INVESTIGATION OF THE EFFECT OF DATA ERROR IN THE ANALYSIS OF BIOLOGICAL TRACER DATA

J. MYHILL

From the Department of Surgery, University of Sydney and the Institute of Medical Research, Royal North Shore Hospital of Sydney, Sydney, Australia

ABSTRACT Data obtained from tracer studies often consist of serial measurements after administration of radioisotope. Very little work has been published on how the error in the data affects the mathematical analysis. Computer simulation was here employed to produce data with error of different magnitude and form for each of several values of rate constant and amplitude. The data were terminated when the value of the last point was 5% of the value of the first point, and also in other ways arranged to simulate experimental situations. The sets of simulated data for a two compartment system were analyzed by the gaussian iterative technique. With a rate constant ratio of at least four the technique converged for data errors of 5% or less. The calculated error in the rate constants ranged from 2 to 85%, and in the amplitudes from 1 to 50%, for data error of 0.5 to 10%. The lesser rate constant and amplitude had the greater errors. If a wrong assumption was made in the analysis about the variation of data error over the time interval of measurement, then the calculated values of parameter standard deviations were greatly in error. The results can be used to decide what experimental accuracy is needed for a given accuracy of model parameters for a variety of biological problems.

INTRODUCTION

Some aspects of biological function can be studied by the use of radioactive tracers. Typically the results of such a study may be a series of measurements of radioactivity extending over an interval of time after administration of the radioisotope. It is often both valid and of value to represent the aspects of a biological system which are under investigation by a simplified compartmental model and to fit the solution function of the model to the experimental data points in order to estimate the model parameters. If the system is in a steady state then it has been previously shown that the behavior of labeled substances can be represented by linear differential equations (Sheppard and Householder, 1951; Berman and Schoenfeld, 1956) and that the solution for the activity vs. time curve in each compartment consists of a sum of exponentials:

$$f(t) = \sum_{i=1}^{n} N_i e^{-\lambda_i t}.$$

How the errors and omissions in the data affect the estimation of the rate constants, λ_i, and the amplitudes, N_i, and through them the estimation of the transfer rates and compartment sizes, has generally not been a subject of study.

The collection of data is usually terminated for some experimental reason: the radioactivity may become too low to count accurately; the animal may die or the patient become unavailable for study; or the radioactive label may become detached from its compound. This data truncation, together with the experimental error in each measured data point and the fewness of points, creates uncertainty in the values of the model parameters. This is the case even where a valid, unequivocal model of the system is assured. Preliminary results have been previously reported (Myhill, 1965, 1966).

METHOD OF ERROR ANALYSIS

(i) *Computer Simulation of Data*

Data points were simulated on a high speed digital computer (English Electric KDF9) using the equation for $f(t)$ and some random number generators. At first uniform random variates were generated using a multiplicative congruential method (Behrenz, 1962) and random normal variates approximated by summing uniform variates. These methods were soon replaced by another multiplicative congruential method for uniform variates and Box and Muller's method for normal variates (Pike and Hill, 1965). The second pair of methods was faster on the computer. In addition, if the first method for uniform variates was used as input to the first method for normal variates, then normally distributed random numbers with a falsely large standard deviation were obtained. This was found to be caused by a correlation ($r \approx 0.2$) between successive values of the input uniform variates.

Since no other studies of this type have been reported it was decided to start with a simulation of data from a sum of two exponentials. It was important to formulate the problem so that the results would be independent of the particular values of N_i and λ_i used in the simulation.

It can be shown that if the standard deviation of the data error is a constant percentage of the value of $f(t)$ at each point, then the problem is unchanged by a multiplication of the N_i by a constant factor. That is, the absolute values of the N_i are not important, and only their ratio will influence the results. It can also be shown that if the λ_i are multiplied by a constant factor, and t is divided by the same factor, then the problem is unchanged. In these studies the maximum time ($t_{\max}$) was chosen for each $f(t)$ to make $f(t_{\max})$ a constant proportion of $f(0)$, and the value of $t_{\max}$ uniquely determined the set of data time values. Larger λ_i gave rise to a proportionately smaller $t_{\max}$. Thus the absolute values of the λ_i are not important, and only their ratio will influence the results.

