International Journal of Epidemiology

Vol. 10, No. 3
Printed in Great Britain

A Multivariate Assessment of Alcohol Consumption

JOHN B WHITFIELD,* JANET K ALLEN,* MICHAEL ADENA,** HUGH G GALLAGHER† and WILLIAM J HENSLEY*

Whitfield JB [Department of Biochemistry, Royal Prince Alfred Hospital, Camperdown, New South Wales 2050, Australia], Allen JK, Adena M, Gallagher HG and Hensley WJ. A multivariate assessment of alcohol consumption. *International Journal of Epidemiology* 1981, **10**: 281–288.

Subjects attending a large, multiphasic health screening centre in Sydney, Australia estimated their alcohol consumption and specimens of their blood were analysed. The most useful univariate estimates of alcohol consumption were erythrocyte mean corpuscular volume and plasma aspartate-aminotransferase, gamma-glutamyl-transpeptidase, triglycerides and uric acid. The most statistically significant of these tests have been combined to form a multivariate predictor of alcohol intake which is more successful in identifying heavy-drinkers than single tests. To describe this population further, and to aid comparisons between populations, information about non-drinkers has also been provided.

Alcohol abuse is a major health problem, in terms of mortality and morbidity, psychosocial disruption, and economic effects.[1,2] Early detection of alcohol abuse through objective measurements, e.g. laboratory tests, would be desirable for 2 reasons. Intervention at an early stage might be beneficial and the natural history and pathology of alcohol-related diseases could be studied. Most individuals do not seek medical attention in these early stages for alcohol-related problems, but the problem could be detected if they saw a physician for other reasons.

Patients frequently under-estimate their long-term alcohol consumption, either deliberately or inadvertently. Many groups have attempted to detect alcoholism or heavy-drinking. Some of the tests used are: blood alcohol,[3] erythrocyte mean corpuscular volume, MCV,[4] plasma gamma-glutamyl-transpeptidase, GGT,[5,6] aspartate-aminotransferase, AST,[5] alanine-aminotransferase, ALT,[5] triglycerides, TG,[7] uric acid, UA,[8] total cholesterol, CH,[9,10] branched-chain amino acids[11] and γ-amino-n-butyric acid.[12] GGT is perhaps the most widely used parameter for measuring alcohol intake, but increased GGT may be caused not only by high alcohol consumption, but also by liver disease or by the ingestion of drugs, especially the barbiturates.[13] Elevated GGT is not observed in all heavy-drinkers,[5] but only in 44% of male and 27% of female admitted heavy-drinkers, while it is elevated in about 6.5% of male and female light-drinkers.[14] Although striking correlations were found for single test results, we felt it worthwhile to combine the results.

* Department of Biochemistry, Royal Prince Alfred Hospital, Camperdown, New South Wales 2050, Australia.
** Department of Population Biology, Research School of Biological Sciences, The Australian National University, Canberra ACT 2601, Australia.
† Medicheck Referral Centre, 65 Bathurst Street, Sydney, New South Wales 2000, Australia.

METHODS

Subjects

The 3597 subjects attended the Medicheck Referral Centre, a large metropolitan multi-phasic health testing centre in Sydney, Australia. A blood sample was collected by venipuncture from the subjects who had fasted for at least 12 hours. The subjects also answered a computer administered questionnaire of generally high validity.[16]

Biochemical

The heparinised blood sample was assayed for sodium, potassium, bicarbonate (BC), total protein, albumin, phosphorus, cholesterol (CH), fasting glucose, urea nitrogen (UN), uric acid (UA), bilirubin, alkaline phosphatase (AP) and calcium, with a Technicon SMA 12/60 using standard Technicon

methods. The enzymes AST and GGT were measured by reaction rate[17,18] and TG was measured enzymatically.[19] The erythrocyte mean corpuscular volume (MCV) was determined using a Coulter-S counter.

