC-PUFA levels in blood is mainly determined based on the composition of membrane phospholipids of blood cells. Their constituent FAs are derived from both dietary sources and endogenous synthesis catalyzed by desaturases and elongases. It is also discussed that physical activity, gender, age, body mass index (BMI) and smoking affect the fatty acid (FA) profiles in blood [1][3][4]. One of the key enzymes in the endogenous synthesis pathways is fatty acid desaturase 1 (FADS1). This enzyme introduces a double bond at the Δ5-position in a 20-carbon FA chain and thus, catalyzes the conversion of eicosatetraenoic acid (ETA) into eicosapentaenoic acid (EPA) and dihomo-γ-linolenic acid (DGLA) into arachidonic acid (AA), respectively. Single Nucleotide Polymorphisms (SNPs) in the *FADS1* gene have been shown to affect the FAs concentrations in blood. There is a prominent SNP, rs174546, in *FADS1* for which T homozygotes have significantly higher concentrations of α-linolenic, linoleic and DGLA than the C homozygotes. Concentrations of AA were significantly lower and docosahexaenoic acid (DHA) levels remained unaffected in the T homozygotes [5][6]. Some controversial findings concerning EPA levels were described with either significantly lower levels in C homozygotes [5] or with an unaltered status [6].

Previous studies aimed at linking selected lifestyle factors and *FADS1* SNPs to PUFA concentrations in blood and even using these selected factors for the prediction of certain PUFA concentrations by defining linear models [6][7][8][9][10]. For the selection of important predictors out of a large set of possible entities, further methods for model development have been introduced recently. As data density increases, Type I error rate also increases and turns multiple hypothesis testing (MHT) for predictor selection [11] into a challenge [12]. While the control of Type I error rate is essential, loss of power needs to be taken into consideration as well [13]. Therefore, methods like Least Absolute

Shrinkage and Selection Operator (LASSO) regression were developed. Rohart et al. [14], for example, found that results from LASSO combined with bootstrapping for robustness, so called bolasso, improved the accuracy of predictions. They aimed at predicting a given phenotype from a metabolomics data set and compared LASSO to bolasso performance, concluding that bolasso provided smaller mean squared prediction errors (MSPE). Lampos et al. [15] successfully used bolasso variable selection to extract information from Twitter for fitting of a regression model to predict flu waves.

Our aim is to develop models with the *FADS1* gene variants and selected foods including supplements as well as physical activity, gender, age, BMI and smoking as predictors for whole blood concentrations of DGLA, AA, EPA, DHA and DPA whose predictions are as accurate as possible. To identify nutritional predictors we examine the applicability of MHT and bolasso. By re-evaluating those models using a different data set we make a first step towards investigating the performance of the developed models.

2. Materials and Methods

Data collection

Data used for model development and testing were collected within the Food4Me Proof-of-Principle study, a pan-European project on personalized nutrition. The study was a web-based randomized controlled trial conducted in 7 countries (Germany, Greece, Ireland, the Netherlands, Poland, Spain, United Kingdom) with 1,607 participants enrolled [16]. Food intake was recorded online via a validated Food Frequency Questionnaire (FFQ), reflecting the participants' diet over one month [17][18]. 137 food items from the originally 162 items in the FFQ were pre-selected as potential predictors by removing country-specific food items like 'stroopwafle'. Additionally, data on supplement intake was collected. Each participant was asked to fill in 3 FFQs over a period of 6 months. Each FFQ was accompanied by a Baecke questionnaire to estimate the participants' Physical

Activity Level (PAL) online during the last month [19]. Participants additionally provided capillary whole blood samples using Dried Blood Spots (DBS) cards, a minimal-invasive, rapid and reliable method of FA quantification [20][21]. The samples were analyzed by Vitas Ltd, Oslo, Norway via GC-MS. Individual FAs in blood samples are expressed as relative concentration i.e. percentage of total fatty acid methyl esters (% FAME). Buccal cell samples from participants were collected and the *FADS1* rs174546 genotyped using the KASP^{TM} assay, performed by LCG Genomics, Hertfordshire, United Kingdom. Details on analyses by Vitas Ltd. and LCG Genomics, as well as the study protocol and other measurements are described elsewhere [16]. Figure 1 gives an overview of the data selection for the different approaches used in the present study.

FADS1 characteristics and relation to FAs concentrations

A Hardy Weinberg Equilibrium for the distribution of the *FADS1* SNP was tested for the training and test data set (Figure 1) by using the X^2 test and the HWExact test using the R package 'HardyWeinberg'. Data in the training set were analyzed by ANOVA followed by non-orthogonal planned contrasts to determine relative FAs concentrations according to the *FADS1* genotype. The contrasts compared the levels of DGLA, AA, EPA, DHA and DPA among the genotypes (C homozygotes versus T homozygotes and C homozygotes versus heterozygotes).

Model selection

The modeling was conducted in 3 steps and for all analyses described, food intake in g/day was standardized for comparability by subtraction of the mean and division by the respective standard deviation. The first step aims at finding those out of the 137 pre-selected food items that are significantly associated with DGLA, AA, EPA, DHA and DPA; bolasso regression and MHT were used for each selected FA on the predictor selection data set (Figure 1) in an exploratory fashion. All cases with a reported energy intake exceeding the EER by more than 30% were excluded, as unrealistic values of energy intake, in this case lowest energy intake/EER ratio with 0.2 and highest with 5, and

hence unrealistic values of the intake of certain food items might lead to biased food predictions and biased models. Food items that appeared in more than 95% of the bootstraps of bolasso were considered further. MHT was conducted on the same data set to get further insight into possibly important foods. For those two methods the R package 'mht' and the functions 'bolasso', with 1000 bootstraps and a positive regularization sequence value of .02, and 'mht' were used. The function 'mht' was implemented with maximally 30 variables to be ordered and the maximum number of hypotheses testing set to 5.