The situations where both N_1/N_2 and λ_1/λ_2 equal 10, 6, 4 and 2 were studied, the cases where the rate constants λ_i are closer to equality being the more difficult. If λ_1/λ_2 is less than about four some methods of analysis become less effective (Brownell and Callahan, 1963) or are subject to large bias (Myhill, Wadsworth, and Brownell, 1965). The simulated data were terminated when the value of the last data point was 5% of the value of the initial point. This corresponds to a situation often observed in reports of experiments and is a more difficult case to analyze than if the data were further extended in time (i.e., so that the value of the last data point was less). In an endeavor to approximate to types of data seen in the literature, the cases of 31 and 11 data points equally spaced over the time interval of data collection were investigated; random error with a constant per cent standard deviation was simulated at each point; the statistical distribution of the error was normal and (in other studies) rectangular, to investigate any effects of nonnormal error; and the magnitude of the standard deviation of error in any one set of data was varied from 0.5 to 10% between sets.

(ii) *Numerical Analysis of Simulated Data*

The function, $f(t)$, was fitted to each set of data using a valid least-squares gaussian iterative technique, and the estimated values of the λ_i, N_i, and their estimated standard deviations were obtained. Each set of data was simulated 50 times (with everything the same except the random numbers employed to simulate the error) and analyzed again. By this means the variability and bias of the estimates of the parameters and their standard deviations were studied.

TABLE I

MAXIMUM AMOUNT OF ERROR IN THE DATA THAT ALLOWS CONVERGENCE*

Ratio λ_1/λ_2	10		6		4		2	
Statistical distribution of error in data	Normal	Rectangular	Normal	Rectangular	Normal	Rectangular	Normal	Rectangular
	%	%	%	%	%	%	%	%
31 data points	10	10	10	10	8	10	1.5	2
11 data points	10	10	10	10	5	5	1	1

* Error greater than 10% was not studied.

RESULTS

(i) *Convergence of the Iterations*

The maximum amount of error in the data that permitted convergence of the numerical technique is shown in Table I. In each case the true parameter values (used in generating the relevant data) were used as starting values for the iteration. From the point of view of starting values, these results thus represent the most favorable situation possible to achieve.

No special techniques were used to help the convergence. Each set of data was simulated 50 times, and a failure to converge for at least one set was interpreted as meaning that convergence was not obtained for those data. Thus the results are somewhat less favorable than might be obtained in practice.

Failure of convergence was judged to have occurred if, in the process of calculation, a number exceeded the machine limit ($\sim 10^{38}$) (this was almost always the case) or, rarely, if the change in parameter values was not less than 10^{-4} and the change in variance not less than 10^{-6} after 25 iterations. The parameter values were in the range 0.01–1.00.

(ii) *Normally and Rectangularly Distributed Error*

For a given value of standard deviation in the data error, the distribution of error did not influence the standard deviations of the λ_i or N_i.

(iii) *Estimates of Parameter Standard Deviations*

Each fit of the function, $f(t)$, to a set of data points was effected using the gaussian iterative process of successive approximations. A straightforward least-squares fit was not possible since $f(t)$ is not linear in the parameters. Using the variance/covariance matrix obtained at the last iteration it was possible to obtain an estimate of the standard deviation (s_p) of each parameter, λ_i or N_i.

The same data was then simulated 50 times, with everything the same except the random numbers used to assign the errors. From the 50 sets of results the mean parameter values and the mean standard deviations (means of the s_p) were calcuated.

Another estimate of the s_p was obtained by direct calculation from the 50 values obtained for each parameter. This estimate is subject to a moderate sampling error, but to run more than 50 simulations of the same data would have required prohibitively excessive computer time.

In a linear problem the two estimates of standard deviation would agree within their statistical errors, but in the present nonlinear problem solved by successive approximation it was considered possible that they could differ. That is, the standard deviation (s_p) computed from the variance/covariance matrix after a fit to one simulated set of data could be biased.

Since the two estimates of standard deviation are not independent, but related in some complicated way, no exact statistical test of their difference (or ratio) was available on the basis of which precise probability statements could be readily made. In the two paragraphs below, if the two estimates differed by about four times the standard error of the mean s_p then a bias was considered to exist. The reason why more results appeared to be biased from rectangularly distributed error than from normally distributed error may possibly lie in the inadequacy of this criterion.

Table II shows at what level of data error (rectangular distribution) the bias is apparent in the computed standard deviation of λ_1, and also by how much the true standard deviation is underestimated. For normally distributed error bias was only apparent with $\lambda_1/\lambda_2 = 6$, with 11 points and with 6% or more data error, in which case the standard deviation was underestimated by 22%.

The standard deviation of λ_2 was also underestimated by 28% in the same situation. The only other occasion on which λ_2 was underestimated (by 21%) occurred when $\lambda_1/\lambda_2 = 2$, 11 points, and 1% or more data error.