Statistical

To derive suitable prediction equations for the rate of ethanol consumption, the following 11 steps were carried out using the SPSS computer package:[20]

(1) Regression and validation groups The subjects were divided into 2 groups. The first group of about two-thirds of the subjects was used to derive regression equations, and these equations were validated using the second group. For internal control purposes, subjects attending Medicheck are assigned successive patient identification numbers between 10 and 899. Subjects with a patient number less than 600 were considered as the regression group, and the remaining subjects comprised the validation group. The regression group was used in steps (2) to (6), while the validation group was reserved for steps (7) to (10).

(2) Dependent variable From the responses to questions on alcohol consumption in the computer administered questionnaire, the approximate number of alcoholic drinks consumed per month was inferred. The type of alcoholic beverage was not considered, because standard drinks of different beverages each contain about 15 ml of ethanol. Admitted daily alcohol consumption of 6 or more drinks (>90 ml ethanol daily) defined a heavy drinker.

The accuracy of drinks per month as a measure of the rate of alcohol consumption is not the same at all levels of consumption, violating the assumption of homogeneous variance in linear regression analysis. In addition, interpretation of a negative value of this measure, as might arise from a prediction equation, is difficult. Accordingly, this measure was logarithmically transformed to approximate homoscedacity. This corresponds to a constant percentage error variance in the original scale. Furthermore, negative predicted values have a natural interpretation in terms of fractions of a standard drink per month.

Subjects claiming not to drink alcohol were excluded from the remaining steps in the derivation of prediction equations, partly because the concern here was with drinkers not non-drinkers. Also, non-drinkers cannot be represented on the log consumption scale, and have a variance different from that of the drinkers. However, non-drinkers were considered when the regression equations were evaluated (see Results).

(3) Independent variables The 16 blood biochemistries and MCV were screened for biologically implausible values. In addition, subjects whose MCV was less than 70 fl were excluded as possible thalassemics.

Although not essential for the validity of the analysis, linear regression is more robust when the empirical distributions of the independent variables are approximately normal.[21] In each sex, the variables CH, UA, AP, AST, GGT and TG required logarithmic transformation, while the remaining 11 variables required no transformation. An additional advantage of transformation to approximate normality is that it often enhances the linearity of the relationships between the variables.

(4) Multiple regression equations For each sex, a stepwise multiple linear regression for the logarithm of the imputed number of alcoholic drinks consumed per month in terms of the variables described in step (3) was performed. The sexes were treated separately because preliminary analyses combining them were unsatisfactory. Because of the large number of independent variables, the 'F-to-enter' for new variables being considered as part of an expanded regression equation was set at 10.83, the 0.1% level of a $F_{1,\infty}$ distribution. If, say, 15 of the independent variables are not relevant to prediction, this would give a probability of $(1-.001)^{15} = 0.985$ of none of these superfluous variables appearing in the final regression equation. The variables MCV, $\log_{10}$ (GGT), and $\log_{10}$ (UA) comprised the regression to estimate the logarithm of the imputed rate of alcohol consumption in males. For each variable, the F statistic for the variable to be dropped from the equation was high. For females, only the first 2 variables appeared.

(5) Residuals plots The residuals from these regression equations for each sex were plotted against the date of the subject's visit to Medicheck and also against the subject's biochemical variables and MCV. There were no consistent, non-random patterns in any of these residuals plots, indicating that the regression equations were adequate descriptions of the data. In particular, the necessity for quadratic or interaction terms was ruled out. The residuals were also homoscedastic with respect

to the predicted logarithm of the rate of alcohol consumption, supporting the choice of the logarithmic transform in step (2).

(6) Correlations of the residuals with other variables Correlations between the residuals and the biochemical variables not in the regression equation were calculated. Because of the large number of correlation coefficients examined, their significance should be interpreted with care. Two correlations, that with phosphate in males and with glucose in females, were significant at the 5% level, but with 14 or 15 correlations in each sex, these are not unexpected.

(7) Residuals in the validation group Residuals for subjects in the validation group were calculated using the regression equations found at step (4). These residuals were examined in the following 3 steps.

