Factors Affecting Plasma Levels of Retinol and Beta Carotene

Applied Regression Analysis (36-707), Project 1

February 21, 2002

Abstract

Previous observational studies have suggested that low plasma concentrations of carotenoids may increase the risk of developing certain kinds of cancer, but the determinants of these concentrations are presently unknown. We analyzed a cross-sectional study designed to investigate whether certain personal characteristics and dietary factors might influence plasma concentrations of retinol and beta-carotene. Building multiple regression models to predict each micronutrient, we found plasma retinol levels tended to increase with age and alcohol consumption, whereas beta-carotene levels tended to increase with vitamin use and fiber intake, and decrease with body mass and smoking habits. However, even the best models were unable to explain most of the variability in plasma concentrations, and they performed poorly when predicting plasma concentrations for a randomly selected test subset of the data. There is great variability in plasma concentrations of both micronutrients which has yet to be explained.

Introduction

Because previous observational studies have suggested that low plasma concentrations of carotenoids, including retinol and beta-carotene, may increase the risk of developing certain types cancer, there is a need to identify those factors that influence plasma concentrations of these micronutrients. We analyzed a cross-sectional study designed to investigate whether certain personal characteristics and dietary factors might influence plasma concentrations of retinol and beta-carotene. The goals of the analysis were

- to confirm the significance of several predictors suggested by the researchers who designed the study,
- to identify clinically relevant variables when building multivariable models to predict concentrations of each plasma micronutrient,
- to test the predictive value of these models against a reserved subset of the data, and
- to present and comment upon the final models, as well as to make recommendations for future study.

Data description

Overview, collection of the data

The data consisted of demographic variables (age and sex), health related variables (smoking status, body mass index, and vitamin use), diet related variables (calories, fat, fiber, and cholesterol consumed per day, alcoholic drinks consumed per week, and dietary beta-carotene and retinol consumed per day) as well as the outcome variables of plasma micronutrient levels (beta-carotene and retinol) for 315 patients at an urban research hospital. All patients had undergone an elective surgical procedure during a three-year period, to biopsy or remove a lesion of the lung, colon, breast, skin, ovary, or uterus that was found to be non-cancerous.

Unusual features of the data

The dataset contained two unusual features. First, one patient was an extreme outlier in terms of both alcohol consumption (203 drinks per week) and caloric intake (6662.2 calories per day). The analyses reported here do not include this patient. Including the patient had the effect of masking the effects of alcohol on plasma retinol levels, but it did not change the significance of any other results.

In addition, one patient had a reported plasma beta-carotene concentration of zero ng/ml. Because analyses were done using log transformed beta-carotene levels as the dependent variable, and because it was unclear whether this value was due to a data-entry error or some other factor, this patient was excluded from analyses involving plasma beta-carotene.

Manipulation of the data

The number of calories and grams of fat a patient consumed per day were used to calculate the percentage of dietary calories that came from fat for each patient.

$$PercentFat = 9(GramsFat)/NumberCalories$$

In addition, the number of drinks per week was recoded to be a categorical variable indicating the patient's overall drinking pattern. A patient either did not drink at all, drank moderately (between one and ten drinks per week), or drank heavily (over 10 drinks per week). These categories were chosen in order to make it easier to interpret the results in terms of clinical recommendations, and in order to eliminate violations of the model assumptions which arose when using number of drinks as an independent variable.

Variable descriptions

Demographics

Patients ranged in age from 19 to 83 years old, with a mean age of 50.1 years. The age variable was close to normally distributed, although with fewer patients below 30 years old than would be expected under a normal distribution. Two hundred seventy-three (86.9%) of the 314 patients included in the analysis were women.

Health

One-hundred fifty-seven (50%) of the patients had never smoked, 115 (36.6%) were former smokers, and 42 (13.4%) were current smokers. Values of the body mass (Quetelet) index ranged from 16.33 to 50.40, with a mean of 26.17 and a median of 24.74. Using this indicator, 101 (32.2%) of the patients would be considered obese. Regarding vitamins, 122 (38.9%) use them fairly often, 82 (26.1%) use them occasionally, and 110 (35%) do not use them.