In step two, models were fitted on a training data set (Figure 1) using the food items selected by MHT, next to PAL, gender, age, BMI, supplementation of unsaturated FA and *FADS1*. Supplementation data was included as dummy variable for taking any supplements with unsaturated FAs as ingredients. Afterwards, models with the food items selected by bolasso were fitted. The adjusted R^2 of those two models were compared. When the adjusted R^2 values were very close, ANOVA was used to determine whether the model containing more predictor foods was significantly better. The model with the higher adjusted R^2 value or the lower number of predictors, when no significant difference between the two models was found by ANOVA, was selected. Log-, exponential-, square root- and square-transformation of dependent and independent variables was tested to improve normality of the residuals, if necessary. For each FA, the model with the highest adjusted R^2 overall was selected.

Model interpretation and diagnostics

Standardized and studentized residuals, leverage, Cook's distance, DFFit and variance inflation factor were considered as model diagnostics for the selected models. Unusual cases, i.e. cases with standardized or studentized residuals greater than 3, leverage greater than $2(k + 1)/n$ (k is the number of predictors in model, n the number of cases used for fitting the model) or high DFFit compared to the majority of other values, were excluded if they did exhibited undue influence on the model, that is, if their Cook's distance was higher than 1 [22].

Model testing

Finally in step three, the models obtained using data from t0 (training data set) including the estimated coefficients were tested on data from t6 (test data set, Figure 1). In contrast to the training, model testing was performed on a data set not excluding cases with a reported energy intake exceeding EER by more than 30%. This was done to include also extreme cases and evaluate model performance on those. Absolute predicted and observed FA concentrations were compared using summary statistics, Spearman correlations, due to deviations from normality, squared base error rates (SBER) in relation to MSPE, R^2 and calibration plots. SBER measures the mean squared prediction error when the null model is used for prediction i.e. when the model is set to predict the mean of the observations in the training set. The MSPE is calculated using the values predicted by the model and the corresponding observations. The relative difference between prediction and observation was analyzed through summary statistics and the relative mean prediction error (MPE). For formulae of the SBER, MSPE and relative MPE, see supporting information.

For all analyses the software R, version 3.1.0 [23], was used. P-values below 0.05 were considered significant.

3. Results

***FADS1* characteristics and relation to FAs concentrations**

The FADS1 SNP allele frequency in the training as well as in the test was T = 0.33 and C = 0.67 which is in line with finding from the 1000 Genomes Project [24] . Their respective genotypes for training (nCC = 315, nTC = 316, nTT = 73) and test set (nCC = 435, nTC = 457, nTT = 101) were found to be in Hardy-Weinberg equilibrium, as the null-hypotheses could not be rejected (p values >0.05). Increasing concentrations for the n-6 FA AA and a strong decrease for DGLA levels were observed for the T homozygous over the heterozygous to the most common C homozygotes. A trend was also

observed for the effects on the n-3 FAs DPA and EPA though not significant for both, or for DHA levels (Figure S1, supporting information).

Model selection

The foods selected by bolasso regression for effects on blood FA levels were in cases of DGLA and DPA a subgroup of those identified by MHT. For blood DHA, EPA and AA levels more foods however were selected based on bolasso (Table 1). Adjusted R^2 values were very close for AA and DHA as outcomes, but F-tests showed no significant improvements when taking the larger food set into the model (data not shown). Therefore, the model with the smaller set of food items was chosen. Bolasso did not prove superiority in any model, therefore MHT models were selected. Only for the outcome of the EPA model a transformation, in this case square root transformation of MHT, improved the adjusted R^2. The finally selected models together with coefficients and levels of significance of the predictors are compiled in Table 2.

Model interpretation and diagnostics

For DGLA, two fish items from the FFQ (non-smoked oily fish and smoked fish) were found associated with significantly decreased relative blood concentration while non-wholegrain cereals as cornflakes, pizza and crisps were associated with significantly increasing levels. Age and supplementation had a negative and BMI a positive association with DGLA concentrations. DGLA concentrations were also significantly higher in women, but no effects of PAL and smoking were found. In C homozygotes of the *FADS1*, the relative DGLA blood concentration was significantly lower than in heterozygotes and T homozygotes. The model explained 32% of variance in DGLA and was highly significant ($F(20, 683) = 16.4$, $p < 0.001$). In comparison, a model containing solely *FADS1* as a predictor explained 15% of the variance ($F(2, 701) = 63.8$, $p < 0.001$).

Higher fish and tea consumption as well as age and supplementation were associated with lower, and poultry consumption with higher relative AA blood concentration without any gender, PAL, BMI

or smoking effects. *FADS1* C homozygotes had a significantly higher blood concentration of AA than heterozygotes and T homozygotes. 26% of the variance in AA concentration was explained by this model ($F(17, 686) = 13.9$, $p < 0.001$). When *FADS1* was used as a univariate predictor, 14% of the variance was explained ($F(2, 701) = 58.4$, $p < 0.001$).

