Paper SA-08

Regression calibration with multiple imputations for red blood cell fatty acids

James V. Pottala, OmegaQuant, Sioux Falls, SD

Abstract

Red blood cell (RBC) fatty acids have positive and negative associations with metabolism and physiology, and subsequently with chronic diseases. Improper freezer storage temperature (i.e. -20°C instead of -80°C) resulted in a loss of polyunsaturated fatty acids (PUFA), which biased the gas chromatography measurements in several thousand samples. External validation experiments involving storage temperature and aliquot size were conducted to mimic the conditions of the samples, and General Linear Mixed Models (GLMM) were used to develop bias-corrected regression calibration equations. Then multiple imputations incorporating stochastic draws for the variance, slopes, and residuals were implemented to accurately reflect the uncertainty in the true fatty acid values. The expanded dataset can be analyzed by imputation with standard statistical methods.

Introduction

Higher blood levels of the essential omega-3 PUFAs, which cannot be efficiently produced *in vivo* and must be consumed, have been shown to reduce risk for sudden cardiac death[1,2] and all-cause mortality[2,3]. There is also evidence that the essential omega-6 PUFAs are inversely associated with risk for coronary heart disease[4]. Hence the study of these fatty acids is of vital importance.

During a recent investigation in which approximately 8,700 RBC samples were analyzed, a particular fatty acid (whose levels are metabolically not nutritionally regulated) was found to be implausibly low by gas chromatography. The issue was later determined to be oxidative degradation due to a few weeks of improper storage temperature, which created biased measurements. Before the data could be assessed for relationships between PUFAs and clinical outcomes, the bias needed to be corrected.

Methods

In order to determine the combined effects of storage temperature and aliquot size on the apparent loss of PUFAs, we used ninety-seven 1.8 mL RBC samples that had always been stored at -80ºC. The samples were thawed and one aliquot was taken for immediate FA analysis. The rest of each sample was divided into six aliquots: three (80 uL) and three (250 uL). The six aliquots were all placed at -20ºC. After 13 days one 80 uL and one 250 uL sample were removed, and analyzed for FA composition. Other aliquot pairs were removed and analyzed on days 20 and 27 of -20ºC storage (Figure 1).

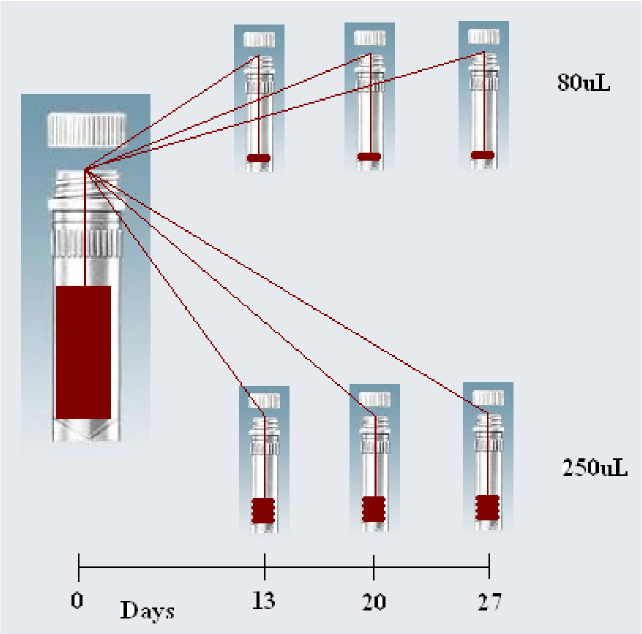

Figure 1. RBC -20°C storage experiment design

General linear mixed models (GLMM) were used to incorporate the correlation among the repeated aliquots from samples using a random intercept model with restricted maximum likelihood estimation for each fatty acid. The mean response profile over time was modeled using a piecewise linear function with a knot at 13 days, and allowed different slopes by aliquot size.

The fatty acids are exposure variables, and the improper freezer condition created biased measurements. The -20C storage experiment was designed to establish regression calibration equations. Because these equations are estimated, it is important to include the imprecision due to this estimation in the bias-corrected data. Values for the bias-corrected data were imputed by drawing parameters of the calibration equations from their sampling distributions. This was repeated for the entire cohort study m=10 times. The variance of a statistical estimand is the average variance over the repeated imputations (within variability) + the variance of the posterior mean (between variability); for finite imputations the between variability needs to be multiplied by [1 + (1 / m)][5].