Diet

Calories patients consumed per day ranged from 445.2 to 4374, with a mean of 1781 calories and a median of 1665 calories. The distribution was slightly right skewed and had two high outliers. Fat consumption ranged from 14.4 to 235.9 grams per day, with a mean of 76.76 grams and a median of 72.9 grams. This distribution was also slightly right skewed, and there were three high outliers. The calculated variable, percentage of calories from fat, ranged from 16.3% to 63%, with mean and median both equal to about 38%. Daily fiber intake ranged from 3.1 to 36.8 grams, with a mean of 12.79 grams and a median of 12.10 grams; the distribution was right skewed with five high outliers. Daily cholesterol intake ranged from 37.7 to 900.7 mg, with a mean of 241.3 mg and a median of 206.2 mg; again the distribution was right skewed with five high outliers.

The variable indicating weekly alcohol consumption was extremely right skewed, with many high outliers. Many patients did not drink at all, and the median number of drinks per week was 0.3 drinks. However, the mean number was 2.643 drinks. With patients categorized as described above, 111 (35.4%) of them did not drink at all, 180 (57.3%) had between one and ten drinks per week, and 23 (7.3%) had more than ten drinks per week.

Distributions for dietary retinol and dietary beta-carotene were both right skewed. Dietary retinol ranged from 30 to 6901 mcg per day, with mean and median of 831 and 707 mcg per day; dietary beta-carotene ranged from 214 to 9642 mcg per day, with mean and median of 2183 and 1795 mcg per day. Retinol had six high outlying values, and beta-carotene had eight.

Plasma micronutrient levels

Both plasma retinol and plasma beta-carotene levels were right skewed as well. Plasma retinol ranged from 179 to 1727 ng/ml, with mean and median of 603.7 and 566.0 ng/ml. Plasma beta-carotene ranged from zero to 1415 ng/ml, with mean and median of 603.7 and 566.0 ng/ml. Because of the right skew in both these variables, and because it is an assumption of ordinary least squares regression that the outcome variable be normally distributed, all analyses in which plasma micronutrient levels were outcomes were performed on the log-transformed values of plasma retinol and plasma beta-carotene. The comparison of the original and log-transformed micronutrient variables to a normal distribution can be seen in Figure 1.

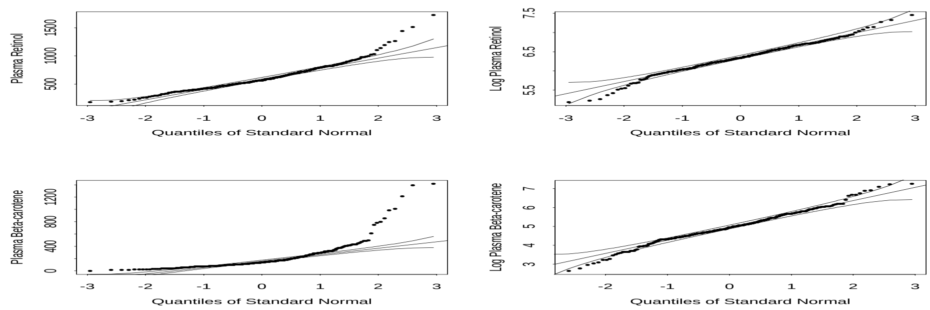

Figure 1: Log Transformations of Plasma Micronutrients

Analysis and results

The following is a summary of rationale behind the analyses, with primary results. For full details please see the Appendix.

Bivariate relationships

It had been suggested by the researchers who designed the study that plasma retinol levels are related to alcohol intake, and that plasma beta-carotene levels are related to body mass (Quetelet index) and vitamin use, as well as cholesterol, calorie, and fiber intake. The following analyses tested these variables as predictors of the respective micronutrients, using the full dataset.

Retinol

Effect of alcohol. The number of alcoholic drinks consumed per week was a significant predictor of log plasma retinol level, but a plot of the residuals from this analysis against the fitted values showed non-constant variance in the residuals, indicating a violation of model assumptions. Because this was not a problem when the analysis was for alcohol consumption category, and because the results are easier to interpret for the categories, only the results of this analysis are presented here.