For EPA, amongst other, fish consumption, supplements and age were associated with the strongest positive effects. Intakes of wine, tea and avocado were also positively associated with blood EPA. Other important predicting items were pizza, olive oil and smoking strongly associated with lowered EPA blood concentrations. The difference between the C homozygotes and the T homozygotes was significant ($p=0.04$), but not for C homozygotes and heterozygotes. The model explained 43% of the variance in relative EPA concentration ($F (27, 676) = 18.96$, $p < 0.001$). When *FADS1* was the only predictor in the model, less than 1% of the variance in relative EPA concentration was explained ($F(2, 701) = 0.9$, $p = 0.4$).

Among the significant predictors for DPA levels, olive oil was the strongest food predictor associated with decreased levels. Smoked fish and butter, but also berries, creaeals and ice-cream were identified as significantly increasing DPA blood concentrations. Chocolates, biscuits and sweet alcoholic drinks as well as BMI and smoking were found to decrease the respective concentrations. Additionally, gender was significantly associated with the relative DPA blood concentration. There was no significant effect of the *FADS1* gene. 29% of the DPA variance could be explained by this model ($F (23, 680) = 10.5$, $p < 0.001$) in contrast to less than 1% when *FADS1* was the only predictor ($F(2, 701) = 1.9$, $p = 0.15$).

For DHA, 6 different types of fish as well as zero fat skimmed milk and avocado had an increasing effect while pizza intake was associated with decreased DHA levels. Supplementation and age had a positive, PAL, smoking and BMI a strong negative effect on DHA blood concentration. The *FADS1* allele was not significantly related to DHA. 35% of the variance in blood DHA concentration was explained by this model ($F(22, 681) = 16.33$, $p < 0.001$). The explained variance decreased to less than 1% when *FADS1* was the only predictor in the model ($F(2, 701) = 0.9$, $p = 0.4$).

The differences between R^2 (Table 2) and adjusted R^2 (

Table 1) in the models are within the range of 0.01 to 0.03. Model diagnostics showed strayed cases in the training set with large deviations between FA concentrations predicted by the model and observed concentrations. However, none of those cases showed undue influence on the model and no indications for multicollinearity were found (data not shown).

Model testing

For all individual FA models, summary statistics of predictions matched those of the observations relatively well in the range between the 1st and 3rd quartile. Based on these quartiles in the summary statistics, the AA model seemed to perform best, the EPA model worst. Spearman correlation coefficients of observed versus predicted concentrations ranged around 0.55 (Table 3) and MSPE was about one quarter lower than SBER for all FAs. For EPA, the relative MPE was noticeably higher than for all other FAs where it was found to be below 5%. Scatter plots revealed a consistent trend for over-prediction of low observed values and under-prediction for high values (Figure S2, supplementary information). All the full models explained about 25% of the variance in blood FA (Table 3). This percentage was reduced to around 14% when FADS1 was the only predictor in the model for AA and DGLA. For EPA, DPA and DHA, FADS1 genotype explained less than 1% of the variance.

4. Discussion

Based on genotypes in *FADS1,* food intake data including supplement intake, PAL, gender, BMI, age, and smoking, we established mathematical models that predicted relative n-3 and n-6 blood concentrations of participants in the Food4Me study. Our statistics showed a strong effect of the *FADS1* variants for predicting n-6 FA concentrations. In T compared to C heterozygotes higher relative concentrations of AA and lower DGLA levels and only a trend towards higher concentrations of DPA and EPA, but no differences in DHA levels were found. Previous studies reported similar findings showing that these genetic influences on blood FA concentrations are very robust and independent of the sample matrix [6][5]. The fact that the genotype of *FADS1* becomes visible mainly in the n-6 FA levels may depend on an almost 15-fold higher intake of n-6 than of n-3 FA in a typical Western type diet [25]. For DHA levels, there was no significant association with *FADS1* genotype and for this FA, the blood concentration was determined mainly by dietary intake as also described previously [6][10][25].

Specific food items were identified as relevant predictors of blood FAs concentrations using MHT and bolasso. Bootstrapping was used, as it provides provably evidence of the consistent selection of the same predictors [26], suggesting that the food items selected by bolasso are indeed associated with blood FA levels. Generally, fewer foods were identified as predictors for n-6 FA than for n-3 FA concentrations. Oily fish is a major source of n-3 FAs [27] and thus, as expected, higher intakes of several types of fish showed strong positive associations with higher n-3 FA levels while decreasing those of n-6 FAs. Olive oil intake showed an opposing effect, possibly by the high n-6 to n-3 ratio in this food [28]. Intakes of several other food items revealed significant associations with blood FA concentrations although no plausible cause could be found. For example, wine, avocado and berries were associated with an increase in relative blood concentrations of n-3 FAs, whereas pizza was found to decrease levels. Also, higher intakes of poultry, pizza and cereals were associated with increased n-6 FAs, while tea was associated with decreased relative concentrations of n-6 FAs.

However, consumption of certain types of tea [29][30], wine [32][33] or specific berries [34] in connection to lipid metabolism have been described before. As tea, wine and berries contain virtually no FAs, associations with FAs in blood might be caused by specific ingredients influencing the FA metabolism in a yet unknown way or such foods might be surrogates for certain dietary patterns.

Our models also include gender, BMI, age, smoking, supplement intake and PAL as additional predictors or confounders. Comparing the significance of the association of these factors with the selected PUFA age and for n-3 smoking seems to be important predictors, whereas gender and PAL revealed lower effects.

