Chance correlations in structure-activity studies using multiple regression analysis

John G. Topliss, and Robert J. Costello

J. Med. Chem., **1972**, 15 (10), 1066-1068 • DOI: 10.1021/jm00280a017

Notes

Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis

John G. Topliss* and Robert J. Costello

Departments of Medicinal Chemistry and Computer and Scientific Information, Research Division, Schering Corporation, Bloomfield, New Jersey 07003. *Received February 17, 1972*

In recent years, with the availability of high-speed computers, numerous publications have appeared[1-5] involving the use of the statistical technique of multiple regression analysis in searching for structure-activity correlations in series of biologically active compounds. In this technique a number of possible variables, comprising physicochemical parameters or molecular orbital parameters, are tested for possible correlation with a suitable measure of biological activity as the dependent variable.

In a typical example the correlation of activity with four possible independent variables with 16 observations may be checked by multiple regression analysis resulting in the finding that activity correlates highly with a combination of two of these variables. The statistical information which is usually provided[6] in reporting such a correlation consists of n, the number of observations, r, the multiple correlation coefficient, r^2, which is a measure of the explained variance, and s, the standard deviation. The statistical significance of the correlation equation and of each independent variable is also given in the form of a p value or as an F statistic from which a p value can be readily determined.

However, a point which appears to be generally overlooked is the influence of the total number of variables screened for possible correlation with activity on the statistical significance of the obtained correlation. Thus, in a hypothetical case in which six possible variables may be listed as possibly affecting activity it may be found on statistical examination using multiple regression analysis that two variables correlate significantly with activity. The standard statistical parameters relating to the selected equation take into account only the two variables contained in the equation and do not consider whether these two variables were selected from 6, 60, 600, or an infinite number of variables. Clearly, however, the greater the number of variables tested, the greater role chance will play in the observed correlation.

Experimental Section

In the present study this phenomenon† has been examined by setting up simulated correlations using random numbers. Numbers obtained from a random number generator represented a set of observations (dependent variable) and corresponding sets of variables (independent variables) to be examined for possible correlation with the observations. A fixed number of independent variables were chosen for one series of determinations. In this series several determinations were made using different numbers of observations. Each determination was the average result of 100 trials. The entire process was then repeated for different numbers of independent variables. Computations were performed on an IBM 360-85 computer using a stepwise multiple regression program (BMDO2R) with variables entered and removed at the 0.1 significance level.

†Chance correlations of this type have also been recognized by both H. U. Hostettler and C. Hansch, private communications.

Table I. Relationship, for 5 Variables, between Number of Observations, r^2, and Average Number of Variables Included

No. of observations	r^2	Average no. of variables included
5	0.90	1.90
10	0.65	1.25
15	0.58	1.20
20	0.47	1.10
30	0.42	1.18
50	0.31	1.18
80	0.26	1.20
120	0.28	1.23
250	0.13	1.07

Table II. Relationship, for 10 Variables, between Number of Observations, r^2, and Average Number of Variables Included

No. of observations	r^2	Average no. of variables included
5	1.00	2.70
10	0.74	2.00
15	0.70	1.90
20	0.52	1.92
30	0.42	1.60
60	0.41	1.77
80	0.32	1.45
120	0.26	1.50

Table III. Relationship, for 20 Variables, between Number of Observations, r^2, and Average Number of Variables Included

No. of observations	r^2	Average no. of variables included
5	1.00	
10	0.98	5.50
15	0.81	4.50
30	0.55	2.80
60	0.40	2.60
120	0.30	2.30
270	0.20	2.20

Table IV. Relationship, for 30 Variables, between Number of Observations, r^2, and Average Number of Variables Included

No. of observations	r^2	Average no. of variables included
20	0.80	6.30
30	0.68	4.80
60	0.48	3.70
120	0.32	3.60
240	0.23	3.00

Results

The results have been tabulated in Tables I–IV and expressed graphically in Figures 1 and 2. In Figure 1 average r^2 values from 100 trials (ordinate) were plotted against the number of observations (abscissa) for different fixed numbers of variables tested. It can be seen that impressive chance correlations were obtained when the number of observations was small compared to the number of variables tested. The variables included were significant at the minimum level of 0.1, as required by the program, and usually at much higher levels. With an increasing number of observations the degree of chance correlation was steadily reduced.