The effect of alcohol category on log plasma retinol can be seen in the parallel boxplots in Figure 2. A regression using indicator variables for alcohol cateogry yielded the regression equation

$$logPlasmaRetinol = 6.2855 + (0.0776)Moderate + (0.2298)High,$$

indicating that those patients who never drank had a mean plasma retinol level of $e^{6.2855} = 536.73$ ng/ml, whereas moderate drinkers had mean $e^{6.2855+0.0776} = 580.04$ ng/ml, and heavy drinkers had mean $e^{6.2855+0.2298} = 675.40$ ng/ml. The F-test for the overall effect of alcohol category was significant at the 0.01 level.

One can see in Figure 2 that although the means are different, as indicated in the regression equation, there still is considerable variation in log plasma retinol within each category.

Beta-carotene

Effect of body mass. Using the body mass (Quetelet) index to predict log plasma beta-carotene also yielded residuals with non-constant variance, so the Quetelet index was log transformed as well. This regression can be described by the equation

$$logPlasmaBetaCarotene = 8.2508 - 1.0152(logQueteletIndex),$$

Figure 2: Log Plasma Retinol by Alcohol Consumption Category

where the coefficient on log Quetelet index was significant with $p < .0001$. This curvilinear relationship can be seen in Figure 3.

It is worth noting that there is a subset of patients with plasma beta-carotene levels over 600 ng/ml and Quetlet index less than 30 whose plasma concentrations are not well-predicted by this equation.

Effect of cholesterol. Regressing log plasma beta-carotene on cholesterol intake also gave problems with non-constant variance in the residuals, so log cholerol was used as the independent variable. The equation descibing this relationship is

$$logPlasmaBetaCarotene = 5.7863 - 1.01544(logCholesterol),$$

where the coefficient on log cholesterol was marginally significant with $p = 0.0563$. This curvilinear relationship can be seen in Figure 4.

Again, not all patients were well accounted for by the model, particularly those with plasma beta-carotene above 600 ng/ml.

Effect of calories. Regressing daily caloric intake on log plasma beta-carotene did *not* show calories to be a significant predictor of log beta-carotene. However, if one examines the interaction between sex and number of calories on log beta-carotene, there are some mildly suggestive results. Specifically, the summary information for this regression is:

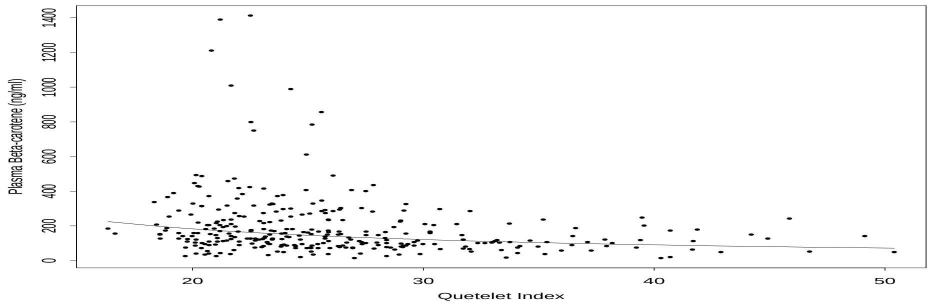

Figure 3: Plasma Beta-carotene by Quetelet Index

```
Coefficients:
                    Value Std. Error  t value Pr(>|t|)
   (Intercept)     5.2229    0.2238   23.3419   0.0000
      Calories    -0.0002    0.0001   -1.6893   0.0922
           Sex    -0.2349    0.2238   -1.0499   0.2946
  Calories*Sex     0.0002    0.0001    1.7498   0.0811

Residual standard error: 0.7416 on 309 degrees of freedom
Multiple R-Squared: 0.02806
F-statistic: 2.973 on 3 and 309 degrees of freedom, the p-value is 0.03197
```

The relationships may be easier to see in Figure 5. The p-value for the overall F-test indicates there is some predictive value in including the effects of calories and sex and their interaction, although the marginal significance of some of the coefficients makes it difficult to see exactly what the relationship is between these variables and log plasma beta-carotene. If one believes there is an interaction between patient sex and calories consumed, then it appears from the graph that caloric intake matters for males and not for females. But if this interaction is not significant, then controlling for patient sex may only make the effect of calories easier to see. It would be reading too much into this data to make conclusions either way.