(8) Validation group residuals plots Scattergrams were plotted of the validation group residuals with the date of the subject's visit, each of the subject's biochemical variables, and MCV. In each plot, the residuals showed no non-random patterns. In addition, these plots were compared with the corresponding plots from step (5) by superposition over an x-ray illuminator. The pairs of plots were similar. These residuals plots, in a sample of subjects different from that used to derive the regression equation, provide excellent validation of the derived equation.

(9) Correlations of the validation group residuals with the independent variables Correlations with all biochemical variables and MCV were calculated. The correlations with variables in the regression equations were not significant. All correlations corresponding to those that were significant at the 5% level at step (6) were not significant. If $\log_{10}$ (UA) was dropped from the male regression equation, the validation group residuals were significantly correlated with $\log_{10}$ (UA) indicating that this variable is necessary in the regression.

(10) Homogeneity of the regression parameters A regression analysis in the validation group comparing the regression from the regression group with one estimated independently in the validation group[22] showed that the parameters from the 2 regressions were homogeneous (males: $F_{4,554} = 0.87$, N.S. $p > 0.05$; females: $F_{3,250} = 2.45$, N.S. $p > 0.05$).

(11) Final estimates of the regression parameters The 2 groups were combined and the regression parameters re-estimated from data from all the subjects. Note that these estimates are optimal for this sample of subjects only, and the regression equation may not perform as well for other subjects (but see step (10) above).

Standard errors of proportions Standard errors of proportions were calculated using the formula: standard error $(p^* (1 - p) /n)^{1/2}$ where p is the proportion and n the sample size.

RESULTS AND DISCUSSION

For men, the values of MCV, uric acid, triglycerides, GGT and AST increase with increasing admitted alcohol intake.[14]

For most of tests the variances increase with increasing alcohol consumption. Therefore, various

TABLE 1 *Correlations between Log_{10} (alcohol consumption) and transformed biochemical results for male and female drinkers.*

Test	Men (n = 1503)	Women (n = 738)
MCV	0.388***	0.335***
Log_{10} (GGT)	0.355***	0.203***
Log_{10} (AST)	0.211***	0.069
Log_{10} (UA)	0.217***	0.111**
Log_{10} (TG)	0.136***	0.032
Log_{10} (CH)	0.152***	−0.019
Age	0.064*	0.041
Multivariate Predictor	0.476***	0.364***