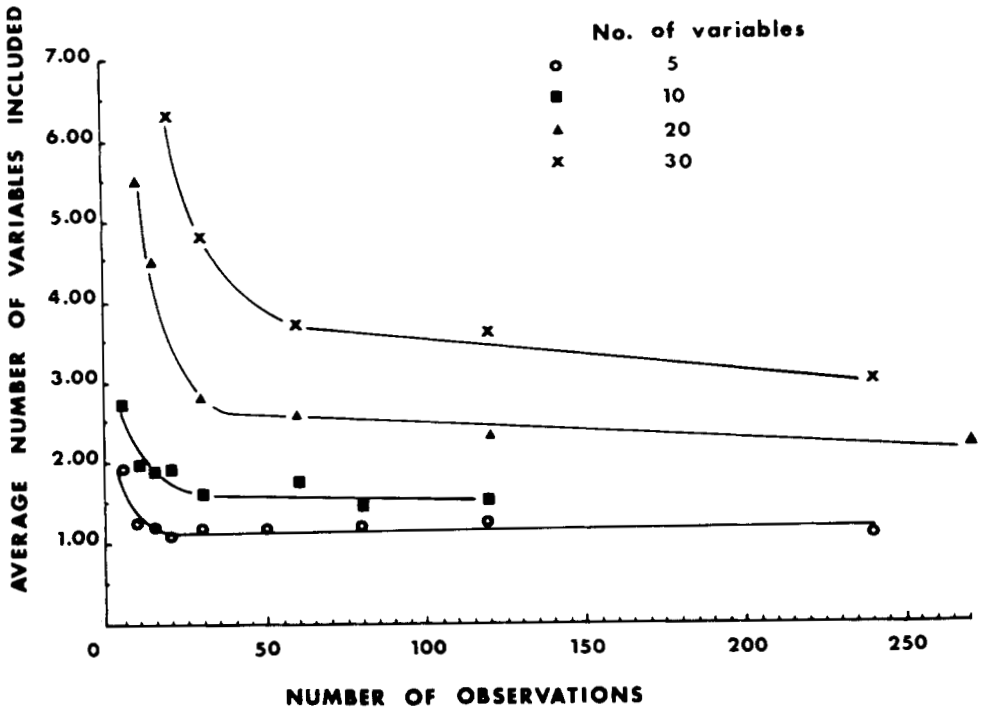

Figure 1. The relationship between r^2 and the number of observations.