Results

There are development and production versions of SAS® code. *During* the statistical analyses there are considerations for methodological approach, distributional assumptions, functional forms of covariates, correlation among responses and/or predictor variables, influential observations, interactions, etc. *After* the statistical analyses there is a single (or a few) models, which are the basis of the study's findings and need to be clearly reported and interpreted. The macro program below is a simplified version of production SAS® code used to demonstrate the steps for implementing regression calibration with multiple imputations, and it produces a 'final results' table of the bias in the fatty acid measurements. However due to a coding simplification explained in step 2, which results in some model over-fitting, the final results table is not biologically accurate but provided for pedagogical purposes.

Step 1 – The below code defines the macro *%regcal*, for regression calibration, with two keywords *filename* and *outputpath*. It also initializes a temporary dataset *corr* that will be used to store the intra-correlation coefficients for each fatty acid, since the samples represent 97 clusters of up to 7 measurements each.

The macro variable *&varlist* includes the names of 24 fatty acids. The fatty acids have been listed generically as saturated (SAT), monounsaturated (MONO), trans fats (TRANS), and polyunsaturated (PUFA); because the results of the cohort study using these data are being prepared for journal submission.

```
%macro regcal(filename=,outputpath=);
data corr;
   length fa $10.;
run;

%let varlist=aSAT1 aSAT2 aSAT3 aSAT4 aSAT5 aSAT6 bMONO1 bMONO2 bMONO3 bMONO4 cTRANS1
cTRANS2 cTRANS3 dPUFA1 dPUFA2 dPUFA3 dPUFA4 dPUFA5 dPUFA6 dPUFA7 dPUFA8 dPUFA9 ePUFA10
ePUFA11;
```

Step 2 – A macro loop is used to fit a general linear mixed model (shown below) separately to each fatty acid. By processing was not used because in the actual analysis each fatty acid can have a mean response profile including zero to four slope parameters, which have been selected to minimize an objective fit measure (e.g. Akaike Information Criterion[6]). The below *model* statement demonstrates using 4 slope parameters as shown in Figure 2.

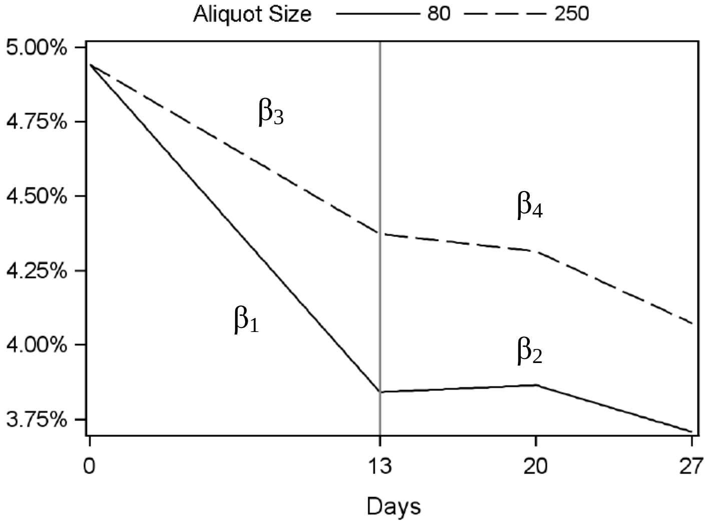

Figure 2. Example PUFA mean response profile

$$Y_{ij} = \beta_0 + \beta_1 Days_i + \beta_2 (Days_i - 13)_+ + \beta_3 Days_i * I(250uL_i) + \beta_4 (Days_i - 13)_+ * I(250uL_i) + b_{0i} + e_{ij}$$