*$0.01 < p < 0.05$, **$0.001 < p < 0.01$, ***$p < 0.001$

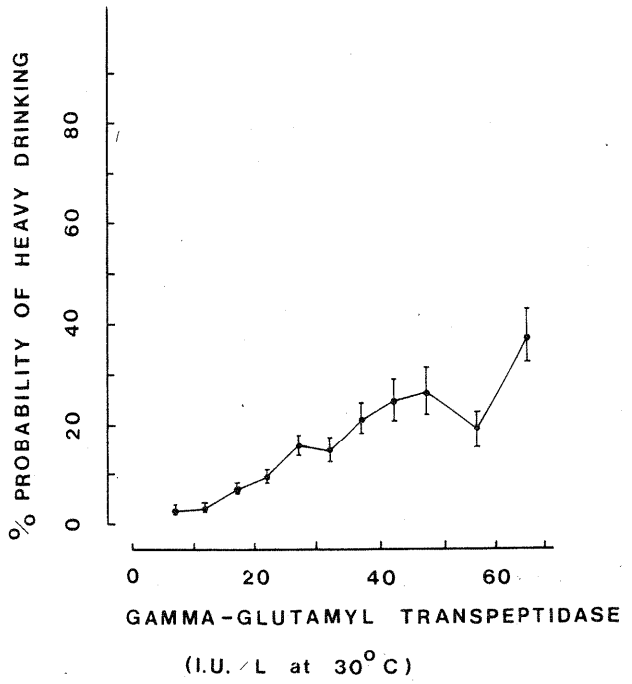

FIGURE 1a

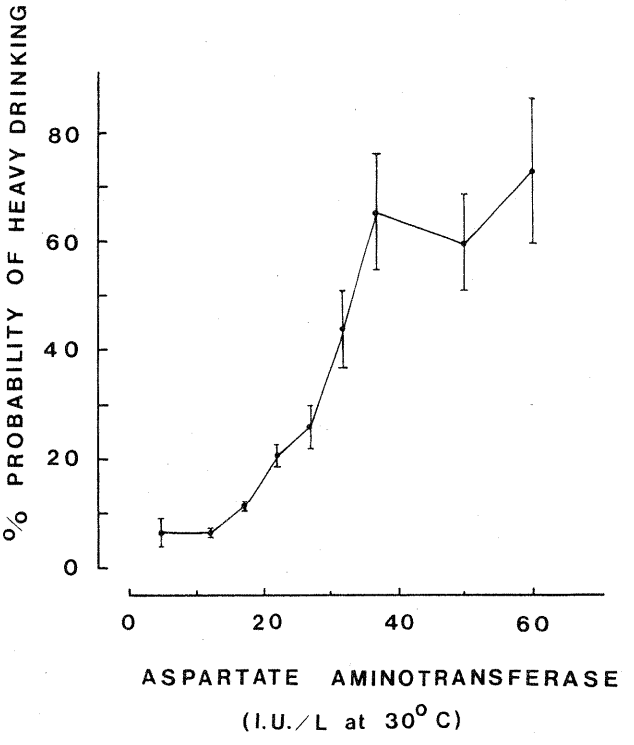

FIGURE 1b

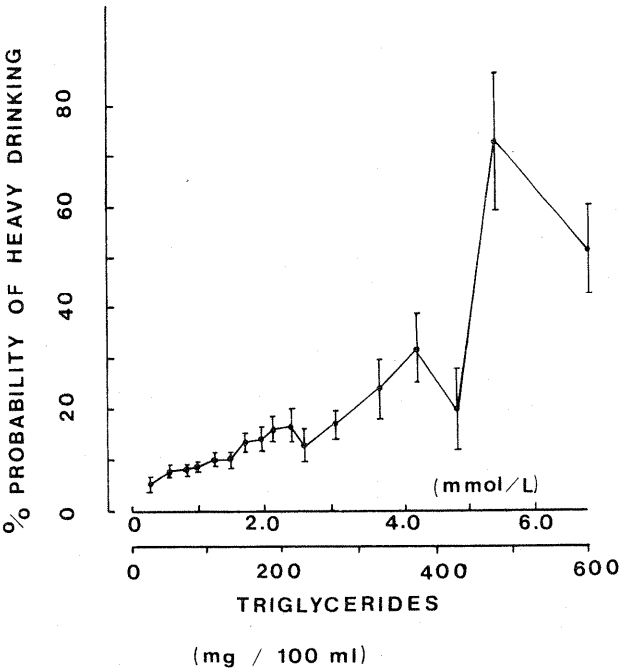

FIGURE 1c

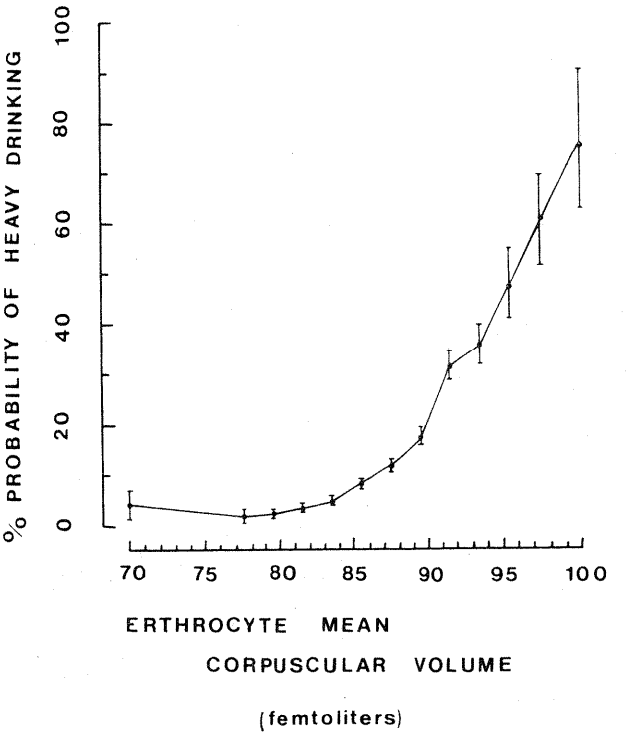

FIGURE 1d

FIGURE 1 *The empirical percentage probability in males of being an admitted heavy-drinker versus test results for: (a) gamma-glutamyl transpeptidase, (b) aspartate aminotransferase, (c) triglycerides, (d) erythrocyte mean corpuscular volume and (e) uric acid. Both drinkers and non-drinkers are included. GGT is grouped in units of 5 IU per litre, AST is grouped as 5 IU per litre, triglycerides in groups of 20 mgm per 100 ml, MCV in groups of 2 femtolitres and uric acid in groups of 5 mgm per 100 ml.*

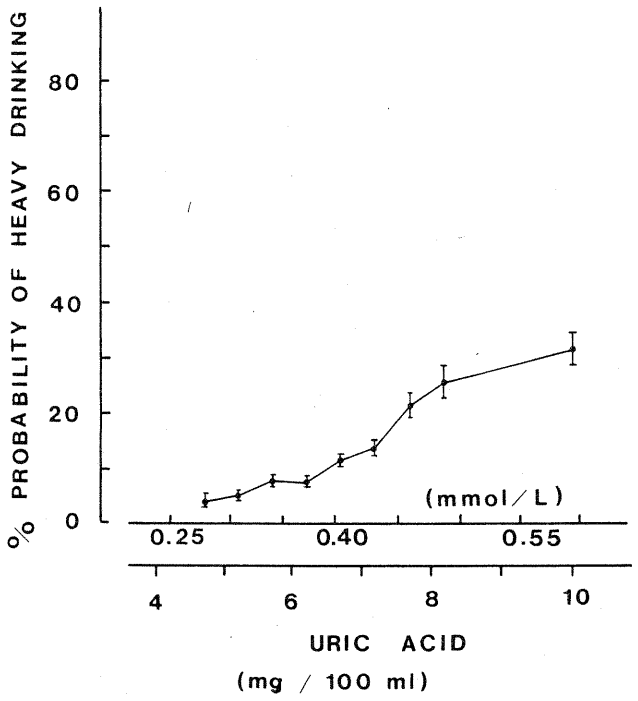

FIGURE 1e

transformations of the data were tested until homoscedacity was attained as discussed in Methods. The base 10 logarithm of the variables GGT and UA of the predictor itself, estimated ethanol consumption yield the most useful equation.

Combining the most significant parameters gives a multivariate estimation of monthly ethanol consumption, EETOHC, for men:

$$\text{LOG (EETOHC)} = 0.3572 \text{ LOG (GGT)} + 0.0392 \text{ MCV} + 0.8082 \text{ LOG (UA)} - 2.0477.$$

and for women:

$$\text{LOG (EETOHC)} = 0.2170 \text{ LOG (GGT)} + 0.0348 \text{ MCV} - 2.1441.$$

MCV is measured in femtolitres, GGT in IU per litre at 30°, and uric acid in mmol per litre.

For men, the correlation coefficient between the test and alcohol consumption increases from 0.388 for MCV, the best single variable to 0.476 using the multiple regression equation. For women, the correlation coefficient increases from 0.335, for MCV, to 0.364 using the 2 variables in this regression equation. Although these improvements in the correlation coefficient are small, the increase in the proportion of variance explained, R^2, is approximately 50% for males. However, there remains a large error in estimating an individual's alcohol intake which precludes using the multivariate regression equation as a direct assessment of an individual's alcohol consumption, but, as will be shown below, it may be used successfully to estimate the probability that an individual is a heavy drinker.