In contrast to other studies (e.g. [6][7][8][9][10]), our models are not just tested on the data set used for fitting (training set) but on another, partly independent, data set (test set). This enables considerably better evaluation of the predictive power and suitability of the model [35]. Overall, our models displayed a correlation coefficient of 0.48 to 0.6 between observed values and those predicted by the models in the test set (Table 3). Tanaka et al. [9] identified in a genome wide association study that the SNP rs174537 in the *FADS1* gene could explain up to 18.6% of the variance in plasma AA levels in minor as compared to the major allele carriers. When using the rs174546 SNP in the *FADS1* gene, our model explained 14% in the training set. Our full model though (including selected food items, anthropometric and lifestyle factors) was able to explain 22% of the variability in AA levels in the test set and even 26% in the training set. In contrast, Zietemann et al. [6] were able to explain less than 5% of the variance by the rs174546 SNP even after adjusting for multiple lifestyle and health parameters. This means that adding selected food items as predictors into the model adds a valuable and until now unaccounted portions of explained variance for AA prediction. At best 18% and 22% of the variability in DGLA levels could be explained in the analysis by Zietemann et al. [6], depending on the number of factors adjusted for in the model. Our models explained up to 16% of the variance when only *FADS1* variants were used as a predictor in the training set. The full

model on the contrary explained 32% in the training set and 28% in the test set which is an indicator of diet being an important factor for DGLA level prediction. For the n-3 FAs, Schaeffer et al. [10] were able to explain 7%, 5% and 3% of the variance in EPA, DPA and DHA respectively, when using 11 SNPs of the 11-locus haplotype (containing rs174546) and even less when using only 5 SNPs. When only the rs174546 SNP was used as a predictor in our models, we were able to explain less than 1% of the variance in these FAs in the training set. When including or selected food items, anthropometric and lifestyle factors on the other hand, we could explain 43%, 29% and 35% in the training set and 33%, 22% and 25% in the test set. All in all, our selected food items are valuable predictors for FA level determination. Also, there is a remarkable portion of explained variance in a test set, although we performed data cleaning by excluding all cases with a reported energy intake exceeding the EER by more than 30% within the trainings set, but not within the test set. Additionally, the generalizability of our models is assumed, as there is no substantial difference between R^2 and adjusted R^2. As advantage over other studies, we also make a step further in evaluating the performance of our models by not just applying them to the training set but also to a test set. Our models prove to be valuable in predicting FA levels as we see good results concerning the predictive power.

It needs to be emphasized though that we examined just one variant of *FADS1* and additional variants in the gene cluster are known [10]. Also, other genes involved in fat digestion, absorption and metabolism are likely to influence blood FA concentrations as well. In addition to genotype and PUFA intake, further dietary factors such as the intake of saturated fat or cholesterol but also disease states such as diabetes or hypertension and medication are known to be associated with altered blood FA profiles [36].

Other limitations in our analysis are that the food intake data were collected via FFQ which are prone to misreporting which is why we included data cleaning for the training set. In addition, for the FFQ used in the Food4Me project several food items, some with quite different PUFA contents,

were pooled into food groups such as 'other vegetable oils' which included all oils except olive oil. Another example is tea consumption which revealed an association with FA patterns in former studies, but only for specific types like green or mate tea [29][30]. Combining all types of tea into one group might impede to find the dominant food item and impart an under-estimated impact. Taking all these uncertainties into account, we nevertheless demonstrate that those rather crude methods to assess food intake can deliver appropriate measures with significant associations.

The over-prediction of low FA blood concentrations and under-prediction for high levels that we observed may call for a data transformation. However, as there appears to be a systematic bias and a linear relationship, such a transformation could be misleading. Violation of normality of the residuals for some FAs appears to represent a challenge for linear models but, as concluded by Lumley et al. [37], when the sample size is sufficiently large, the assumption of normality is not required. This leaves us with the need to extent analysis to identify other determinants such as overall dietary patterns and other lifestyle parameters not yet included to further enhance predictive power.

In conclusion, our analysis based on MHT and bolasso regression identified food items that were associated significantly with blood n-3 and n-6 FA concentrations as were the genetic variants of the *FADS1* gene and age. Comparing the MHT to the respective bolasso model, MHT models were superior, showing higher adjusted R^2 and no advantage over bigger bolasso models, respectively. For the limited number of genetic, phenotypic and dietary parameters taken into model this result may be considered as valuable and high compared to other findings, especially as it is based on the evaluation on a test set in contrast to other studies. The models developed may be tested in further independent data sets and can be extended to increase the predictive power by including for example more genetic variants and other lifestyle factors.

YM, IT, ERG, LB, JAL, JAM, WHMS, HD, MJG and JCM contributed to the research design. JCM was the Food4Me Proof of Principle study leader. CCM, SNC, RSC, CW, CBO, HF, CFMM, ALM, RF, SK, CPL, CM, GM, MG, AS, MCW and JCM conducted the intervention. JH and SK wrote the paper and performed the statistical analysis for the manuscript. JH and SK are joint first authors. All authors contributed to a critical review of the manuscript during the writing process. All authors approved the final version to be published.

The Food4Me study is supported by the European Commission under the Food, Agriculture, Fisheries and Biotechnology Theme of the 7th Framework Programme for Research and Technological Development, Grant Number 265494.

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Trial registration—NCT01530139 (http://clinicaltrials.gov/show/NCT01530139).

The authors have declared no conflict of interest.

References

[1] Das, U. N., Essential fatty acids and their metabolites could function as endogenous HMG-CoA reductase and ACE enzyme inhibitors, anti-arrhythmic, anti-hypertensive, anti-atherosclerotic, anti-inflammatory, cytoprotective, and cardioprotective molecules. *Lipids Health Dis.* 2008, 7: 37.

