pected. However, it was correctly predicted that all of the aromatic BGH's examined would be L1210 active, although the rank ordering of these was clearly different from that observed. With the view that such ranking might reflect the different screening protocols employed, it was then desirable to apply regression analysis directly to the biologic data. However, with only eight tumor-active aromatic BGH's available, there are barely enough data points to adequately examine the importance of one variable, and it has been clearly demonstrated that in vivo antitumor activity is modeled successfully only by multivariable analysis. Although no single term could be found significant at the 5% level, in modeling ILS values for the aromatic BGH, the parameter which enters first in a forward, stepwise, multiple linear regression is log $[C_{50}(G\text{-}C)]/[C_{50}(A\text{-}T)]$, the measure of DNA discriminating ability being employed. This latter parameter is also the single most important variable in modeling of the antitumor activities of the BQAH.[3]

While it has been suggested that the antitumor activity of the aromatic BGH's is related to their ability to inhibit DDP in vivo,[8,9] the present study shows that this property is related to the in vivo toxicity of the general class of BGH and may indeed be a useful predictor of this toxicity. However, if the aromatic BGH's are congeneric with the BQAH, then it would be expected that the antitumor selectivity is dependent on in vivo binding to an as yet undefined, alternating A-T rich site(s) in the tumor cell DNA.

Experimental Section

R_m values were determined by the chromatographic method detailed earlier for the BQAH,[3] employing Merck cellulose F_{254} DC sheets as support. UV-absorbing compounds were detected by their fluorescence quenching of the cellulose support. UV-transparent compounds, for example, synthalin (**18**), were located by spraying with pentacyanoaquoferriate reagent, prepared as follows: Equal volumes of cold 10% aqueous solutions of sodium nitroprusside, potassium ferricyanide, and sodium hydroxide were mixed together, in that order, and the red-orange solution stood at room temperature until the color faded to a clear yellow-green (20–25 min). H_2O (4 volumes) and Me_2CO (3 volumes) were then added, and the solution was used immediately.

Where adequate quantities of the BQH salts were available, weighed amounts were employed to prepare standard aqueous solutions of 2 mM strength. For those agents in short supply, aqueous solutions were prepared and the concentration of these was determined from the UV data quoted by Mihich et al.,[9] final dilutions being made to provide 2 mM solutions.

Acknowledgment. We are grateful to Mrs. E.-M. Falkenhaug for performance of the many C_{50} determinations. We are indebted to Dr. E. Mihich for samples of many BGH, without which this study could not have been undertaken. This work was supported by the Auckland Division of the Cancer Society of New Zealand (Inc.) and, in part, by the Medical Research Council of New Zealand.

References and Notes

(1) G. J. Atwell and B. F. Cain, *J. Med. Chem.*, **10**, 706 (1967).
(2) B. F. Cain, G. J. Atwell, and R. N. Seelye, *J. Med. Chem.*, **12**, 199 (1969).
(3) W. A. Denny, G. J. Atwell, B. C. Baguley, and B. F. Cain, *J. Med. Chem.*, **22**, 134 (1979).
(4) A. C. Counsilman, B. F. Cain, and B. C. Baguley, *Eur. J. Cancer*, **10**, 75 (1974).
(5) B. C. Baguley and E. M. Falkenhaug, *Nucleic Acids Res.*, **5**, 161 (1978).
(6) E. Mihich, *Cancer Res.*, **23**, 1375 (1963).
(7) E. Mihich, *Cancer*, **20**, 880 (1967).
(8) E. Mihich, "Antineoplastic and Immunosuppressive Agents", A. C. Sartorelli and D. G. Johns, Ed., Springer-Verlag, New York, 1975, p 766.
(9) C. Dave, M. J. Ehrke, and E. Mihich, *Chem.-Biol. Interact.*, **16**, 57 (1977).
(10) C. Dave, M. J. Ehrke, and E. Mihich, *Proc. Am. Assoc. Cancer Res.*, **12**, 40 (1971).
(11) C. Dave, M. J. Ehrke, and E. Mihich, *Chem.-Biol. Interact.*, **12**, 183 (1975).
(12) C. Dave and L. Caballes, *Fed. Proc., Fed. Am. Soc. Exp. Biol.*, **32**, 736 (1973).
(13) M. T. Hakala, *Biochem. Pharmacol.*, **20**, 81 (1971).
(14) A. C. Sartorelli, A. T. Iannotti, B. A. Booth, F. H. Schneider, J. R. Bertino, and D. G. Johns, *Biochim. Biophys. Acta*, **103**, 174 (1965).
(15) E. G. Podrebarac, W. H. Nyberg, F. A. French, and C. C. Cheng, *J. Med. Chem.*, **6**, 283 (1963).
(16) G. Cavallini, E. Massarani, D. Nardi, L. Mauri, and D. Mantegazza, *J. Med. Pharm. Chem.*, **4**, 177 (1961).
(17) C. Zimmer, *Prog. Nucleic Acid Res. Mol. Biol.*, **15**, 285 (1975).
(18) B. C. Baguley, unpublished data, this laboratory.
(19) M. J. Waring, *J. Mol. Biol.*, **54**, 247 (1970).
(20) J. Soucek, M. J. Ehrke, and E. Mihich, *Cancer Res.*, **30**, 2187 (1970).