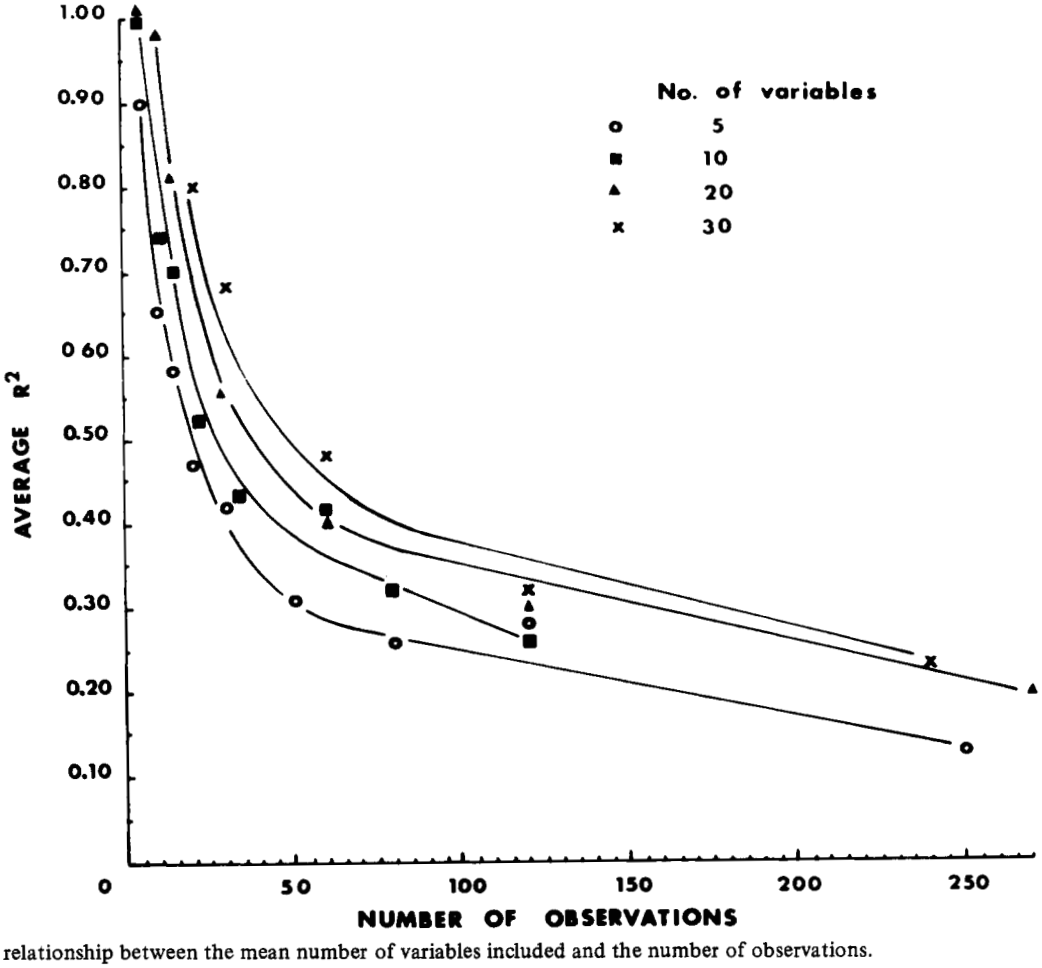

Figure 2. The relationship between the mean number of variables included and the number of observations.

The relationship between the mean number of variables included and the number of observations for different fixed numbers of variables tested is illustrated in Figure 2. The results show that the number of variables correlating by chance increases as the number of observations decreases. Also, for a set number of observations the number of variables included increases as the number of variables tested increases.

These studies reveal a potential problem of some magnitude in structure-activity correlations where many possible variables must be considered. This is particularly so in MO type correlations where many parameters can be calculated for a compound and frequently there is no valid reason to choose one over another. In these cases in order to adequately reduce the risk of chance correlations a large number of observations must be employed and these are not always available. Situations could easily arise in which the number of possible variables could not be adequately supported by the number of observations. Correlations obtained under these conditions would have greatly reduced significance. Some recently reported[7-9] correlations using MO parameters need to be reexamined in this context.

In Hansch type correlations the situation is less difficult since only a limited number of variables need be considered, representing possible hydrophobic, electronic, and steric effects. However, misleading correlations can still arise with an insufficient number of observations.

In Free and Wilson type correlations the phenomenon under discussion does not arise since each substituent is treated as a significant variable and therefore variables are not tested for possible inclusion using a multiple regression procedure.

The data presented allow an assessment to be made of the probable degree of chance correlation, when observations are examined for correlation with varying numbers of independent variables, as a function of the number of observations and the number of variables. Thus, for a given number of variables to be tested, the required number of observations to avoid undue risk of chance correlations can be estimated. For example, if $r^2 = 0.40$ is regarded as the maximum acceptable level of chance correlation then the minimum number of observations required to test five variables is about 30, for 10 variables 50 observations, for 20 variables 65 observations, and for 30 variables 85 observations.

Acknowledgment. The authors are indebted to Mr. M. Miller, Department of Computer and Scientific Information, Schering Corp., for the statistical studies.