$$b_{0i} \sim N\left(0, \sigma_b^2\right);\ e_{ij} \sim N(0, \sigma^2);\ i = 1, \ldots, 97;\ j = 0, 13, 20, 27\ days$$

```
%do i=1 %to 24;
   %let v=%scan(&varlist, &i);
   ods output Estimates=mixest&i VCorr=vcorr&i CovParms=covparms&i;
   proc mixed data=&filename method=reml noclprint=10;
      class idnum;
      title "&v";
      model &v = slope1days slope2days slope1days*vial250ul slope2days*vial250ul /
         solution cl ddfm=satterth covb;
      estimate "SlopeA80ul" slope1days 7 / cl;
      estimate "SlopeB80ul" slope1days 7 slope2days 7 / cl;
      estimate "SlopeA250ul" slope1days 7 slope1days*vial250ul 7 / cl;
      estimate "SlopeB250ul" slope1days 7 slope1days*vial250ul 7 slope2days 7
         slope2days*vial250ul 7 / cl;
      random intercept / subject=idnum type=vc vcorr;
   run;

   data corr;
      set corr vcorr&i(in=a);
      if a then FA="&v";
   run;
```

Step 3 – The dataset *reg&i* is used to merge the slope parameter estimates with the residual variance from their corresponding PROC MIXED analysis.

```
   proc sql;
      create table reg&i as
         select a.*, b.estimate as MSE
         from mixest&i as a, covparms&i as b
         where covparm='Residual';
   quit;
```

Step 4 – PROC SQL was used to calculate the baseline data for each fatty acid.

```
   proc sql;
      create table mean&i as
         select mean(&v) as Mean format=5.2,std(&v) as SD format=5.2,count(&v) as N,
           "&v" as FA
         from &filename
         where slope1days=0;
   quit;
```

Step 5 – The below code implements the steps Rubin listed to make stochastic draws from the bias-corrected sampling distributions[7]. For each imputation (m=10) use independent draws for all random variables.

a) Draw χ^2_{n-q} and let $\sigma_*^2 = \hat{\sigma}^2(n-q)/\chi^2_{n-q}$

b) Draw q independent N(0,1) and let $\beta_* = \hat{\beta} + \sigma_*[V]^{1/2}Z$

c) Draw n independent N(0,1) and impute the missing value as $Y_{i*} = Y_i + X_i\beta_* + Z_i\sigma_*$ (This step is carried out in a separate macro on the cohort study data, but an example is presented at the end of this section.)

```
   data beta&i;
      set reg&i;
      array ChiSq{10} _temporary_;
      length fa $10.;
      rename label=beta_type;
      if label in ('SlopeA250ul', 'SlopeB250ul') then AliquotSize=250;
         else AliquotSize=80;
      FA="&v";
      call symput("fa",fa);
      call streaminit(20110711);
      do ImputeID=0 to 10;
         if ImputeID=0 then do;
            RanSlope=0;
            RanMSE=0;
         end;
         else do;
            if label='SlopeA80ul' then chisq{ImputeID}=rand('CHISQUARE',95);
            X2=chisq{ImputeID};
            RanMSE=mse*95/x2;
            Z=rannor(20110131&i);
            RanSlope=-sum(estimate,sqrt(RanMSE/mse)*stderr*z);
         end;
         output;
      end;
   run;
```

Step 6 – The below data step creates a calibration matrix, which allows a single fatty acid measurement to be matched by aliquot size and then creates 11 values, i.e. the original values plus 10 imputations. An example using the calibration matrix is provided following the macro.

```
   proc sort data=beta&i;
      by ImputeID AliquotSize;
   run;
   data cal&i;
      set beta&i;
      by ImputeID AliquotSize;
      keep ImputeID AliquotSize RandomSlope1_&fa RandomSlope2_&fa RandomMSE_&fa;
      retain RandomSlope1_&fa .;
      if beta_type in ('SlopeA80ul','SlopeA250ul') then RandomSlope1_&fa=ranslope;
         else RandomSlope2_&fa=RanSlope;
      RandomMSE_&fa=RanMSE;
      if last.AliquotSize then do;
         output;
         RandomSlope1_&fa=.; RandomSlope2_&fa=.; RandomMSE_&fa=.;
      end;
   run;
%end;
```