Chance Factors in Studies of Quantitative Structure–Activity Relationships

John G. Topliss* and Robert P. Edwards

Schering-Plough Research Division, Bloomfield, New Jersey 07003. *Received April 2, 1979*

Multiple regression analysis is a basic statistical tool used for QSAR studies in drug design. However, there is a risk of arriving at fortuitous correlations when too many variables are screened relative to the number of available observations. In this regard, a critical distinction must be made between the number of variables screened for possible correlation and the number which actually appear in the regression equation. Using a modified Fortran stepwise multiple-regression analysis program, simulated QSAR studies employing random numbers were run for many different combinations of screened variables and observations. Under certain conditions, a substantial incidence of correlations with high r^2 values were found, although the overall degree of chance correlation noted was less than that reported in a previous study. Analysis of the results has provided a basis for making judgements concerning the level of risk of encountering chance correlations for a wide range of combinations of observations and screened variables in QSAR studies using multiple-regression analysis. For illustrative purposes, some examples involving published QSAR studies have been considered and the reported correlations shown to be less significant than originally presented through the influence of unrecognized chance factors.

During the past decade, quantitative structure–activity relationships (QSAR) have been increasingly used in drug-design studies.[1] Typically, a number of possible independent variables, usually physicochemical parameters relating to a series of compounds, are evaluated for correlation with activity values using multiple-regression analysis.[2] The correlation equation which emerges from this analysis may contain only a small number of inde-

0022-2623/79/1822-1238$01.00/0

pendent variables out of many evaluated. The independent variables in the equation and the overall equation itself may be highly significant by standard statistical criteria. However, these criteria only relate to variables in the final equation and do not take into account how many independent variables were actually screened for possible inclusion in the equation. Clearly, the larger the number of possible independent variables considered, the greater the possibility that a correlation will occur purely by chance and this will not be reflected in the standard statistical criteria for the equation. It is therefore very important to consider not only the number of independent variables in the correlation equation itself but how many such possible variables were screened in the statistical analysis.

By way of illustration, take a case where there are ten observations and corresponding activity values, A_1 to A_{10}, and eight variables, V_1 to V_8, are considered. Multiple-regression analysis yields eq 1, where $r^2 = 0.85$, $F_{2,7} = 19.8$

$$A = aV_1 + bV_2 + c \quad (1)$$

($p < 0.005$), and V_1 and V_2 are each significant at the level $p < 0.01$. However, the statistical test for significance of the equation only relates to the variables included in the actual correlation equation and does not take account for the fact that eight variables were screened for possible correlation. The p value of <0.005 for the equation shown thus overstates to some extent the probability that the relationship given by the equation is real rather than occurring by chance. The question is: how misleading is this p value? The studies to be described will attempt to provide some answer to this question.

Earlier work on this problem was reported some years ago.[3] However, these studies were too limited in scope to provide more than a demonstration that, indeed, there was a problem from chance correlations under certain conditions and to provide very general guidelines. Accordingly, in view of the interest in this problem following publication of the earlier study, plans were made to generate more extensive and detailed data.

Method

The basic approach was to set up a large number and wide range of simulated correlation studies using random numbers for both observations and screened variables and assess the incidence and degree of any resulting correlations.

Computer System. The program used was developed by modifying an existing Fortran multiple-regression analysis program (BMD 02R), the essence of the modifications being to generate the data to be analyzed using a pseudo-random number generator and to accumulate statistics from each of the individual simulations and produce output reports, frequency distributions, and summary tables. The program was run on an IBM 360/158 computer performing under OS/VS2.

Random Number Generator. The pseudo-random numbers were generated using the IBM RANDU subroutine[4] which employs the multiplicative congruential method[5] and returns a vector of random numbers, each uniformly distributed in the range 0 to 1. The random numbers were scaled by a factor of 1000. Starting from a user-defined seed value (1 478 653 was used for this study), RANDU generates a stream of pseudo-random numbers and will produce 2^{29} terms before repeating. By comparison of the observed against the expected frequency distribution (uniform) of numbers, RANDU is an adequate pseudo-random number generator. To better approximate pure randomness, each computer run had a different externally imposed initial seed value. In addition, only every fourth pseudo-random number was sampled from the stream.

Computer Program Methodology. The first step in each simulation is to generate the data matrix with r rows (one for each observation) and $n + 1$ columns (n independent variables and one dependent). The values of r and n are specified when the program is run. The data are then analyzed using a stepwise multiple-regression technique.[2] At each stage of this procedure, the independent variable most highly correlated with the dependent variable is brought into the model, provided that the partial F-test value for that variable is significant at the 10% level. Each independent variable in the regression model is then reexamined to see if it is still making a significant contribution. Any variable whose partial F-test value is not significant at the 10% level is dropped from the model. This process continues until no more variables enter the model and none are rejected.

From each simulation, the number of variables entering the regression (0 to n) is stored, as well as the r^2 statistic in the case of at least one variable entering the regression analysis. This whole process is repeated for each simulation, the number of simulations to be carried out being specified when the program is run. The results of all the simulations are then consolidated to give the following statistics: (1) the number and relative frequency for 0, 1, 2, 3, etc. variables entering the regression analysis; (2) the number and relative frequency for at least one variable entering the regression analysis; (3) for each significant correlation observed: (a) mean r^2 statistic, (b) median r^2 statistic, (c) minimum r^2 statistic, (d) maximum r^2 statistic; (4) the number and relative frequency of observed correlations by r^2 interval: 1.0–0.9, 0.9–0.8, 0.8–0.7, etc.

Scope of Study. The number of variables screened was 3–15, 20, 30, and 40, and the number of observations was 5–30, 50, 100, 200, 400, and 600. Altogether, some 300 combinations of variables and observations were examined. The number of runs for each combination ranged from 120 to 2190. It should be noted that the study was originally designed for far fewer combinations of variables and observations and runs per combination. However, due to the unexpected availability, for a short period, of a computer which was being phased out, the scope of the study was greatly expanded.