[2] Calder, P. C., n− 3 polyunsaturated fatty acids, inflammation, and inflammatory diseases. *The American journal of clinical nutrition*, 2006, *83*(6), S1505-1519S.

[3] Andersson, A., Sjödin, A., Olsson, R., Vessby, B., Effects of physical exercise on phospholipid fatty acid composition in skeletal muscle. *Journal of Physiology-Endocrinology And Metabolism* 1998. 274(3): E432-E438.

[4] Nikkari, T., Luukkainen, P., Pietinen, P. Puska, P., Fatty acid composition of serum lipid fractions in relation to gender and quality of dietary fat. *Annals of medicine*. 1995, 27(4): 491:498.

[5] Bokor, S., Dumont, J., Spinneker, A., Gonzalez-Gross, M., Nova, E., Widhalm, K., Moschonis, G., Stehle, P., Amouyel, P., De Henauw, S., Molnàr, D., Moreno, L. A., Meirhaeghe, A., Dallongeville, J.; HELENA Study Group, Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fatty acid ratios. *J Lipid Res. Aug.* 2010, 51(8):2325-33.

[6] Zietemann, V., Kröger, J., Enzenbach, C., Janzen, E., Fritsche, A., Weikert, C., Boeing, H., Schulze, M. B., Genetic variation of the FADS1 FADS2 gene cluster and n-6 PUFA composition in erythrocyte membranes in the European Prospective Investigation into Cancer and Nutrition-Potsdam study. *Br. J. Nutr*. 2010, 104: 1748–1759.

[7] Rzehak, P., Heinrich, J., Klopp, N., Schaeffer, L., Hoff, S., Wolfram, G., Illig, T., Linseisen, J., Evidence for an association between genetic variants of the fatty acid desaturase 1 fatty acid desaturase 2 (FADS1 FADS2) gene cluster and the fatty acid composition of erythrocyte membranes. *Br J Nutr*. 2009, 101.01: 20-26.

[8] Koletzko, B., Lattka, E., Zeilinger, S., Illig, T., Steer, C., Genetic variants of the fatty acid desaturase gene cluster predict amounts of red blood cell docosahexaenoic and other polyunsaturated fatty acids in pregnant women: findings from the Avon Longitudinal Study of Parents and Children. *American journal of clinical nutrition*. 2011, 93(1): 211-219.

[9] Tanaka, T., Shen, J., Abecasis, G. R., Kisialiou, A., Ordovas, J. M., Guralnik, J., Singleton, A., Bandinelli, S., Cherubini, A., Arnett, D., Tsai, M. Y., Ferrucci, L., Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI study. *PLoS Genet.* 2009, 5:e1000338.

[10] Schaeffer, L., Gohlke, H., Müller, M., Heid, I. M., Palmer, L. J., Kompauer, I., Demmelmair, H., Illig, T., Koletzko, B., Heinrich, J., Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. *Hum. Mol. Genet.* 2006, 15: 1745–1756.

[11] Rohart, F., Multiple hypothesis testing for variable selection, *arXiv preprint* arXiv: 1106-3415, 2012.

[12] Dudoit, S., Popper Shaffer, J., Boldrick, J. C., Multiple hypothesis testing in microarray experiments. *Statistical Science.* 2003, Vol. 18, No 1: 71-103.

[13] Shaffer, J. P., Multiple hypothesis testing. *Annual Review of Psychology*. 1995, 46: 561.

[14] Rohart, F., Villa-Vialaneix, N., Paris, A., Canlet, C., Molina, J., Milan, D., Laurent, B., Sancristobal, M., Phenotypic prediction based on metabolomics data: lasso vs Bolasso, primary data vs wavelet transformation. *World Congress on Genetics Applied to Livestock Production, hal.archives-ouvertes.fr.* 2012, 3-55. Date of access: March 2015.

[15] Lampos, V., De Bie, T., Cristianini, N., Flu detector – Tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases; Lecture Notes in Computer Science.* 2010, Vol 6323: 599-602.

[16] Celis-Morales, C., Livingstone, K. M., Marsaux, C. F., Forster, H., O'Donovan, C. B., Woolhead, C., Macready, A. L., Fallaize, R., Navas-Carretero, S., San-Cristobal, R., Kolossa, S., Hartwig, K., Tsirigoti, L., Lambrinou, C. P., Moschonis, G., Godlewska, M., Surwiłło, A., Grimaldi, K., Bouwman, J., Daly, E. J., Akujobi, V., O'Riordan, R., Hoonhout, J., Claassen, A., Hoeller, U., Gundersen, T. E., Kaland, S. E., Matthews, J. N., Manios, Y., Traczyk, I., Drevon, C. A., Gibney, E. R., Brennan, L., Walsh, M. C., Lovegrove, J. A., Martinez, J. A., Saris, W. H., Daniel, H., Gibney, M., Mathers, J. C., Design and baseline characteristics of the Food4Me study: a web-based randomised controlled trial of personalised nutrition in seven European countries. *Genes and Nutrition.* 2015, 10:450.

[17] Forster, H., Fallaize, R., Gallagher, C., O'Donovan, C. B., Woolhead, C., Walsh, M. C., Macready, A. L., Lovegrove, J. A., Mathers, J. C., Gibney, M. J., Brennan, L., Gibney, E. R., Online Dietary Intake Estimation: The Food4Me Food Frequency Questionnaire. *J Med Internet Res.* 2014, 16(6): e150

[18] Fallaize, R., Forster, H., Macready, A. L., Walsh, M. C., Mathers, J. C., Brennan, L., Gibney, E. R., Gibney, M. J., Lovegrove, J. A., Online Dietary Intake Estimation: Reproducibility and Validity of the Food4Me Food Frequency Questionnaire Against a 4-Day Weighed Food Record. *J. Med. Internet Res.* 2014, 16(8): e190.