Step 7 – The above *%end* statement concludes the macro loop, which executes once for each fatty acid. The below data steps combine the individual fatty acid datasets as needed.

```
data mwsug11.fa_beta;
   set beta1-beta24;
run;

data mwsug11.cal_matrix;
   merge cal1-cal24;
   by ImputeID;
run;

data fa_mean;
   length FA $10.;
   rename mean=BLmean sd=BLsd;
   set mean1-mean24;
run;
```

Step 8 – The below steps build columns 1-4 of Table 1.

```
data fa_corr;
   set corr(where=(row=2));
   rename col1=WithinSubjectCorr;
   keep fa col1;
run;

proc sql;
   create table mean_corr as
      select *
      from fa_mean as a left join fa_corr as b on a.fa=b.fa;
quit;
```

Step 9 – The below steps build columns 5-8 of Table 1.

```
data slopes;
   set mwsug11.fa_beta(where=(ImputeID=0));
   by fa;
   keep fa WkSlopeA80ul WkSlopeB80ul WkSlopeA250ul WkSlopeB250ul;
   retain WkSlopeA80ul WkSlopeB80ul WkSlopeA250ul .;
   if beta_type='SlopeA80ul'     then WkSlopeA80ul   =estimate;
   if beta_type='SlopeB80ul'     then WkSlopeB80ul   =estimate;
   if beta_type='SlopeA250ul'    then WkSlopeA250ul  =estimate;
   if beta_type='SlopeB250ul'    then WkSlopeB250ul  =estimate;
   if last.fa then do;
      output; WkSlopeA80ul=.; WkSlopeB80ul=.; WkSlopeA250ul=.;
   end;
run;
```

Step 10 – PROC SQL was used to build the final dataset of results, which gets printed as an .rtf file for journal publication and archive.

```
proc sql;
   create table mwsug11.bias as
      select a.fa as FattyAcid, blmean*100 as Mean format=5.2,
         blsd*100 as SD format=5.2,
         withinsubjectcorr as IntraCorr label="Intra-Correlation" format=4.2,
         WkSlopeA250ul/blmean as Bias250uLwk02 label="0-13d, 80uL" format=percent6.1,
         WkSlopeB250ul/blmean as Bias250uLwk24 label="14-27d, 80uL"
            format=percent6.1,
         WkSlopeA80ul/blmean as Bias80uLwk02 label="0-13d, 250uL" format=percent6.1,
         WkSlopeB80ul/blmean as Bias80uLwk24 label="14-27d, 250uL" format=percent6.1
      from slopes as a left join mean_corr as b on a.fa=b.fa;
quit;

title "Table 1. Bias (% mean change per week) in red blood cell fatty acid
measurements";
ods rtf file="&outputpath.\Table 1 Bias.rtf";
proc print data=mwsug11.bias noobs label;
run;
ods rtf close;
```

Step 11 – Delete temporary datasets

```
proc datasets;
   delete beta1-beta24 mixest1-mixest24 vcorr1-vcorr24 mean1-mean24 cal1-cal24
      covparms1-covparms24 reg1-reg24;
run;
quit;

%mend;
```

Step 12 – Call the macro *%regcal* and watch Table 1 get generated (demo).

%regcal(filename=mwsug11.expdata,outputpath=C:\Biostat\0-SAS\MWSUG 2011\Papers)

Table 1. Bias (% mean change per week) in red blood cell fatty acid measurements

Fatty Acid	Mean*	SD	Intra-Correlation	0-13d, 250uL	14-27d, 250uL	0-13d, 80uL	14-27d, 80uL
SAT1	0.44	0.11	0.41	(2.3%)	1.2%	2.2%	(4.5%)
SAT2	23.08	1.34	0.72	2.1%	2.0%	5.7%	1.4%
SAT3	17.48	1.02	0.73	2.5%	2.4%	6.8%	1.3%
SAT4	0.11	0.08	0.15	6.0%	(4.1%)	0.9%	2.0%
SAT5	0.06	0.02	0.20	26%	1.9%	21%	22%
SAT6	0.13	0.05	0.21	23%	2.9%	29%	6.3%