Results

Representative data obtained in the study are presented in Table I. The columns in the table give the number of independent variables; the number of observations; the number of runs; the frequency of a chance correlation; the average number of variables entered (for those runs for which at least one variable is entered); the maximum, minimum, and mean r^2 values; and the frequency of those runs where $r^2 \geq 0.5$ and 0.8. It should be noted that p values for the observed correlations ranged from a minimum value of 0.10 up to 0.00001.

For the situation with ten independent variables screened and ten observations, some two-thirds of the runs gave a correlation and for these an average of 2.20 variables entered. r^2 ranged from 0.28 to 1.00 with a mean value of 0.64. Almost half of the total runs made gave correlations with r^2 values of 0.5 or higher, while about one-fourth of the total runs made gave correlations with r^2 values of 0.8 or higher. As the number of observations increased, there was a downward trend in r^2 values, so that for 15 observations only 4% of the total runs gave correlations with $r^2 \geq 0.8$ and for 30 observations no correlations with $r^2 \geq 0.8$ were found.

For 20 variables screened and only 15 observations, practically all of the runs resulted in some level of chance correlation with about 44% of the total runs resulting in a correlation with $r^2 \geq 0.8$. As the number of observations increased to 30, the number of runs yielding correlations with $r^2 \geq 0.8$ declined to about 1%.

At this point, it must be noted that the degree of chance correlation found in this study was substantially less than that reported in an earlier study.[3] Mean r^2 values averaged about one-half those reported in the earlier study, ranging from three-fourths down to one-fourth. The cause of this discrepancy could not be determined, since the basic records from the earlier study were not retained. However, it must be concluded that the earlier study was in error.

Table I. Data from Simulated Correlations Using Random Numbers

no. of ind var	no. obs	no. of runs	freq chance corr.	av. no. of var. entered	max r^2	min r^2	mean r^2	freq $r^2 > 0.5$	freq $r^2 > 0.8$
3	5	2190	0.26	1.26	1.00	0.65	0.83	0.26	0.15
3	10	2190	0.27	1.15	0.95	0.30	0.47	0.09	0.01
3	20	2190	0.28	1.14	0.63	0.14	0.24	0.01	0.00
5	5	120	0.44	1.87	1.00	0.61	0.80	0.44	0.28
5	7	398	0.39	1.54	1.00	0.45	0.73	0.34	0.15
5	10	2190	0.40	1.41	0.99	0.30	0.54	0.21	0.04
5	20	2190	0.41	1.30	0.79	0.14	0.28	0.03	0.00
10	10	120	0.67	2.20	1.00	0.28	0.64	0.46	0.23
10	12	398	0.71	2.09	1.00	0.23	0.57	0.41	0.14
10	15	398	0.71	1.92	0.98	0.18	0.46	0.28	0.04
10	30	2190	0.74	1.84	0.74	0.08	0.23	0.03	0.00
15	10	120	0.88	4.42	1.00	0.26	0.79	0.74	0.54
15	15	120	0.87	3.33	1.00	0.19	0.60	0.56	0.28
15	17	398	0.86	2.77	0.98	0.16	0.53	0.46	0.14
15	20	2190	0.85	2.61	0.97	0.13	0.45	0.34	0.06
15	30	2190	0.85	2.48	0.86	0.08	0.30	0.11	0.00
20	15	120	0.94	5.00	1.00	0.17	0.73	0.76	0.44
20	20	120	0.97	3.53	0.98	0.13	0.56	0.58	0.17
20	30	2190	0.95	3.53	0.94	0.08	0.39	0.29	0.01
20	50	1991	0.95	3.41	0.73	0.04	0.24	0.03	0.00
30	50	1991	0.99	5.59	0.76	0.04	0.37	0.19	0.00
30	100	1991	0.99	5.17	0.52	0.02	0.18	0.00	0.00
40	50	199	1.00	7.14	0.90	0.04	0.49	0.44	0.02
40	100	199	1.00	7.04	0.47	0.03	0.24	0.01	0.00

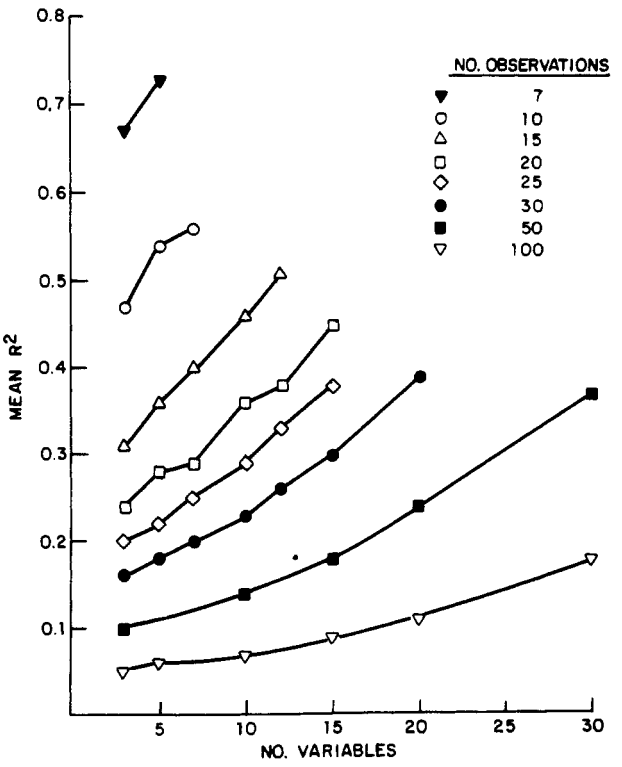

Figure 1. Relationship between mean r^2 values and number of variables screened.