[19] Baecke, J.A., Burema, J., Frijters, J.E., A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *Am J Clin Nutr*. 1982, 36(5):936-42.

[20] Bailey-Hall, E., Nelson, E.B., Alan S. Ryan, A.S., Validation of a Rapid Measure of Blood PUFA Levels in Humans.Lipids (2008) 43:181–186

[21] Bell, J.G., Mackinlay. E.E., Dick, J.R., Younger, I., Lands, B., Gilhooly, T., Using a fingertip whole blood sample for rapid fatty acid measurement: method validation and correlation with erythrocyte polar lipid compositions in UK subjects. *Br J Nutr.* 2011, 106(9):1408-15

[22] Field, A., Miles, J., Field, Z., *Discovering Statistics using R*, SAGE Publications, 2012: 269.

[23] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2014, *http://www.R-project.org/*

[24] 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A., An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012, 491: 56–65

[25] Simopoulos, A. P., The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases. *Exp Biol Med (Maywood)*, 2008, 233(6):674-88.

[26] Bach, F. R., Bolasso: Model Consistent Lasso Estimation through the Bootstrap. *Proceedings of the 25th international conference on Machine learning, arXiv: http://arxiv.org/pdf/0804.1302.pdf:* 2008, 33-40. *Date of access: March 2015.*

[27] Strobel, C., Jahreis, G., Kuhnt, K., Survey of n-3 and n-6 polyunsaturated fatty acids in fish and fish products. *Lipids Health Dis*. 2012, 11: 144.

[28] Mailer, R., Chemistry and quality of olive oil. *NSW DPI primefacts.* 2006.

[29] Ahmed, S. T., Lee, J.W., Mun, H. S., Yang, C. J., Effects of supplementation with green tea by-products on growth performance, meat quality, blood metabolites and immune cell proliferation in goats. *J Anim Physiol Anim Nutr.* 2014. DOI: 10.1111/jpn.12279.

[30] Martins, F., Suzan, A. J., Cerutti, S. M., Arçari, D. P., Ribeiro, M. L., Bastos, D. H., Carvalho Pde, O., Consumption of mate tea (lex paraguariensis) decreases the oxidation of unsaturated fatty acids in mouse liver. *Br J Nutr.* 2009. 101(4):527-32.

[31] Qin, B., Polansky, M. M., Harry, D. , Anderson, R.A., Green tea polyphenols improve cardiac muscle mRNA and protein levels of signal pathways related to insulin and lipid metabolism and inflammation in insulin-resistant rats, *Mol. Nutr. Food Res.*, 2010, 54, S14–S23

[32] Urquiaga, I., Guasch, V., Marshall, G., San Martín, A., Castillo, O., Rozowski, J., Leighton, F., Effect of Mediterranean and Occidental diets, and red wine, on plasma fatty acids in humans. An intervention study. *Biological Reseach*. 2004, 37 (2): 253-61.

[33] Rifler, J. P., Lorcerie, F., Durand, P., Delmas, D., Ragot, K., Limagne, E., Mazué, F., Riedinger, J., d'Athis, P., Hudelot, B., Prost, M., Lizard, G., Latruffe, N., A moderate red wine intake improves blood lipid parameters and erythrocytes membrane fluidity in post myocardial infarct patients. *Mol. Nutr. Food Res.*, 2012, 56(2), 345-351.

[34] Basu, A., Rhone, M., Lyons, T. J., Berries: emerging impact on cardiovascular health. *Nutrition reviews* 2010,68.3: 168-177.

[35] Tropsha, A., Gramatica, P., Gombar, V., The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *BSAR & Combinatorial Science*, 2003, 22(1): 60-77.

[36] Das, U. N., Essential fatty acids: Biochemistry, physiology, and pathology. *Biotechnology J.* 2006, 1:420–439.

[37] Lumley, T., Diehr, P., Emerson, S., Chen, L., The importance of the normality assumption in large public health data sets. *Annual review of public health.* 2002, 23.1: 151-169.

Figure 1: Data selection criteria and data sets defined for different analyses. Data were selected from the Food4Me study with 1,607 participants. Each participant delivered food, blood, Estimated Energy Requirement (EER) and energy intake on 3 measurement days.