TABLE II

VALUE OF ERROR IN DATA FOR WHICH BIAS IS APPARENT IN THE STANDARD DEVIATION OF λ_1 CALCULATED FROM ONE SET OF DATA (RECTANGULAR DISTRIBUTION OF ERROR) AND THE PERCENTAGE UNDERESTIMATE OF THE STANDARD DEVIATION

Ratio λ_1/λ_2	6	4	2
	%	%	%
31 points	10 (18 low)	10 (18 low)	2 (18 low)
11 points	5, 6, *not* 10 (25 low)	5 (7 low)	1 (26 low)

(iv) *Error in the λ_i as a Function of Error in Data and Number of Data Points*

Table III shows the average error in λ_1 and λ_2 as a function of magnitude of error in the data, and number of data points, for normally distributed error. Two results are obvious: 31 points give a lower error than 11 points for the same data error; and the lesser rate constant, λ_2, is affected more than the greater, λ_1, even though the

TABLE III

AVERAGE STANDARD DEVIATIONS AS PERCENT OF TRUE λ_i

Rate constant		λ_1		λ_2	
Number of data points		11	31	11	31
Ratio λ_1/λ_2	Error in data				
	%	%	%	%	%
2	0.5	4.7	3.5	6.2	4.8
	1	8.7	6.7	14.2	10.4
	1.5	—	10.4	—	15.5
4	2	4.8	3.6	9.7	7.4
	5	11.6	8.8	27.5	19.9
	8	—	14.8	—	32.0
6	2	3.2	2.4	9.4	7.1
	5	7.9	5.9	25.3	18.3
	10	16.5	13.1	49.6	37.7
10	2	2.5	1.9	16.8	12.6
	5	6.3	4.7	44.2	32.3
	10	13.2	10.3	85.6	66.5

TABLE IV

AVERAGE STANDARD DEVIATIONS AS PERCENT OF TRUE N_i

Amplitude		N_1		N_2	
Number of data points		11	31	11	31
Ratio N_1/N_2	Error in data				
	%	%	%	%	%
2	0.5	9.5	7.2	19.6	15.0
	1	17.4	13.7	35.9	28.4
	1.5	—	20.1	—	41.8
4	2	3.4	2.3	14.7	11.0
	5	8.0	5.7	34.5	26.7
	8	—	9.1	—	41.6
6	2	1.9	1.3	9.8	7.2
	5	4.7	3.3	24.0	18.0
	10	9.4	7.1	43.9	36.8
10	2	1.5	1.1	10.3	7.5
	5	3.8	2.8	24.9	18.6
	10	7.9	6.0	48.4	39.7

TABLE V

PER CENT CHANGE IN PARAMETER SD'S CALCULATED ON ASSUMPTION OF CONSTANT SD IN DATA OVER THE CORRECT ONES CALCULATED ON BASIS OF CONSTANT PERCENT SD IN DATA (NORMALLY DISTRIBUTED ERROR)

Parameter		λ_1		λ_2		N_1		N_2	
Number of data points		11	31	11	31	11	31	11	31
		%	%	%	%	%	%	%	%
Ratio λ_1/λ_2	Error in data								
2	0.5%	33	*	64	*	33	*	29	*
4	2%	0	24	72	128	51	105	41	83
6	2%	−11	15	75	135	28	67	49	97
	5%	*	23	*	140	*	61	*	88
10	2%	−12	16	90	147	6	40	61	110
	5%	*	24	*	151	*	32	*	111

* No comparison possible since convergence not obtained.

per cent standard deviation in the data was constant, i.e., the tail of the data was no more noisy than the initial portions.

(v) *Error in the N_i as a Function of Error in Data and Number oj Data Points*

Table IV shows the average error in N_1 and N_2 as a function of magnitude of error in the data, and number of data points for normally distributed error.

(vi) *Effect of Wrong Error Assumption on Analysis*

In practice it is sometimes not known whether the error in some biological data has a constant per cent standard deviation (*SD*), or simply a constant *SD*, over the time interval of data collection. The same simulated data (with constant percent *SD*) was therefore analyzed as though it had a constant *SD* and the effects observed. Firstly a bias was sometimes, though not often, observed in the average values of λ_i estimated. Secondly the computed standard deviations were biased as shown in Table V.

DISCUSSION

The tracer activity curve in any steady-state two-compartment open system can be described by a sum of two exponential terms. The relationships between the transfer rates and compartment sizes, on the one hand, and the rate constants (λ_i) and amplitudes (N_i), on the other hand, are expressible by simple algebraic equations. The form of the equations depends on the configuration of the system, and these equations can be used to transform information on the errors obtained in λ_i and N_i to information about the associated errors in the transfer rates and compartment sizes. By this means this error information can be applied to a range of biological problems, mainly tracer studies in biochemistry and physiology.