Effect of vitamin use. Regressing log plasma beta-carotene on indicator variables for vitamin use gives the following equation:

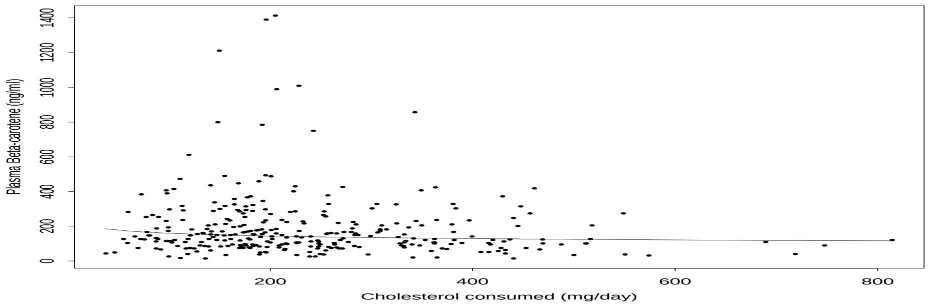

Figure 4: Plasma Beta-carotene by Cholesterol (mg/day)

$$logPlasmaBetaCarotene = 4.7178 + 0.2920(Occasional) + 0.4308(Frequent)$$

This indicates that the mean plasma beta-carotene level for patients who never use vitamins is $e^{4.7178} = 111.92$ ng/ml, whereas the mean for occasional vitamin users is $e^{4.7178+0.2920} = 149.87$ ng/ml, and the mean for frequent users is $e^{4.7178+0.4308} = 172.19$ ng/ml. The overall F-test for the effect of vitamin use is significant with $p < .0001$.

Effect of fiber. Finally, log plasma beta-carotene was regressed on daily fiber intake, yielding the following equation:

$$logPlasmaBetaCarotene = 4.5314 + 0.0336(Fiber),$$

indicating that each additional daily gram of fiber increases mean plasma beta-carotene levels by $e^{0.0336} = 1.034$ ng/ml. The coefficient for fiber was significant at the 0.0001 level. The curvilinear relationship between fiber and plasma beta-carotene can be seen in Figure 6. There were no major assumption violations apparent in plots of the residuals.

Model selection

The goal of model selection in this analysis was to find a set of variables to predict plasma levels of each micronutrient. In order for a variable to be included

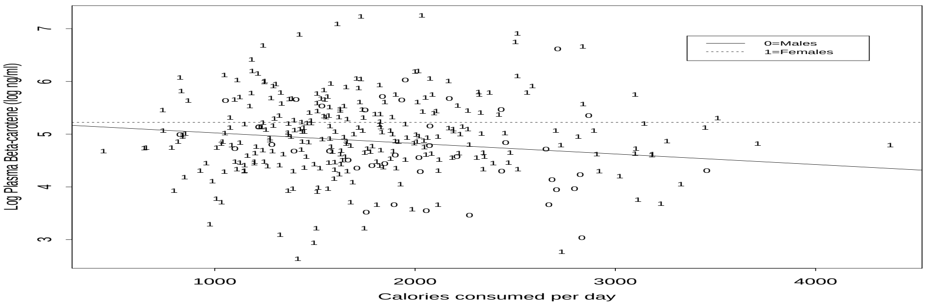

Figure 5: Log Plasma Beta-carotene by Sex and Daily Calories

in either model, it had to have coefficient(s) significantly different from zero, but it also had to be clinically relevant. That is, even if it was statistically significant as a predictor, it had to account for a relatively sizeable difference in micronutrient levels. This was assessed by splitting the data into random model-building and model-validating subsets. The model-building subset consisted of 200 (63.7%) randomly selected cases. Beginning with the predictors that had been associated with each micronutrient in the bivariate analyses, a model was built for each micronutrient, testing the significance of each new variable in the model using nested F-tests to compare the model with and without that variable. When all variables had been tested, there was one preferred model for each micronutrient. This model was compared to the results of an exhaustive stepwise regression procedure, and in each case the preferred model and an alternative model from the stepwise results were retained. These two models for each micronutrient were then compared by examining the predictive value of each against the remaining third of the data (see the model validation section).