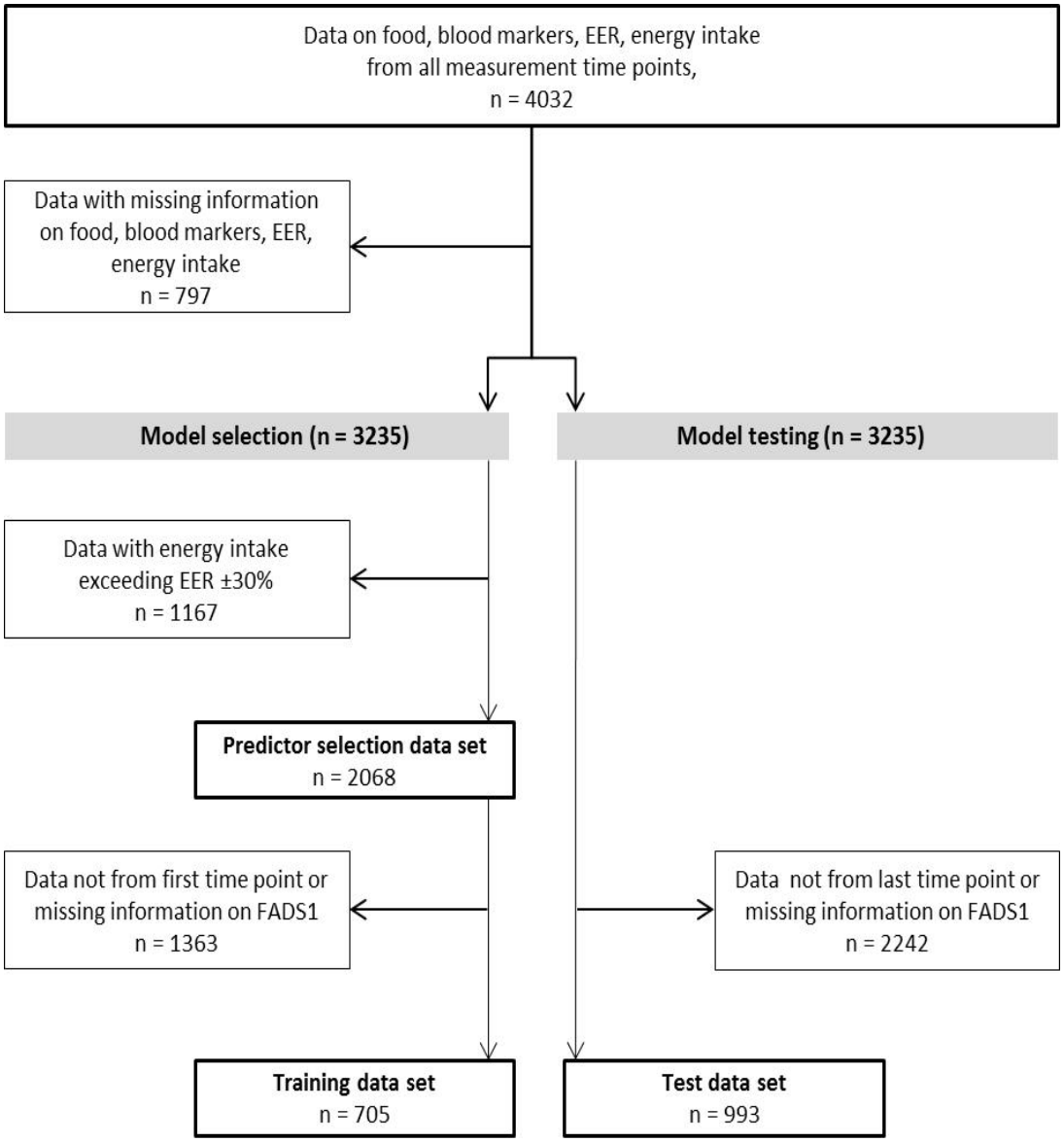

Table 1: Adjusted R^2 values for model variants: MHT, bolasso, transformation. Number of selected food items and adjusted R^2 values for models including the foods selected based on bolasso or by MHT for individual FAs. The * indicates the model chosen for further analysis.

Fatty acid/fatty acid group	Number of selected food items	adjusted R^2
DGLA		
MHT *	12	0.31
bolasso	5	0.30
AA		
MHT *	9	0.24
bolasso	12	0.24
EPA		
MHT	12	0.38
bolasso	19	0.37
MHT sqrt(EPA) *	12	0.41
DPA		
MHT *	19	0.26
bolasso	7	0.23
DHA		
MHT *	14	0.32
bolasso	15	0.32

Table 2: Regression coefficients, standard error and p-values of predictors included in models Food items (standardized) selected from FFQ data used for modelling selected either in 95% of lasso bootstraps or by MHT for each fatty acid. Food items ordered according to the magnitude of their regression coefficient with positive regression coefficients before negative ones when they have the same absolute value. FAs stated in % FAME; ns: not significant; p-value ≤ 0.05 and > 0.01: *; p-value ≤ 0.01 and >0.001: **; p-value ≤ 0.001: ***

Predictors	Coefficient (se)	p-value
DGLA (R^2 = 0.32)		
FADS1 T:C	0.2 (0.02)	***
FADS1 T:T	0.5 (0.04)	***
Woman	0.05 (0.02)	*
PAL	-0.09 (0.1)	ns
Age	-0.004 (0.001)	***
BMI	0.01 (0.002)	***
Smoking	0.01 (0.04)	ns
Supplements	-0.1 (0.03)	**
Smoked Fish	-0.04 (0.01)	**
Non Smoked Oily Fish	-0.05 (0.01)	***
Pizza	0.03 (0.01)	**
Crisps	0.02 (0.01)	*
Cornflakes	0.05 (0.01)	***
Offal	-0.02 (0.01)	ns
Nuts and Seeds	-0.01 (0.01)	ns
Soups	-0.02 (0.01)	ns
Tofu	0.01 (0.01)	ns
Cereals	0.02 (0.01)	ns
Garlic	0.001 (0.01)	ns
Sausages	-0.01 (0.01)	ns
AA (R^2 = 0.26)		
FADS1 T:C	-0.8 (0.1)	***
FADS1 T:T	-1.7 (0.2)	***
Woman	0.04 (0.1)	ns
PAL	-0.07 (0.5)	ns
Age	-0.02 (0.004)	***
BMI	0.002 (0.01)	ns
Smoking	-0.1 (0.2)	ns
Supplements	-0.4 (0.2)	**
Wine	-0.02 (0.05)	ns
Chicken	0.2 (0.06)	**
Non Smoked Oily Fish	-0.1 (0.06)	*
Non Smoked Oily Fish Canned	-0.1 (0.06)	*
Smoked Fish	-0.1 (0.05)	*
Eggs	0.2 (0.07)	ns
Tea	-0.2 (0.05)	***
Grapefruit	-0.1 (0.05)	ns
Coleslaw	0.1 (0.06)	ns
DHA (R^2 = 0.35)		
FADS1T:C	-0.1 (0.06)	ns
FADS1T:T	-0.1 (0.09)	ns
women	0.03 (0.06)	ns
PAL	-0.9 (0.3)	**
Age	0.01 (0.002)	***
BMI	-0.03 (0.01)	***
Smoking	-0.3 (0.09)	**
Supplements	0.2 (0.08)	**
Non Smoked Oily Fish	0.2 (0.04)	***
Smoked Fish	0.2 (0.03)	***
White Fish	0.1 (0.03)	***
Pizza	-0.08 (0.03)	**
Fried Fish	0.1 (0.03)	***
Sushi	0.1 (0.03)	**
Zero Fat Skimmed Milk	0.1 (0.02)	***
Avocado	0.1 (0.03)	**
Non Smoked Oily Fish Canned	0.1 (0.03)	**
Melon	0.02 (0.04)	ns
Burgers	-0.03 (0.03)	ns
Medium Fat Cheeses	-0.04 (0.03)	ns
Chips	-0.05 (0.03)	ns
Sugar added to Coffee/Tea	0.01 (0.02)	ns