Fatty Acid	Mean*	SD	Intra-Correlation	0-13d, 250uL	14-27d, 250uL	0-13d, 80uL	14-27d, 80uL
MONO1	0.51	0.21	0.86	1.3%	2.4%	5.2%	(2.7%)
MONO2	14.17	1.07	0.81	2.9%	0.3%	3.5%	1.0%
MONO3	0.22	0.05	0.42	4.3%	0.6%	6.4%	3.3%
MONO4	0.09	0.03	0.17	31%	13%	37%	14%
TRANS1	0.22	0.04	0.14	24%	6.0%	8.2%	9.0%
TRANS2	2.20	0.65	0.94	1.8%	(1.1%)	2.5%	(1.3%)
TRANS3	0.59	0.16	0.17	13%	(13%)	5.6%	(5.0%)
PUFA1	11.09	1.72	0.89	(1.1%)	0.1%	(2.6%)	(1.5%)
PUFA2	0.29	0.05	0.43	(2.4%)	1.6%	1.0%	(1.3%)
PUFA3	0.17	0.07	0.55	(4.5%)	1.1%	(10%)	3.4%
PUFA4	0.11	0.05	0.11	(11%)	(4.5%)	(0.3%)	(22%)
PUFA5	1.66	0.37	0.82	(2.1%)	(0.2%)	(6.9%)	1.1%
PUFA6	15.51	1.36	0.65	(4.6%)	(3.1%)	(9.6%)	(1.9%)
PUFA7	3.81	0.78	0.78	(5.6%)	(4.1%)	(11%)	(1.6%)
PUFA8	0.66	0.46	0.94	(6.7%)	(1.9%)	(9.2%)	(2.8%)
PUFA9	2.38	0.38	0.79	(6.4%)	(2.6%)	(11%)	(2.1%)
PUFA10	0.68	0.20	0.65	(4.1%)	(2.4%)	(9.9%)	(0.4%)
PUFA11	4.34	1.42	0.85	(5.8%)	(3.1%)	(12%)	(1.1%)

* Reported as a % of total fatty acids.

Example: How to Use the Calibration Matrix (Supplementary Table 1)

Let a subject have a saturated fatty acid (SAT2) measured amount of 0.22 (22% of total fatty acids), which was exposed to -20C for 18 Days in an 80uL aliquot sample. If Days<14 then Days2=0, else Days2=Days-13. The same subject also has a PUFA1 amount of 0.10 (10%). Draw m=10 from N(0,1) and do a one-to-many join with the subjects' data as a row vector [80 18 5 0.22 0.10]. Then impute the missing values as $Y_{i*} = Y_i + X_i\beta_* + Z_i\sigma_*$ matching on aliquot size and impute ID, where the random slopes and MSE for each fatty acid are from the calibration matrix. Each new value is calculated as below, and the results from this example are shown in Table 2.

$$NewAmount = OldAmount + \frac{Days}{7} * RandomSlope1 + \frac{Days2}{7} * RandomSlope2 + Z_i * \sqrt{RandomMSE}$$

Table 2. Imputed fatty acid measurements

Impute ID	Z_i	SAT2	PUFA1
0	0	22.00	10.00
1	1.471	22.54	11.31
2	-0.158	19.29	10.41
3	-0.717	18.50	10.00
4	-0.788	17.84	9.92
5	0.629	20.32	10.88
6	0.876	21.01	10.97
7	0.675	21.50	10.88
8	-0.739	18.10	10.03
9	-0.501	18.70	10.20
10	1.060	21.63	11.20