The present study was very carefully checked, and selected combinations set up and performed manually showed similar results.

From data generated in the present study, it is now possible to answer the question raised ealier concerning the p value of eq 1. With ten observations and eight screened variables, the expected frequency of chance correlations comes out to be about 12%. Thus, the validity of the correlation equation is far less certain than the reported p value indicates.

The results are further illustrated in Figures 1–8. The relationship between r^2 and the number of variables screened for varying numbers of observations is illustrated in Figure 1. It can be seen that for a constant number of observations r^2 increases as the number of variables screened increases. Also, for a given number of variables screened, r^2 increases as the number of observations decreases.

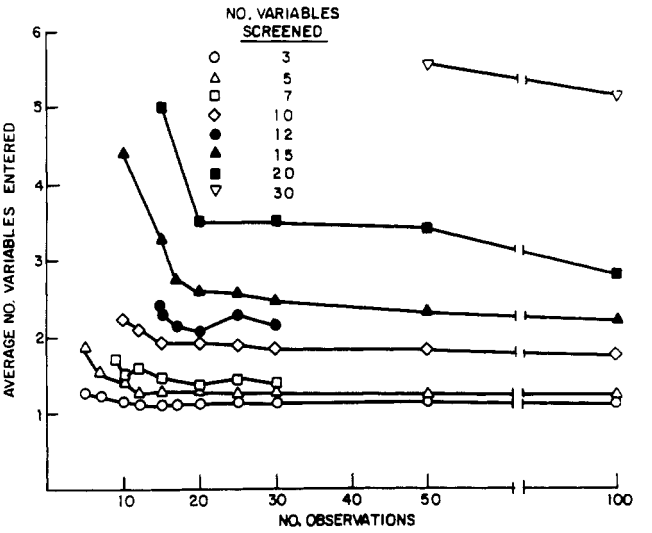

Figure 2. Relationship between the average number of variables (independent) entered in the correlation equation and the number of observations for different numbers of screened variables.

The relationship between the average number of variables (independent) entered in the correlation equation and the number of observations for different numbers of screened variables is illustrated in Figure 2. For a particular number of variables screened, the average number of variables entering the equation is relatively constant, showing a slight increase with a decrease in the number of observations over a wide range in the number of observations. However, when the number of observations approaches the number of variables screened there is a steep increase in the average number of variables entering. As might be expected, there is an increase in the number of variables entered as the number of variables screened increases.

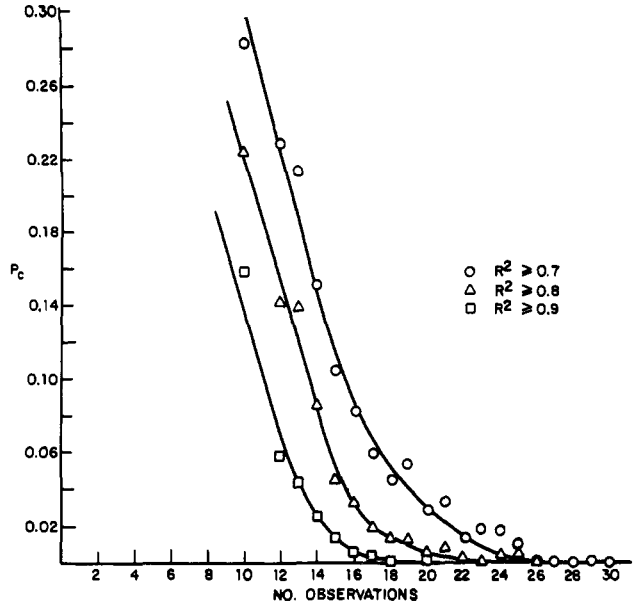

Figure 3. Relationship between the number of observations and the probability of a chance correlation (P_c) for ten screened variables with $r^2 \geq 0.7$, 0.8, and 0.9.

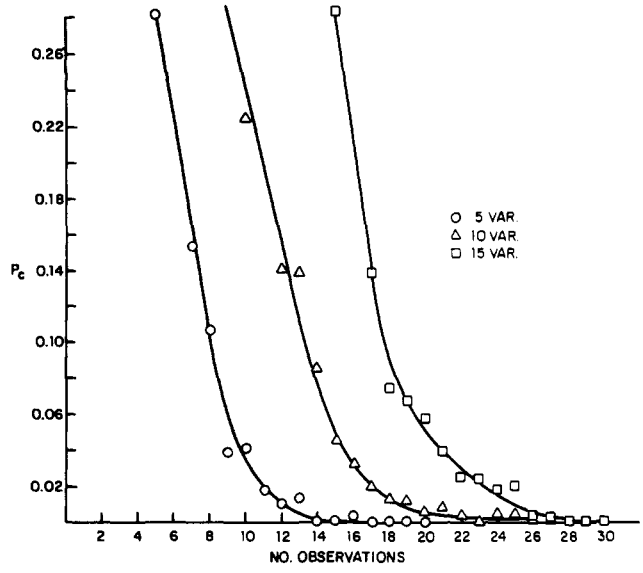

Figure 4. Relationship between the number of observations and the probability of a chance correlation for 5, 10, and 15 screened variables with $r^2 \geq 0.8$.

