Journal of Medicinal and Pharmaceutical Chemistry

VOL. 2, No. 5 (1960)

Review

Sequential Analysis, Screening and Serendipity*

JOHN T. LITCHFIELD, JR., M.D., *Experimental Therapeutics Research Section, Lederle Division, American Cyanamid Company, Pearl River, New York*

In 1754 Walpole related the fairy tale of the Three Princes of Serendip who in their travels were constantly making, by *accident* or *sagacity*, delightful discoveries of things for which they were not searching. From this the term serendipity was coined. An outstanding example of serendipity is the voyage of Columbus in which he sailed west from Spain in order to reach Japan and the Indies and, instead, discovered a new continent.

It is unlikely that there is anyone engaged in medicinal chemical research who does not have a burning desire to be a discoverer. Each new compound conceived and synthesized probably carries the hope of realizing this desire, a hope which is generally extinguished promptly when the compound is put to test. At rare intervals the desired event occurs; the logically conceived compound is found to be active and all of the other logically conceived compounds which were not active are forgotten. This latter group of miscarriages is often put on the shelf, yet it is known from past experience that one or more of these compounds may possess a totally unexpected type of activity. An example of serendipity is provided by a compound which was synthesized with the idea that it might be a superior antioxidant for rubber but which was a failure for this purpose; it was unexpectedly effective against tuberculosis in mice.

This coincidence of a suitable test with the right compound occurs so infrequently that it would be rewarding if the chances for such happy accidents could be increased.

* Presented at Gordon Research Conference on Medicinal Chemistry, August 7, 1959.

Are there ways in which these chances can be favourably altered? The answer to this cannot be certain; but some logical things can be done. These all represent efforts to increase the chances that a suitable test will be applied to the right compound in the shortest possible time.

The term activity as used in this presentation is defined as follows.* From the practical view, new drugs may have desirable actions such as the lowering of blood pressure in hypertension or the producing of acute pulmonary edema for killing rats, or they may have undesirable actions such as an emetic effect in the case of a drug intended for treating hypertension or a repellent effect in the case of a drug intended for use as a rat poison. Therefore, to have a common term for these different properties, the word 'activity' will be used in the general sense and, for the most part, without reference to its desirability or undesirability. Even though degree of activity may range from very low to very high, it will be indicated in this paper only as the dichotomy: 'interesting' *vs.* 'uninteresting activity'. The term 'inactivity', meaning the total absence of 'activity', will scarcely be mentioned because the absence of 'activity' can never be proven. For example, if activity of a compound is not observed in a particular test, the possibility must always be admitted that it might be observed in a modified or repeated test.

The majority of interesting classes of drugs have actually been found by some kind of test unrelated to the purpose for which the compound was synthesized by a chemist. Even some of the exceptions are more apparent than real because the original concept for these particular compounds traces back to a fortuitous observation. An example of this is the carbonic anhydrase inhibitor, acetazolamide, which was developed in a logical study relating structure, pK_a and ability to inhibit the enzyme carbonic anhydrase thereby causing excretion of base and diuresis. However, the genealogy of this compound as shown in Table I can be traced back to quite fortuitous observations[1–9] that administration of sulphanilamide caused the urine to become alkaline and that

* A portion of this paper was published in *Evaluation of Drug Toxicity*, edited by A. L. Walpole and A. Spinks; J. and A. Churchill, Ltd., London, 1958. Permission of the copyright owners, Imperial Chemical Industries, Ltd., to reproduce this material is gratefully acknowledged.

blocking the free sulphonamide group eliminated carbonic anhydrase inhibitory activity. It is worth noting that a weak effect on the kidneys which was regarded as undesirable in the case of sulphanilamide is regarded as desirable when present to a higher degree in acetazolamide.

Table I. Genealogy of acetazolamide

1932	Discovery of carbonic anhydrase.[1]
1937	Sulphanilamide produces acidosis, loss of fixed base and rise in urinary pH.[2-4]
1940	Substituents on $-SO_2NH_2$ nitrogen block carbonic anhydrase inhibitory activity.[5]
1941	Carbonic anhydrase found in kidney.[6]
1942	Sulphanilamide raised pH of frog urine by carbonic anhydrase inhibition.[7]
1950	Synthesis of acetazolamide and other heterocyclic sulphonamides.[8, 9] Their principal action: to promote diuresis and loss of base.