Supplementary Table 1. Calibration Matrix*

Aliquot Size	Impute ID	SAT2 Random Slope1	SAT2 Random Slope2	SAT2 Random MSE	PUFA1 Random Slope1	PUFA1 Random Slope2	PUFA1 Random MSE
80	0	0	0	0	0	0	0
250	0	0	0	0	0	0	0
80	1	-0.011695	-0.005048	.000281641	0.002422	0.001175	.000038767
250	1	-0.004131	-0.002447	.000281641	0.001204	-0.000772	.000038767
80	2	-0.014477	-0.003483	.000299120	0.003106	0.001434	.000041173
250	2	-0.006683	-0.005152	.000299120	0.001918	0.000373	.000041173
80	3	-0.013006	-0.002595	.000332949	0.002972	0.001526	.000045829
250	3	-0.003792	-0.004666	.000332949	0.000877	-0.000359	.000045829
80	4	-0.016003	-0.002106	.000338703	0.002783	0.001017	.000046621
250	4	-0.004793	-0.004655	.000338703	0.001385	0.000489	.000046621
80	5	-0.016832	-0.003969	.000331820	0.002816	0.002114	.000045674
250	5	-0.007226	-0.007935	.000331820	0.001288	-0.000457	.000045674
80	6	-0.013755	-0.001579	.000236879	0.002908	0.001780	.000032606
250	6	-0.003583	-0.003933	.000236879	0.001063	-0.000276	.000032606
80	7	-0.011603	-0.004895	.000445437	0.002242	0.002218	.000061313
250	7	-0.002459	-0.004955	.000445437	0.001416	0.000176	.000061313
80	8	-0.014734	-0.002301	.000361884	0.003330	0.001675	.000049812
250	8	-0.002187	-0.005910	.000361884	0.001304	0.000350	.000049812
80	9	-0.014751	-0.004695	.000278559	0.003147	0.001952	.000038343
250	9	-0.004341	-0.002952	.000278559	0.000731	-0.000297	.000038343
80	10	-0.014348	-0.004528	.000364939	0.002775	0.001548	.000050233
250	10	-0.006696	-0.004450	.000364939	0.001846	-0.000455	.000050233

***** The entire calibration matrix has 74 columns, i.e. aliquot size, impute ID, and (random slope1, random slope2, random MSE) * 24 fatty acids. However, for demonstration only 2 fatty acids are shown above (SAT2 and PUFA1).

Conclusion

The improper freezer temperature (i.e. -20°C instead of -80°C) allowed oxidation from the air inside the sample vial to interrupt the carbon double bonds in the PUFAs. The average weekly decreases in PUFA content during the first two weeks were -4.9% and -7.4% for the 250uL and 80uL aliquots, respectively (Table 1). Over the next two weeks, the PUFA average content decreased by an additional -1.6% and -2.7%, respectively. Hence most of the loss occurred early, and the 80uL aliquot declined more rapidly than the 250uL. The saturated and monounsaturated fatty acid proportions increased in a compensatory manner as the PUFAs decreased, since the total amount of fatty acids is constrained to 100%.

The FA composition of the damaged RBC samples was largely rehabilitated by application of multiple linear regression equations and multiple imputation techniques. While these techniques are intended to reduce bias, the cost is increased variability in the imputed data. RBC samples should always been stored at -80C for analyzing fatty acid composition.

References

1. Albert CM, Campos H, Stampfer MJ, et al. Blood levels of long-chain n-3 fatty acids and the risk of sudden death. *N Engl J Med* 346:1113-1118, 2002.

2. Wang C, Harris WS, Chung M, et al. n-3 Fatty acids from fish or fish-oil supplements, but not alpha-linolenic acid, benefit cardiovascular disease outcomes in primary- and secondary-prevention studies: a systematic review. *Am J Clin Nutr* 84:5-17, 2006.

3. Pottala JV, Garg S, Cohen BE, Whooley MA, Harris WS. Blood Eicosapentaenoic and Docosahexaenoic Acids Predict All-Cause Mortality in Patients With Stable Coronary Heart Disease: The Heart and Soul Study. *Circ Cardiovasc. Qual Outcomes* 3:406-412, 2010.

4. Harris WS, Mozaffarian D, Rimm EB et al. Omega-6 Fatty Acids and Risk for Cardiovascular Disease: A Science Advisory from the American Heart Association Nutrition Committee. *Circulation* 119:902-907, 2009.

5. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 91:473-489, 1996.

6. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov EBN, Csaki F, editors. 2nd International Symposium on Information Theory and Control. Budapest: Akademia Kiado, 267-281, 1973.

7. Rubin DB. Multiple imputation for non-response in surveys. New Jersy: John Wiley and Sons, 1987.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

James V. Pottala
OmegaQuant®
2329 N. Career Avenue, Suite 233
Sioux Falls, SD 57107
(605) 271-6917, Ext. 8783
potta119@umn.edu