Retinol

The initial model for predicting log plasma retinol concentrations included only alcohol consumption category. Including age in the model indicated that age was a significant predictor which increased R-squared from 0.0295 to 0.1028. It was therefore retained in the model. Nested F-tests for including the variables sex, smoking status, body mass, vitamin use, calories, fiber, choesterol, percentage of calories from fat, and dietary retinol were all non-significant, so these variables

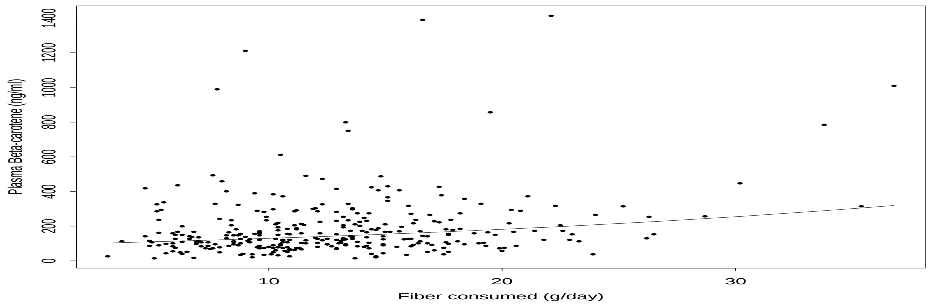

Figure 6: Plasma Beta-carotene by Fiber Intake

were not added to the model. (See the Appendix for details on the F-tests.)

The two-variable model including alcohol consumption and age performed well in the stewise procedure (BIC = -77.8, adjusted R-squared = 0.089), but so did a model including only age (BIC = -77.2, adjusted R-squared = 0.055). These two models were compared using predictions on the test data (below).

Beta-carotene

The initial model for predicting log plasma beta-carotene levels included log body mass, log cholesterol intake, vitamin use, and fiber intake. Again based on nested F-tests to determine which variables to add to the model, smoking status was added, whereas age, percentage of calories from fat, alcohol consumption category, and dietary beta-carotene were not. (See Appendix for details.)

Next, the preferred, five variable model, which included log body mass, log cholesterol intake, vitamin use, fiber intake, and smoking status was compared to the results of exhaustive stepwise regression. Of the top ten models in terms of BIC, log body mass was in all of them, vitamin use was in five, fiber was in seven, and smoking status was in six. Log cholesterol, however, was not in any of the top ten models.

In order to determine whether to retain log cholesterol in the model, I examined a partial regression plot (Figure 7), which indicated that log cholesterol does little to explain the variance left in the residuals once we have already regressed log plasma beta-carotene on the other four predictor variables. So the preferred model was revised to include only log body mass, vitamin use, fiber intake, and smoking status. In the next section this is compared to a model which also includes log cholesterol.

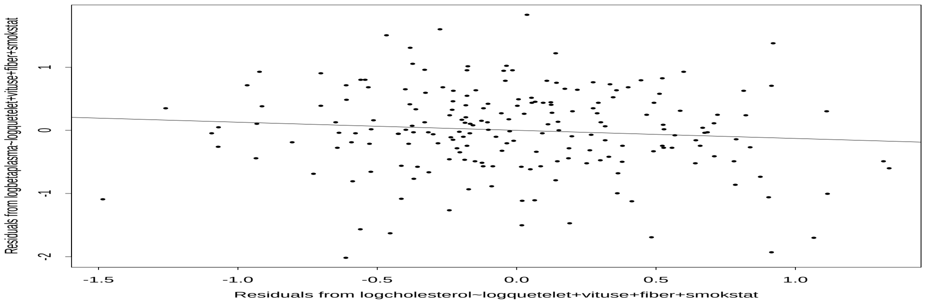

Figure 7: Partial Regression Plot for Log Cholesterol

Model validation, summary of final models

In order to test the predictive power of the preferred model against the alternative model for each micronutrient, predictions were generated based on the respective independent variables for the test data (n=114). Then, plots of the fitted values against the actual values were compared for each model. Finally, the residuals were compared for each model. The final model for each micronutrient is summarized below.