In Figure 3, the relationship between the number of observations and the probability of a chance correlation[6] with ten screened variables for various r^2 values is shown. From the graph corresponding to $r^2 \geq 0.8$, it can be seen that the probability of encountering a chance correlation at this level is about 22% for ten observations, reaching 0 for 23 observations.

The graph shown in Figure 4 gives the relationship between the number of observations and the probability of a chance correlation with $r^2 \geq 0.8$ for 5, 10, and 15 variables. Thus, for 5 screened variables the probability of encountering a chance correlation is about 22% for 6 observations, 10% for 8 observations, 3% for 10 observations, and 1% for 12 observations.

A guideline often stated is that a certain number of observations are required in order to have much confidence in a given correlation equation. This rule of thumb is deficient on two counts: first, variables screened, rather than variables included in the equation, should be considered; second, the number of observations per variable is not a linear function of the number of variables, as it pertains to a specified level of chance correlation. In Figure 5 are plotted the number of observations per variable against the number of variables for chance correlation levels of 1% for r^2 values greater than or equal to 0.5, 0.8 and 0.9. Considering the $r^2 \geq 0.8$ curve, the number of observations per variable is 3.7 for 3 variables, dropping to 1.75 for 13 variables.

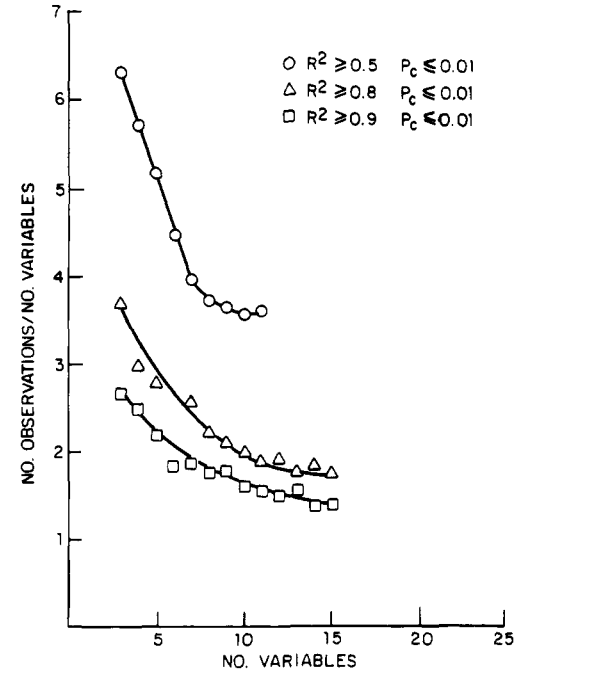

Figure 5. Number of observations per number of variables as a function of the number of variables for specified levels of chance correlation.

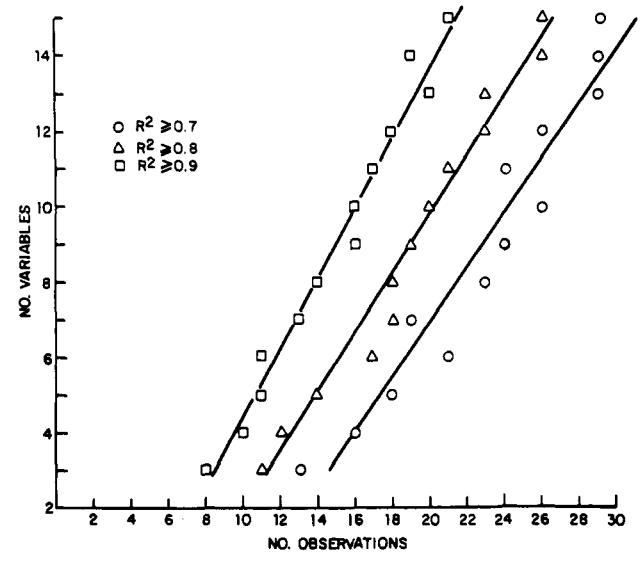

Figure 6. Relationship between the number of observations and the number of variables for chance correlation level $P_c < 0.01$, with $r^2 \geq 0.7$, 0.8, and 0.9.

Figure 6 shows the relationship between observations and variables for a chance correlation probability of 1% or less for various r^2 levels. The linear relationships shown are each statistically significant at the $p < 0.0001$ level. This graph permits the determination of the number of observations required to screen, for example, ten variables while keeping the probability of encountering a chance correlation with $r^2 \geq 0.8$ at the 1% level or less. From the