Knowledge concerning the relation between chemical structure and pharmacological activity of drugs is very limited. So is knowledge of the mode of action of most drugs. Therefore, the search for new and better drugs has many illogical features which all appear in the procedure called blind testing or screening. This, however, does not preclude that the procedures as such should be as systematic as possible, a fact which has been recognized in recent years. Some basic principles which are involved in blind but systematic screening will be considered at this point.

The difference between screening and evaluation may be examined in terms of the questions and answers underlying these two kinds of tests as shown in Table II.

Table II. Difference between screening and evaluation

Screening	
Qualitative Question:	Answer:
Is the drug active as tested?	Yes, no, or not certain.
Evaluation	
Quantitative Question:	Answer:
How much drug does it take to produce a given effect?	Dose, in units of weight.

Screening deals with what may be termed a qualitative question and answer. Consider the difference between qualitative and quantitative chemical analysis. The former is based on the dichotomy—present or absent, while the latter is based on measurement of the amount present. In the type of simplified test for drug activity under consideration, the analogous question is asked: Is interesting activity present or absent? This is entirely different from the question: How much drug will be required to produce a given effect? or, in other words, how active is the drug?

Next compare the modest effort required to show by qualitative analysis whether one particular substance is present or not with

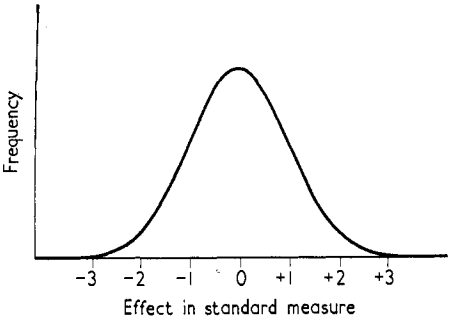

Fig. 1. Activity of isotonic saline (or any inert material) in a hypothetical test. This is a normal curve relating frequency of observing values which deviate from the mean value for saline. Standard measure expresses these deviations as multiples of the standard deviation of the mean effect of saline which is zero.

the relatively large effort required to isolate and determine the actual quantity present. The analogy carries over quite precisely to drug screening where it may be entirely feasible to decide whether interesting activity is present or not with a test on only one or two animals, while it may take 50 to 100 animals to relate dosage to activity, and as many as 400 animals to compare the activities of two drugs quantitatively.

In a blind screening programme, large numbers of essentially unselected compounds are examined to find out if any possess a particular action. This is based on the assumption that amongst available compounds, in general, there must be some which have

interesting activity although the frequency with which these occur in the general population of compounds is quite unknown.

It is now evident that for operational purposes there must be a definition of what constitutes interesting activity. As a beginning, interesting activity in a test might be defined as a degree of activity more intense than that exhibited by isotonic saline. Fig. 1 portrays the activity of isotonic saline in a hypothetical procedure and it is evident that on one occasion the activity may be much greater than on another. It can be deduced from this that many compounds whose activity does not exceed that of isotonic saline will belong to this same frequency distribution. Of course, the compounds of interest are those whose distribution is offset to one side or the other. Usually, but not always, effects in only one direction are of interest. The problem in every case is how to decide that the observed activity of a particular compound is significantly different from that of isotonic saline.

It is well known that the more nearly equal two degrees of action, the greater the number of animals or tests needed to show a difference.[10] It follows, therefore, that if the degree of activity to be defined as interesting is rather similar to that of saline, a large effort will be required to show that an observed effect from a drug is significantly different from that of the saline control. Suppose the problem is approached in terms of showing, at least 97 per cent of the time, that saline is of no interest. This would define interesting activity as an effect which exceeds that of saline by almost two standard deviations. At the same time, uninteresting activity might be equated to the mean effect of saline.