Retinol

Recall that for predicting log plasma retinol, the preferred model was one with coefficients for both age and alcohol consumption category, whereas the alternative model included a coefficient for age only. The predictions are plotted against the true values for each of these models in Figure 8. Including alcohol consumption in the model improved these predictions moderately, and it decreased the range of the residuals, so this variable was retained.

The summary of the final model describing log plasma retinol is then:

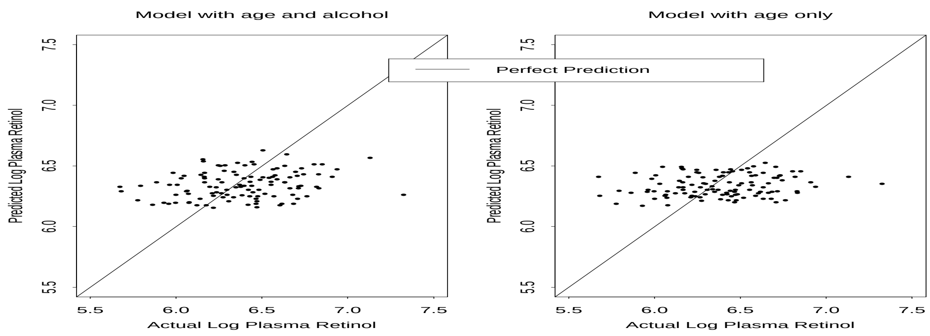

Figure 8: Comparison of Models for Predicting Log Plasma Retinol

```
Coefficients:
               Value Std. Error t value Pr(>|t|)
(Intercept)  5.8990  0.0966    61.0898  0.0000
     alcMod  0.1357  0.0519     2.6155  0.0096
      alcHi  0.2371  0.1031     2.2989  0.0226
        age  0.0067  0.0017     4.0008  0.0001

Residual standard error: 0.3421 on 196 degrees of freedom
Multiple R-Squared: 0.1028
F-statistic: 7.486 on 3 and 196 degrees of freedom, the p-value is 9.053e-05
```

Increasing alcohol consumption tends to increase plasma levels, as does age. Specifically, moving from the non-drinking to the moderately drinking group increases the predicted plasma retinol concentration by a factor of $e^{0.1357} = 1.15$, whereas moving from the non-drinking to the heavy drinking group increases the predicted plasma retinol concentration by a factor of $e^{0.2371} = 1.27$. Each additional year of age increases predicted plasma retinol concentration by a factor of $e^{0.0067} = 1.01$.

Examination of various diagnostic plots of the residuals showed no violations of model assumptions.

Beta-carotene

For predicting log plasma beta-carotene, the preferred model was one with coefficients for log body mass, vitamin use, fiber, and smoking status, whereas

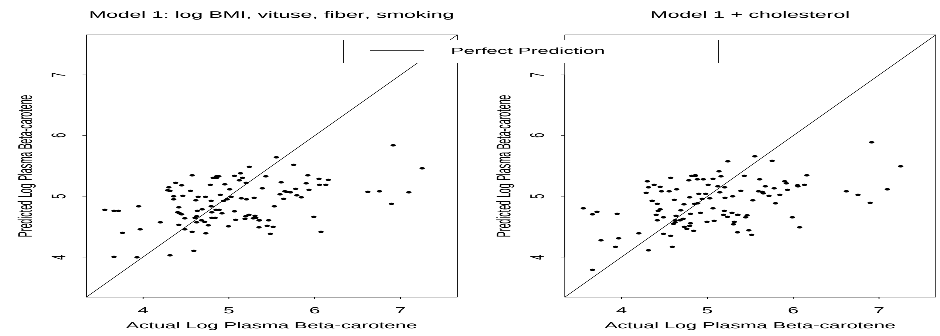

Figure 9: Comparison of Models for Predicting Log Plasma Beta-carotene

the alternative model also included a coefficient for log cholesterol intake. The predictions based on these models can be seen in Figure 9. Including log cholesterol has an almost imperceptible effect on the predicted values, and it does not substantially reduce the range of the residuals, so it is not included in the final model.