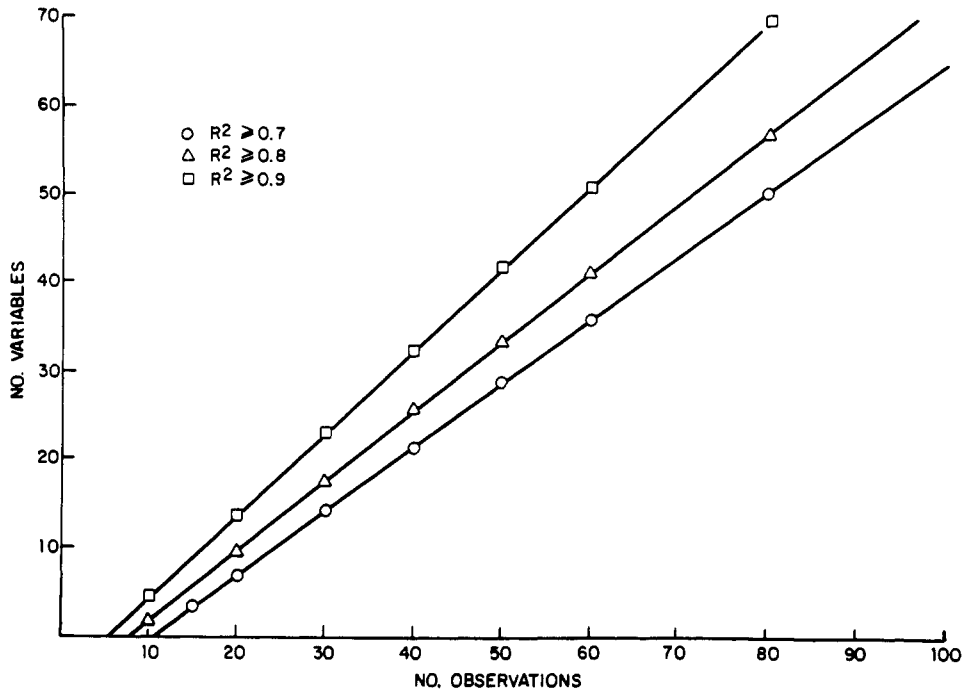

Figure 7. Relationship between the number of observations and the number of variables for chance correlation level $P_c < 0.01$, with $r^2 \geq 0.7$, 0.8, and 0.9 (extrapolated).

graph it can be estimated that this number of observations is about 20. For $r^2 \geq 0.9$, the number required is less, about 16.

Extrapolation of the previous graph allows one (Figure 7) to make similar estimations for larger numbers of screened variables. Thus, some 56 observations would be required to screen 40 variables while keeping the probability of encountering a chance correlation with $r^2 \geq 0.8$ at the 1% level or less.

The preceding graphs make possible the estimation of the number of observations needed to screen a specified number of variables while maintaining the probability of encountering a chance correlation of defined r^2 level at 1% or less. In Figure 8, the graph shows the effect on the number of required observations as the tolerated chance correlation probability increases to 5 or 10%. Again the depicted linear relationships are each statistically significant at the $p < 0.0001$ level. Thus, for 10 variables, 20 observations are needed at a chance correlation level of 1%, which reduces to 15 at the 5% level and 13–14 at the 10% level.

Some limitations must be pointed out to the approach described in this study for estimating probable chance correlation levels. In actual practice, screened variables have varying degrees of collinearity and, not infrequently, some variables are highly collinear. Generally, collinearity will be more in evidence in actual practice among a group of screened variables than among random-number simulated screened variables. To the extent that this occurs, it has the effect of reducing the number of independent variables actually operative as such. To reference a group of screened variables used in an actual problem with the random-number data, it is therefore suggested that when the correlation coefficient between two screened variables is 0.8 or higher only one of these variables should be counted in arriving at the effective number of independent variables screened. This approximate correction device should avoid serious overestimation of chance correlation effects.

A second limitation has to do with use of the stepwise multiple-regression method. This method probably misses some correlations and, as such, leads to some underestimation of the incidence of chance correlations.

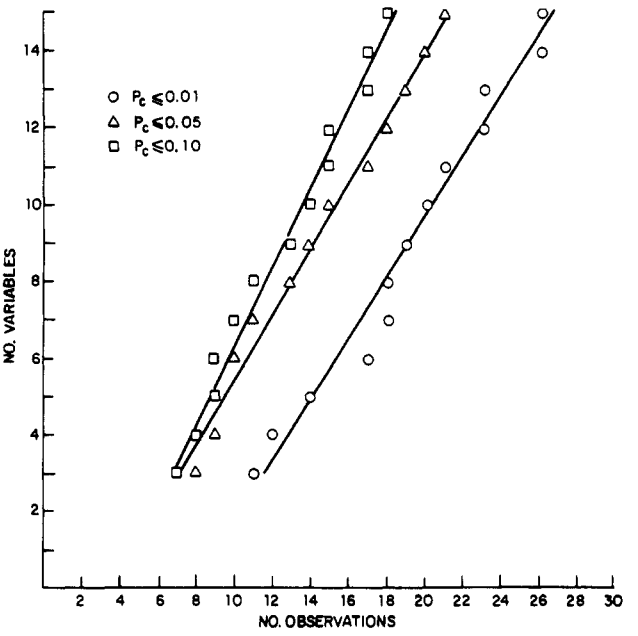

Figure 8. Relationship between the number of observations and the number of variables with $r^2 \geq 0.80$ at specified probability levels.

Overall, the residual bias after correction for the collinearity effect tends to counterbalance the bias from use of the stepwise regression method, so the net estimation error may not be that large.

Applications to Reported Correlation Studies. In the light of the chance correlation phenomenon described, some comments will be made on a number of correlation studies reported in the literature which involved large numbers of screened variables.

The first of these reported by Peradejordi et al.[7] concerned a quantum chemical approach to structure–activity relationships of tetracycline antibiotics. The correlation

eq 2 contained seven independent variables, each significant at the 1% level with the overall equation significant at the 0.1% level. The r^2 value given was 0.986, so that the equation accounts for almost all of the variance in the data. There were 20 observations, and some 18 possible independent variables were screened. The random-number simulated correlation data generated in the current study did not include the combination of 20 observations and 18 variables. The closest reference points were 20 observations and 20 variables, for which the incidence of chance correlations with $r^2 \geq 0.9$ was about 8% and 17 observations and 15 variables with a 5% chance correlation incidence. The approximate risk of a chance correlation is, therefore, 5–8%. The reported correlation equation, although it still may be valid, thus appears far less secure than the reported p value of <0.001 would indicate. It should be noted that we are not saying that there is a 5–8% probability that the reported equation could arise by chance. We are saying that from 20 observations and 18 variables there is a 5–8% probability that some correlation with $r^2 \geq 0.9$ could emerge from chance factors alone.