There are several possible reasons for the large error in estimating an individual's alcohol intake. Firstly, the declared alcohol intake may not accurately reflect the actual intake, and these physiological variables must assess actual intake rather than the amount declared. Secondly, GGT probably correlates not with alcohol consumption itself, but with liver damage arising from alcohol consumption, and the relationship between alcohol consumption and liver damage shows considerable individual variation. Thirdly, there are correlations between the variables and therefore the gain by adding extra variables is not simply additive.

In spite of these difficulties, multivariate regression may be used to identify probable heavy drinkers and is more successful than single variable equations are. Figure 1 shows for men the empirical percentage probability of being an admitted heavy drinker as a function of the results of 5 separate laboratory tests MCV, GGT, AST, uric acid and triglycerides. Generally the higher the test result, the greater the probability of being an admitted heavy drinker, although only MCV, AST, TG and the multivariate predictor are sufficiently related to alcohol consumption to give an empirical probability of being a heavy drinker of 50%.

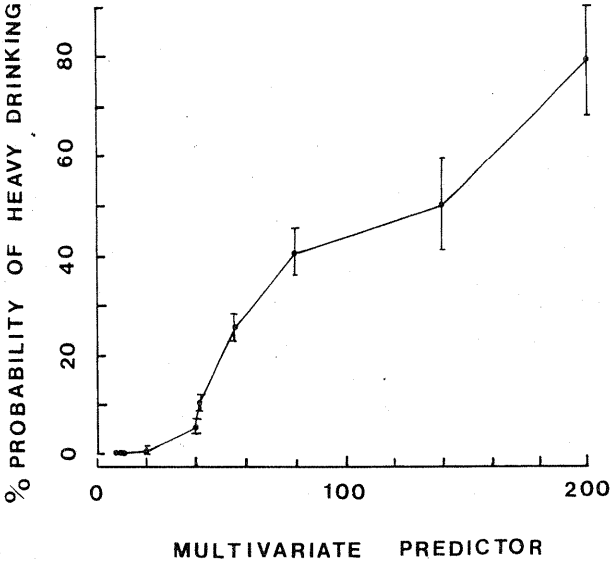

FIGURE 2 *The empirical percentage probability in males of being an admitted heavy-drinker versus the multivariate predictor. Both drinkers and non-drinkers are included. Data from subjects with multivariate predictor values in 10 unit ranges are grouped. Groups are formed from subjects whose values fall between 0 and 9.99, 10 and 19.99, etc.*

TABLE 2 *Empirical classification success for males (both drinkers and non-drinkers). The percentages of correct classifications are shown at the test values giving a) a 50% probability, or b) a 19% probability (that is, double the population risk) of being a heavy drinker.*

a) Test	Sensitivity, heavy-drinkers correctly classified	Specificity, non-heavy-drinkers correctly classified
MCV	84%	54%
Log_{10} (GGT)	88%	55%
Log_{10} (AST)	76%	47%
Log_{10} (UA)	70%	52%
Log_{10} (TG)	66%	51%
Multivariate Predictor	92%	55%

b) Test	Sensitivity, heavy-drinkers correctly classified	Specificity, non-heavy-drinkers correctly classified
MCV	50%	87%
Log_{10} (GGT)	29%	81%
Log_{10} (AST)	39%	83%
Log_{10} (UA)	44%	83%
Log_{10} (TG)	35%	82%
Multivariate Predictor	66%	86%

Figure 2 shows the relationship between the multivariate predictor, estimated ethanol consumption, and the empirical probability of being a heavy drinker.

A comparison of the effectiveness of single tests and the multivariate predictor is shown in Table 2. It might be desirable to follow up all subjects with a risk of being a heavy drinker above a certain level, for example above 50%, or above twice the average population risk. We have compared the tests using sensitivities and specificities determined from the entire sample. This will tend to over-estimate the performance of the tests.

With either cut-off value, the multivariate predictor gives a more favourable combination of sensitivity and specificity than any of the other tests do. This trade-off between sensitivity and specificity is illustrated more clearly in Figure 3.