References

(1) C. Hansch, *Annu. Rep. Med. Chem.*, **1966,** Chapter 34 (1967).
(2) C. Hansch, *ibid.*, **1967,** Chapter 33 (1968).
(3) W. P. Purcell and J. M. Clayton, *ibid.*, **1968,** Chapter 28 (1969).
(4) J. M. Clayton, O. E. Miller, Jr., and W. P. Purcell, *ibid.*, **1969,** Chapter 26 (1970).
(5) A. Cammarata, *ibid.*, **1970,** Chapter 24 (1971).
(6) P. N. Craig, C. Hansch, J. W. McFarland, Y. C. Martin, W. P. Purcell, and R. Zahradník, *J. Med. Chem.*, **14,** 447 (1971).
(7) A. Cammarata and R. L. Stein, *J. Med. Chem.*, **11,** 829 (1968).
(8) A. J. Wohl, *Mol. Pharmacol.*, **6,** 195 (1970).
(9) F. Peradejordi, A. N. Martin, and A. Cammarata, *J. Pharm. Sci.*, **60,** 576 (1971).

Structure-Activity Correlation for Substrates of Phenylethanolamine *N*-Methyltransferase (PNMT)

Ray W. Fuller* and Max M. Marsh

Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46206. *Received March 20, 1972*

Fujita and Ban[1] have reported a mathematical correlation of structure-activity relationships among PNMT substrates, using literature data.[2,3] The data they selected were measurements of substrate activity at a single substrate concentration; the concentration was very high, one at which inhibition by excess substrate occurs to different degrees among the various substrates.[4,5] Thus, the group contributions calculated by Fujita and Ban[1] probably relate mainly to *inhibitory* influences rather than to interactions favoring substrate activity. For instance, they showed a negative contribution of the 4-hydroxyl group. Such a negative contribution contrasts with the effect of the 4-hydroxyl group in the α-methylphenethylamine series that we recently reported as PNMT inhibitors.[6] There the 4-hydroxyl conferred an even greater affinity for PNMT than we were able to account for by σ and π values. We report here that correlation of structure with affinity of phenylethanolamines as PNMT substrates rather than with activity at a single excessively high concentration leads to conclusions different from those of Fujita and Ban.

We have calculated $-\log K_m$ values for phenylethanolamines as substrates for PNMT from rabbit adrenal.[5] The K_m values were calculated by the method of Wilkinson[7] from measurements of reaction velocity at 4–7 different substrate concentrations. All enzyme assays were done by a previously described method[8] in which the transfer of the labeled methyl group was measured after precipitation of *S*-[*methyl*-^{14}C]adenosylmethionine with Reinecke salt. The statistical evaluations were made by single and multiple linear regression analysis using Lilly Program S21, a modified and updated version of an original program submitted to the IBM 1620 user's group.

From the observed results with the six phenylethanolamines listed in Table I, with meta or para substituents on the ring, we derived an equation (eq 1) that fit the data at a level of significance $P = 0.01$. The square of the correlation coefficient was 0.909. The correlation was not improved

$$-\log K_m = 1.240\pi + 4.339 \quad (1)$$
$$(\pm 0.243) \quad (\pm 0.196)(\pm 0.152)$$

by adding a π^2 term. Standard errors of the terms are included in parentheses. As shown in the table, $-\log K_m$ values calculated from this equation agreed well with those derived from experimental observations. Hammett σ values for the substituents were not useful in the correlation. That these should not greatly influence the methylation reaction could be surmised from the fact that the pK_a's are not greatly influenced by aromatic substitution.

Two ortho-substituted phenylethanolamines, not included in the derivation of the equation, were estimated reasonably well by it. The calculated and observed $-\log K_m$ values were 5.06 and 4.80, respectively, for *o*-chlorophenylethanolamine and 4.39 and 4.10, respectively, for *o*-fluorophenylethanolamine.

Three hydroxy derivatives were, on the other hand, not

Table I. Observed and Calculated Activity of PNMT Substrates

Substrate	π^b	$-\log K_m$ value[a] Obsd	Calcd	Difference
Phenylethanolamine	0.0	4.10	4.34	+0.24
3-Fluorophenylethanolamine	0.19	4.60	4.57	−0.03
4-Fluorophenylethanolamine	0.14	4.80	4.51	−0.29
3-Bromophenylethanolamine	0.91	5.20	5.47	+0.27
4-Bromophenylethanolamine	0.90	5.60	5.46	−0.14
3,4-Dichlorophenylethanolamine	1.38	6.10	6.05	−0.05

[a] Units of K_m values were in molar substrate concentration. [b] π values from Fujita, *et al.*[10]