$$\log K_1' = 18.3975 + 56.1733Q_{O_{10}} + 16.9155E_{O_{10}} + 48.7956Q_{O_4} - 1.1063E_{O_{11}} + 71.3254Q_{O_{12}} + 18.3655E_{O_{12}} + 3.3880Q_{C_6} \quad (2)$$

Gibbons et al.[8] reported on quantitative structure–activity relationships among selected pyrimidinones and Hill reaction inhibition. The correlation eq 3 arrived at was

$$R_pI_{50} = -0.46(\pm 0.11)\pi \text{ ring} + 8.92(\pm 1.85)\sigma_{R_3} + 0.045(\pm 0.008)\sum MR - 0.15 \quad (3)$$

based on 17 observations, contained three independent variables, and had an $r^2 = 0.83$ and a p value of <0.01. Some 20 variables were screened for possible inclusion, which may be equivalent to about 18 variables allowing for collinearity factors. This gives the combination of 17 observations and 18 variables. The nearest reference points available for chance correlations are the combinations of 20 observations and 20 variables and 15 observations and 15 variables, for which the correlation frequencies for $r^2 \geq 0.8$ are 17 and 28%, respectively. It must, therefore, be concluded that there is a substantial risk that the reported equation is spurious, in sharp contrast with the reported p value of <0.01.

A study of Timmermans and van Zwieten[9] on QSAR in centrally acting imidazolines structurally related to clonidine led to the development of eq 4. There were 27

$$\log 1/ED_{30} = -0.00032(\pm 0.00008)(\sum Par)^2 + 0.105(\pm 0.03)\sum Par - 0.695(\pm 0.17)\Delta pK_a^\circ + 5.333(\pm 1.89)HOMO(P) + 6.752(\pm 2.25)EE(P) + 2.494 \quad (4)$$

$$n = 27,\ r^2 = 0.91,\ p < 0.001$$

observations and effectively perhaps 30–35 possible independent variables screened. For 27 observations and 35 variables using random numbers, the frequency of occurrence of correlations with $r^2 \geq 0.9$ was approximately 45%, with approximately 11 variables entering on the average for all significant correlations. For 27 observations and 30 variables there is about a 15% frequency with, on the average, seven variables entering overall. It should be noted that in the correlations obtained with random numbers the average number of variables involved exceeds the five in the equation of Timmermans and van Zwieten. Therefore, the probability that the actual equation reported by them is spurious is less than the 15–45% chance correlation incidence suggests, since r^2 increases as the number of independent variables in the equation increases. Nevertheless, it can be seen that under the conditions of the Timmermans and van Zwieten study chance factors could be important and have to be addressed. Certainly, the stated p value for the equation of <0.001 is misleading.

Discussion

An important point which needs to be considered in assessing the role of chance factors in correlation studies is that some variables screened may have a greater probability of being significant activity correlators than others, whereas the simulation studies treat all screened variables alike. In this sense, a conservative stance has been taken with regard to interpretation of the simulated correlation studies regarding the degree of chance correlation risk inherent with any particular combination of screened variables and observations. It is clear that some parameters will have a much stronger physical or mechanistic basis than others as potential activity determinants. Thus, a π term will be a more probably meaingful activity correlator than a $\pi\sigma$ cross-product term. A good approach in correlation studies where a large number of potential variables could be considered would be to initially select for the correlation study, where possible, a limited group of preferred variables. Any correlation which emerged would then be unlikely to be clouded by chance factors. A second study could then be conducted utilizing the complete set of variables. Variables found to be activity correlators in the second study but not in the first would be viewed as higher risk in terms of chance effects. The extent to which this procedure proves useful will of course depend on how well the total variable set can be justifiably subdivided into primary and secondary sets.

Having established that a particular correlation may have a meaningful risk associated with it from chance factors, the correlation should then be further examined to more precisely define this influence. In the case of correlation eq 5, which has been generated from screening

$$Y = aX_1 + bX_2 + cX_3 + dX_4 + e \quad (5)$$

variables X_1 to X_n, Y values can be replaced by sets (100 or more) of random numbers and the correlation analysis can be repeated with variables X_1 to X_n. In assessing the validity of eq 5, the key point would be the incidence and r^2 values of significant correlations with four independent variables. Such a procedure has been used by Kier and Hall.[10,11] They also searched for correlations where all X values were replaced by random numbers and real Y values were retained. Since this does not preserve the pattern of collinearity present in the real X variable set, this latter procedure may not provide as good an estimate of chance effects. It is also possible to replace selected X variables only, with random numbers. This would be useful in situations where it is suspected that in a correlation equation certain terms, e.g., π and π^2, were real correlators while others were spurious. In this case, real π and π^2 numbers would be retained and the other X variables replaced by random numbers.

In evaluating correlations for chance effects, attention should also be paid to whether some observations have been dropped from the set in order to obtain a better data fit. This is a fairly common occurrence and usually gives much higher r^2 values. However, sometimes the justification given for dropping the poorly fit observations is questionable. The correlation equation obtained from utilizing all of the observations (i.e., after adding back in the dropped observations) should therefore be checked for

possible chance correlations, to give a better overall evaluation of the reliability of the result.

Conclusion

The results of the studies described show that chance correlations are a real phenomenon occurring when the number of variables screened for possible correlation is large compared to the number of observations. For this reason, some correlations are less significant than their standard p values indicate, as has been demonstrated by reference to some reported correlations.