Perhaps the greatest difficulty in using multivariate assessments is the variation between populations in the coefficients used. Populations may be contaminated to varying degrees with the condition being studied (in this case, the prevalence of heavy drinking), or with confounding conditions,

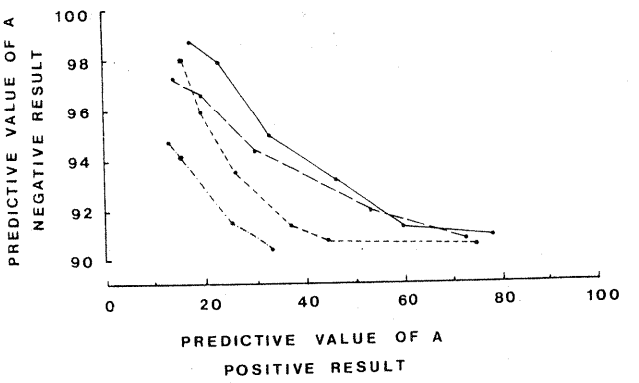

FIGURE 3 *The predictive value of a positive result versus the predictive value of a negative result using various biochemical assessments of heavy drinking in males. Each biochemical assessment is evaluated with several arbitrary cut-off values as the demarcation between heavy drinkers and non-heavy drinkers.*
——— *indicates the multivariate predictor,*
– – – – *MCV,* · · · · · · · *LOG_{10} (GGT) and*
– · – · *LOG_{10} (UA). The optimal predictor would have a 100% predictive value for a positive result and a 100% predictive value for a negative result.*

TABLE 3 *Biochemical and haematological parameters in non-drinkers in Sydney, Australia; means and standard deviations.*

	Men (n = 424)	Women (n = 632)
Mean Corpuscular Volume (fl)	83.9 ± 4.30	84.6 ± 4.50
Gamma Glutamyl-Transpeptidase (IU/l)	17.9 ± 13.6	14.4 ± 14.9
Uric Acid (mmol/l)	0.342 ± 0.065	0.260 ± 0.059

(e.g. liver disease, use of therapeutic drugs). Genetic and environmental differences also contribute to this variation as does laboratory analytical variation. Table 3 shows means and standard deviations for those who do not drink in our population. These figures could be compared with values obtained for other populations and provide a basis for applying this multivariate predictor to other populations. By far the most significant variable in the multiple regression equation is the erythrocyte mean corpuscular volume. Fortunately, among non-drinkers, it appears to be tightly controlled with a narrow distribution. Further, between populations, this variable seems to be relatively constant, as long as proper quality control is maintained.[23,24,25] Therefore, it is expected that these multiple regression equations will be similar to those needed for other populations, although the information given in Table 3 provides a basis for adjusting the equations to other populations.

CONCLUSIONS

Multivariate predictors of alcohol intake have been developed, using the variables erythrocyte mean corpuscular volume, $\log_{10}$ (gamma-glutamyl-transpeptidase) and $\log_{10}$ (uric acid.) These multivariate predictors may be used to identify heavy-drinkers more successfully than predictors based on single tests alone, as shown in Table 2. Further, these multivariate predictors could reasonably be expected to apply to other populations since erythrocyte mean corpuscular volume, the most important variable in the multivariate predictor, appears to be closely regulated, with similar values in different populations.

ACKNOWLEDGEMENT

We are grateful to Professor John Gibson and the staff of the Department of Population Biology at the Australian National University, Canberra for suggestions and the use of computing facilities. Mr D Bryden assisted with the initial data reduction. One of us (MAA) was supported by a grant to Professor Gibson from the AW Tyree Foundation, Sydney.

REFERENCES

1 Work in Progress on Alcoholism. FA Seixas and S Eggleston (eds). *Ann NY Acad Sci* 1976; **273**: 5–77.

2 Luce BR and Schweitzer SO. Smoking and alcohol abuse: a comparison of their economic consequences. *N Engl J Med* 1978; **298**: 569–571.