Predictors	Coefficient (se)	p-value
DPA (R^2 = 0.29)		
FADS1 T:C	-0.04 (0.03)	ns
FADS1 T:T	-0.07 (0.04)	ns
Woman	-0.1 (0.03)	***
PAL	0.15 (0.1)	ns
Age	-0.0001 (0.001)	ns
BMI	-0.006 (0.003)	*
Smoking	-0.1 (0.04)	**
Supplements	0.1 (0.04)	*
Olive Oil	-0.07 (0.01)	***
Smoked Fish	0.04 (0.01)	**
Butter	0.04 (0.01)	***
Cereals	0.04 (0.01)	**
Nuts and Seeds	0.02 (0.02)	ns
Porridge	0.03 (0.01)	**
Berries	0.05 (0.02)	**
Ice Cream	0.04 (0.01)	**
Flapjacks	-0.01 (0.01)	ns
Chocolates	-0.03 (0.01)	*
Sweet Biscuits	-0.03 (0.01)	*
High Fat Cheeses	0.02 (0.01)	ns
Sweet Alcoholic Drinks	-0.03 (0.01)	*
Brown Bread	-0.01 (0.01)	ns
Brown Rice	0.009 (0.01)	ns
Potatoes	0.02 (0.01)	ns
Coffee	0.008 (0.01)	ns
Pears	-0.02 (0.01)	ns
Pork	-0.02 (0.01)	ns
EPA (R^2 = 0.43)		
FADS1 T:C	-0.02 (0.01)	ns
FADS1 T:T	-0.05 (0.02)	*
Woman	0.003 (0.02)	ns
PAL	0.03 (0.07)	ns
Age	0.003 (0.0006)	***
BMI	-0.003 (0.002)	ns
Smoking	-0.09 (0.02)	***
Supplements	0.1 (0.02)	***
Smoked Fish	0.07 (0.008)	***
Tea	0.02 (0.007)	*
Non Smoked Oily Fish	0.05 (0.01)	***
Pizza	-0.03 (0.007)	***
Wine	0.02 (0.007)	*
Olive Oil	-0.03 (0.007)	***
Avocado	0.02 (0.009)	**
Lamb	0.008 (0.007)	ns
Offal	-0.0008 (0.009)	ns
Other Vegetable Oils	0.004 (0.007)	ns
White Bread	-0.01 (0.006)	ns
Non Smoked Oily Fish Canned	0.03 (0.009)	***
Broccoli	0.007 (0.008)	ns
Butter	0.02 (0.006)	*
Flapjacks	-0.0003 (0.006)	ns
Low Calorie Soft Drinks	-0.009 (0.007)	ns
White Fish	0.01 (0.008)	ns
Kiwi	0.006 (0.006)	ns
Sweet Biscuits	-0.02 (0.008)	**

Table 3: Spearman correlation coefficient of observed versus predicted FA concentrations, SBER, MSPE, relative MPE, R^2 and R^2 when *FADS1* is used as univariate predictor of observed values and values predicted by the models in test set. r_s Spearman correlation coefficient, SBER squared base error rate, MSPE mean squared prediction error, relative MPE relative mean prediction error, R^2 FADS1 R^2 when FADS1 was the only predictor in the model.

FA	r_s (95 % CI)	SBER (95 % CI)	MSPE (95 % CI)	relative MPE [%]	R^2	R^2 *FADS1*
DGLA	0.51 (0.46, 0.55)	0.12 (0.11, 0.13)	0.09 (0.08, 0.10)	-1	0.28	0.16
AA	0.48 (0.43, 0.53)	2.10 (1.93, 2.33)	1.68 (1.52, 1.85)	-2	0.22	0.12
EPA	0.60 (0.56, 0.64)	0.23 (0.18, 0.27)	0.17 (0.12, 0.23)	12	0.33	<0.001
DPA	0.50 (0.46, 0.55)	0.13 (0.11, 0.14)	0.10 (0.09, 0.11)	-0.1	0.22	<0.001
DHA	0.54 (0.49, 0.58)	0.84 (0.77, 0.93)	0.64 (0.56, 0.71)	5	0.25	<0.001