Within the conditions here investigated (including $\lambda_1/\lambda_2 = N_1/N_2$, and data points equally spaced) convergence is readily obtained in the gaussian iterative method if the ratio of λ_1 to λ_2 is not less than about four. For lower ratios (i.e. rate constants more nearly equal) a very good experiment is necessary, that is, the standard deviation in each data point should be 1%. If more points are measured then the error may permissibly rise to 1½%, which still represents in most situations a very good experiment. These are the most favorable situations possible; in practice since the best starting values for the iteration are of course unknown, convergence may be difficult to obtain even when these requirements are satisfied.

It would be expected from least squares theory (Kendall and Stuart, 1961) that the calculated values of the standard deviations of the model parameters in a linear model would not be affected by the form of the statistical distribution of the error in the data, but only by its standard deviation. The same conclusion was shown to be valid for models of the form of $f(t)$ when solved by successive approximation, if normal and rectangular error is studied and for magnitudes of error within the range of the present study.

The standard deviations calculated from one set of data were generally unbiased, as would be anticipated. However for larger errors, fewer data points, and lower rations of λ_1 to λ_2 there was a suggestion of some bias. The problem can thus be circumvented by increasing the accuracy of each measurement and the number of points measured. If this is not possible, then the percentage underestimates stated under Results (iii) could possibly be used to correct the calculated standard deviations if the magnitude of the per cent standard deviation in the data error was known or could be determined independently. This correction would not be of high accuracy and any further probability statements on the basis of the corrected parameter standard deviations would need to be made with caution.

For any desired error in λ_i or N_i, the allowable error in the data may be found from Tables III and IV. As the data error increases, the parameter error increases proportionately. For the same error in each point, taking more points results in a smaller parameter error. The errors in λ_2 and N_2 are greater than those in λ_1 and N_1, even though the error in the data had a constant per cent standard deviation. This would seem to be due to the truncation of the data at the point where its value was 5% of the initial point. The parameter errors could be reduced by collecting data extending further in time, as well as by lowering the error in each point or taking more points over the same time interval. The object of the present study, however, was to examine the situation where for experimental reasons the data could not be extended further in time.

If the data are believed to contain error with a constant standard deviation and analyzed as such, whereas really the error has a constant per cent standard deviation, then serious errors in the calculated parameter standard deviations result (Table V). They could not be validly used in any further probability statements. It is therefore essential to find out the variation in standard deviation over the full range of collected data. In practice the gaussian iterative method can be easily carried out with a weighting of each point that represents the actual variation of standard deviation observed in the data.

The results in this paper apply to any two-compartment system from which data have been collected until $f(t) = qf(0)$, $q = 0.05$, in one of the compartments, and for which $N_1/N_2 \approx \lambda_1/\lambda_2$. The results do not depend on the absolute magnitudes of the N_i of λ_i. In practice if $q > 0.05$ then the analysis is more difficult, the errors in N_i and λ_i will be greater, and the situation is worse in other respects. If $q < 0.05$, i.e. the data is sampled for a longer time, the results will be better. If more data points are taken the results will also be better, etc. The figures quoted in this paper can thus form a guideline in evaluating other situations. If three compartments are being studied then the results presented here would be the best that one could hope to achieve for the errors in four of the six parameters involved.

This work was supported by a grant from the New South Wales State Cancer Council.

Received for publication 17 May 1967.

REFERENCES

Behrenz, P. G. 1962. *Comm. ACM* **5**:553.

Berman, M., and R. Schoenfeld. 1956. *J. Appl. Phys.* **27**:1361.

Brownell, G. L., and A. B. Callahan. 1963. *Ann. N. Y. Acad. Sci.* **108**:172.

Kendall, M. G., and A. Stuart. 1961. The Advanced Theory of Statistics. Hafner Publishing Co., Inc., New York. **2**:83.

Myhill, J. 1965. Conference on Computer Applications in Medicine. Melbourne, Australia. October.

Myhill, J. 1966. Proceedings of 6th Annual Meeting of Physics in Medicine and Biology. Melbourne, Australia. August.

Myhill, J., G. P. Wadsworth, and G. L. Brownell. 1965. *Biophys. J.* **5**:89.

Pike, M. C., and I. D. Hill. 1965. *Comm. ACM* **8**:605, 606.

Sheppard, C. W., and A. S. Householder. 1951. *J. Appl. Phys.* **22**:510.