3 Hamlyn AN, Brown AJ, Sherlock S and Baron D. Casual blood-ethanol estimations in patients with chronic liver disease. *Lancet* 1975; **ii**: 345–347.

4 Unger KW and Johnson D Jr. Red blood cell mean corpuscular volume: a potential indicator of alcohol usage in a working population. *Am J Med Sci* 1974; **267**: 281–289.

5 Patel S and O'Gorman P. Serum enzyme levels in alcoholism and drug dependency. *J Clin Pathol* 1975; **28**: 414–417.

6 Whithead TP, Clarke CA and Whitfield AGW. Biochemical and haematological markers of alcohol intake. *Lancet* 1978; **i**: 978–981.

7 Ginsberg H, Olefsky J, Farquhar JW and Reaven GM. Moderate ethanol ingestion and plasma triglyceride levels, a study in normal and hyper-triglyceridemic persons. *Journal of Internal Medicine* 1974; **80**: 143–149.

8 Olin JS, Devenyi P and Weldron KL. Uric acid in alcoholics. *Quarterly Journal of Studies on Alcohol* 1973; **34**: 1202–1207.

9 Barboriak JJ. Hyperlipemia in alcoholics. *Quarterly Journal of Studies on Alcohol* 1974; **35**: 15–19.

10 Ostrander LD, Lamphiear DE, Block WD, Johnson BC, et al. Relationship of serum lipid concentrations to alcohol consumption. *Arch Inter Med* 1974; **134**: 451–456.

11 Shaw S and Lieber SC. Plasma amino acid abnormalities in the alcoholic. *Gastroenterology* 1978; **74**: 677–682.

12 Shaw S, Lue S-L and Lieber CS. Biochemical tests for the detection of alcoholism: comparison of

plasma alpha-amino-n-butyric acid with other available tests. *Alcoholism: Clinical and Experimental Research* 1978; **2**: 3–7.

13 Spencer-Peet J, Wood DCF and Glatt MM. Screening test for alcoholism. *Lancet* 1973; ii: 1089–1090.

14 Whitfield JB, Hensley WJ, Bryden D and Gallagher H. Some laboratory correlates of drinking habits. *Ann Clin Biochem* 1978; **15**: 297–303.

15 Goldberg DM and Ellis G. Mathematical and computer-assisted procedures in the diagnosis of liver and biliary tract disorders. *Adv Clin Chem* 1978; **20**: 49–128.

16 Rawson G. Patient attitudes to screening by an automated multiphasic health testing facility. *Aust Fam Physician* 1975; **4**: 219–222.

17 Karmen A. A Note on the spectrophotometric assay of glutamic oxalacetic transaminase in human blood serum. *J Clin Invest* 1955; **34**: 131–133.

18 Szasz G. A kinetic photometric method for serum gamma-glutamyl transpeptidase. *Clin Chem* 1969; **15**: 124–136.

19 Bucolo G and David H. Quantitative determination of serum triglycerides by the use of enzymes. *Clin Chem* 1973; **19**: 476–482.

20 Nie NH, Hull CH, Jenkins JG, Steinbrenner K and Bent DH. Statistical Packages for the Social Sciences, Second Edition. New York: McGraw-Hill Book Company, 1975.

21 Box GEP and Watson GS. Robustness to non-normality of regression tests. *Biometrika* 1962; **49**: 93–106.

22 Williams EJ. Regression Analysis. New York: John Wiley and Sons, Inc., 1959, p 81.

23 Okumo T. Red cell size as measured by the Coulter model S. *J Clin Pathol* 1972; **25**: 599–602.

24 Hamilton PJ and Davidson LR. The inter-relationships and stability of Coulter S-determined blood indices. *J Clin Pathol* 1973; **26**: 700–705.

25 Prangnell DR and Johnson PH. A new method of quality control for the Coulter model S counter. *J Clin Pathol* 1977; **30**: 487–491.

(Revised version received 21 January 1981)