The present study provides guidelines for the approximate incidence of chance correlations at specified r^2 values for various combinations of observations and screened variables. These data may be used prospectively in planning correlation studies. Thus, for a given lead structure around which further synthesis is being planned, the number of relevant independent variables to be considered may be related to the number of compounds planned for synthesis, so that an unacceptable risk of chance correlation will not be present. Specific correlations obtained under conditions where the guidelines indicate an appreciable risk of chance correlations should be individually checked as described.

Finally, it should be pointed out that there is no intention of advocating correlation studies only under conditions where there is a miniscule risk of chance correlations. However, the importance of having a true idea of the reliability of a correlation equation is obvious. It is one thing to develop and use a correlation equation which is known to be somewhat tenuous but quite another to believe that it has a solid foundation when, in fact, it has not.

Acknowledgment. The authors are indebted to L. Weber for assistance in data tabulation and to M. Miller and A. Saltzman for helpful discussions.

Supplementary Material Available: The complete version of Table I containing data from simulated correlations using random numbers (34 pages). Ordering information is given on any current masthead page.

References and Notes

(1) Y. C. Martin, "Quantitative Drug Design", Marcel Dekker, New York, 1978.
(2) N. R. Draper and H. Smith, "Applied Regression Analysis", Wiley, New York, 1966.
(3) J. G. Topliss and R. J. Costello, *J. Med. Chem.*, **15**, 1066 (1972).
(4) International Business Machines Corporation, "System/360 Scientific Subroutine Package, Version III, Programmer's Manual, Program Number 360A-CM-03X", Manual GH20-0205-4, 5th ed, IBM Corp., White Plains, N.Y., Aug 1970.
(5) International Business Machines Corporation, "Random Number Generation and Testing", Manual G20-8011, IBM Corp., White Plains, N.Y.
(6) The incidence of chance correlations in the random-number simulated correlations is denoted by P_c. This is to be distinguished from a p value for a specific correlation which has the standard statistical connotation.
(7) F. Peradejordi, A., N. Martin, and A. Cammarata, *J. Pharm. Sci.*, **60**, 576 (1971).
(8) K. L. Gibbons, E. F. Koldenhoven, R. E. Nethery, R. E. Montgomery, and W. P. Purcell, *J. Agric. Food Chem.*, **24**, 203 (1976).
(9) B. M. W. M. Timmermans and P. A. van Zwieten, *J. Med. Chem.*, **20**, 1636 (1977).
(10) L. B. Kier and L. H. Hall, *J. Med. Chem.*, **20**, 1631 (1977).
(11) L. B. Kier and L. H. Hall, *J. Pharm. Sci.*, **67**, 1409 (1978).

A Preliminary Structure–Activity Study of the Mixed-Function Oxidase Inhibitor 7,8-Benzoflavone

Stephen Nesnow

Metabolic Effects Section, Genetic Toxicology Program, Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711. *Received April 9, 1979*

A series of substituted and structural analogues of 7,8-benzoflavone were examined for their ability to inhibit benzo[*a*]pyrene oxidation by the mixed-function oxidases found in hepatic microsomes prepared from 3-methylcholanthrene- and phenobarbital-induced rats. Of all the benzoflavones tested, only 6-amino-7,8-benzoflavone possessed significant inhibitory activity toward both classes of induced mixed-function oxidases. Parameters which were found to be necessary for maximal inhibitory activity were the maintenance of an unsubstituted or specifically substituted exocyclic phenyl group on position 2, the preservation of the pyran-4-one ring, and a 6 position which is either unsubstituted or substituted with an oxidizable moiety.

7,8-Benzoflavone (α-naphthoflavone, 7,8-BF, 1), a synthetic flavanoid, is an inhibitor of microsomal mixed-function oxidase activities.[1] These enzymes metabolize drugs, steroids, and xenobiotics, some of which are chemical carcinogens. 7,8-BF has been used for many years in mechanistic studies of the metabolic activation of carcinogens. 7,8-BF inhibits the metabolism, binding to DNA, and tumorigenesis in mouse skin of 7,12-dimethylbenz[*a*]anthracene[2] and inhibits the metabolism and carcinogenicity of 3-methylcholanthrene in mouse embryo cells[3] and mouse skin.[4] The metabolism of benzo[*a*]pyrene by hepatic microsomes isolated from rats induced with 3-methylcholanthrene,[5] 5,6-benzoflavone, and Aroclor 1254[6] is also inhibited by 7,8-BF. In contrast, 7,8-BF stimulates the metabolism of benzo[*a*]pyrene in hepatic microsomes isolated from rats induced with phenobarbital.[5] In mouse skin, 7,8-BF inhibits benzo[*a*]pyrene metabolism and binding to RNA and protein, has only a marginal inhibitory effect on benzo[*a*]pyrene binding to DNA, and has no effect on benzo[*a*]pyrene-mediated tumorigenesis.[2] Several 4′-substituted 7,8-benzoflavones have been reported to inhibit benzo[*a*]pyrene metabolism[7] but none were more potent than the parent compound.

This structure–activity study is a logical and stepwise approach toward an understanding of the mechanism of action of 7,8-BF and concentrates on the examination of three areas of the 7,8-BF molecule: the exocyclic phenyl group, the 6 position of 7,8-BF, and the pyran-4-one ring.

Results and Discussion

The inhibitory activities of the flavones and related compounds were evaluated with hepatic microsomes prepared from rats induced with 3-methylcholanthrene