The summary of the final model describing log plasma beta-carotene is then:

```
Coefficients:
                Value Std. Error  t value Pr(>|t|)
 (Intercept)   8.0089   0.7788    10.2837   0.0000
 logquetelet  -1.0670   0.2286    -4.6664   0.0000
      vitMod   0.2760   0.1256     2.1969   0.0292
       vitHi   0.2756   0.1149     2.3978   0.0175
       fiber   0.0228   0.0094     2.4339   0.0159
 smokeFormer  -0.1825   0.1041    -1.7535   0.0811
smokeCurrent  -0.4366   0.1563    -2.7932   0.0057

Residual standard error: 0.6788 on 192 degrees of freedom
Multiple R-Squared: 0.211
F-statistic: 8.56 on 6 and 192 degrees of freedom, the p-value is 2.996e-08
```

Unit increases in log body mass decrease the predicted plasma beta-carotene concentration by a factor of $e^{1.0670} = 2.91$. Compared to not using vitamins at all, being either a occasional or frequent user of vitamins tends to increase plasma beta-carotene concentrations to a similar degree; moving from the non-using to the occasionally using group increases the predicted plasma beta-carotene concentration by a factor of $e^{0.2756} = 1.317$, whereas moving from the non-using to the frequently using group increases it by a factor of $e^{0.2756} = 1.318$. Each additional gram of fiber per day increases the predicted plasma beta-carotene concentration by a factor of $e^{0.0228} = 1.02$. Finally, smoking behavior tends to decrease plasma concentrations of beta-carotene; compared to a non-smoker, a former smoker has a predicted plasma beta-carotene level which is $e^{0.1825} = 1.2$ times lower, whereas a current smoker has a level which is $e^{0.4366} = 1.55$ times lower.

Again, examination of the model residuals revealed no violations of model assumptions.

Discussion and conclusions

Initial bivariate regression confirmed that alcohol consumption increases mean plasma retinol levels. The multivariate model for predicting plasma retinol concentration added age as an independent variable, and this two variable model performed better than a model including age only when predicting the random test subset of the data. For predicting plasma beta-carotene, initial bivariate regressions confirmed that (log) body mass decreases the mean levels, whereas vitamin use and higher fiber intake increases them. No significant effect was found for calories, although there were marginally significant effects when using an interaction between calories and sex. Although (log) cholesterol was a significant predictor of plasma beta-carotene by itself, in multivariate regression it was removed from the model. Body mass, vitamin use, and fiber were retained, and smoking status was added. When these four variables were in the model, cholesterol had little predictive power for either the original or test data.

Clinical implications

The two best predictors of plasma retinol concentration, age and alcohol intake, present problems in terms of clinical recommendations. Age is not under the patient's control, and the biggest increases in plasma retinol occur when one drinks heavily. The best recommendation would be for patients to drink moderately, as this increases mean retinol levels somewhat and is not associated with significant physical and mental health problems.

Clinical recommendations to increase plasma levels of beta-carotene are a bit more promising. First, as being overweight decreases these levels, patients should maintain a healthy weight. Vitamin use is effective in increasing plasma beta-carotene, but there is no apparent benefit of being a frequent user over an occasional user. Increasing fiber intake is another recommendation that could be made. Finally, although former smokers have decreased levels of beta-carotene compared to non-smokers, they are better off than current smokers, so quitting smoking may have some benefit here as well.

Research implications

It is clear from the values of R^2 for the final models and from the relatively large error in the model predictions (Figures 8 and 9) that there is still much variation in both plasma retinol and plasma beta-carotene which is left to be explained, even after testing the predictive power of a range of demographic, health related, and diet related variables on these concentrations. A greater understanding of how variable cartenoid levels are across time for individuals would give a better indication of whether the current study design, which measured cartenoid levels at one timepoint, was appropriate. If levels vary over time, then taking multiple measurements and predicting mean concentration for each individual may give much better predictive results, by eliminating some random noise from the dependent variables. However, if plasma concentrations are relatively stable over time, then it might be best to investigate other variables for predicting these micronutrients, as the best models are currently inadequate.

Finally, because this study was conducted on a relatively restricted subset of the population, patients who had undergone elective surgery to biopsy or remove a non-cancerous lesion, it is uncertain whether the distributions of cartenoid concentrations in this sample, or the factors associated with changes in these concentrations, would be similar for the population in general.