Fig. 2 may help to clarify this point. This illustrates the effect of two weeks' treatment with normal food on the survival time of mice infected with tuberculosis. A few mice began to die on the second day and about half of the mice were dead by the fourteenth day. However, some mice lived 24 days or longer. If this infection was used for screening compounds, those which reproduced the therapeutic benefits of normal food, namely survival on the average for no more than 14 days, would be of no interest. There would be interest, however, in a compound which produced a therapeutic effect of a degree unlikely to occur with normal food alone, i.e. survival on the average for 20 days or more.

Thus far it is established that screening is based on a qualitative test and that the minimum degree of effect which must be produced by a compound to be of interest can be exactly specified. In addition, effects not better than those produced on the average by saline, food, or the vehicle used for drug administration have been declared to be of no interest.

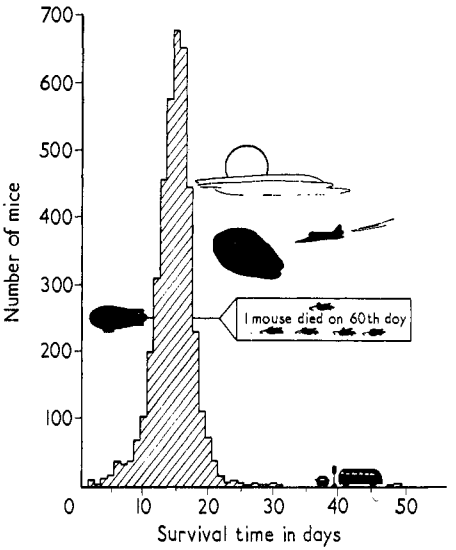

Fig. 2. Frequency distribution of the survival times after infection of 4,094 mice inoculated intravenously with *Mycobacterium tuberculosis* and fed normal diet containing no therapeutic agent. Period represented: June, 1948 to June, 1951.

Therefore, a means is available for deciding that certain compounds are of interest and others not. Actually, even more than this is available because the odds for erroneous decisions under these specified criteria can be calculated. For example, there is the possibility that a compound will be called interesting which is not really any more active than saline. In addition, there is the probability of a more important kind of error; namely that a compound will be called uninteresting when in reality it is more active than saline and should have been called interesting. No matter how the test is conducted, these two kinds of risks will be

present and their magnitude will be determined by the variability of the animals' response and the number of animals used.[11]

If an effect is observed which is between the two criteria, the test has to be repeated because there is not enough information to decide whether or not to be interested. The information from both the first and second stages of testing would then be combined in order to reach a decision, if possible. If it is possible to specify: (1) the level of activity which is of definite interest as well as the level which has no interest, (2) the number of animals to be used in the testing, and if (3) the basic variability of the response of the experimental animals is known, then the risks of either kind of incorrect decision have been fixed and can be calculated. The method of calculating these is a part of sequential analysis as developed by Wald.[11] The method of conducting such a test in several stages is commonly known as the sequential method and will be discussed more fully later.

While it is evident that in using this exact approach quantitative considerations are concerned, it should be noted particularly that the size of the dose of a compound has not even been mentioned. In actual practice, some decision must, of course, be made regarding the dose to be used in screening compounds. Frequently, the dose–effect relationship of already known active drugs enters into this latter decision. As an example, Fig. 3 shows the dose–effect curve of a compound, chlorpromazine, which possesses a selective action on the central nervous system so that caffeine-induced motor activity in the mouse is suppressed. The ordinate represents the number of movements of a jiggle cage during a one-hour period. Each point represents the mean of six counts each on a cage containing five mice. Mice treated with isotonic saline gave rise to a mean count of 12,400 while the drug used in this experiment gave decreasing counts with increasing dosage. If this form of selective action is to be considered for screening, the degree of minimum effect which is interesting might be defined in the following way: (*a*) if a compound under test lowers the count below 6,400 it is of interest, and (*b*) if a compound under test does not lower the count to at least 10,400 it is of no interest. The fact that chlorpromazine has already been discovered and exhibits the effect shown here might lead to a decision to screen compounds at a dose of 16 mg which is twice that needed

with chlorpromazine to elicit interesting activity. In another case the decision might be made to test all compounds at 1 g/kg and to re-test at a lower dose if toxicity is observed. In this case it might be found that an unduly high proportion of samples required re-testing because of lethal effects and that a lower dose level would be more practical.

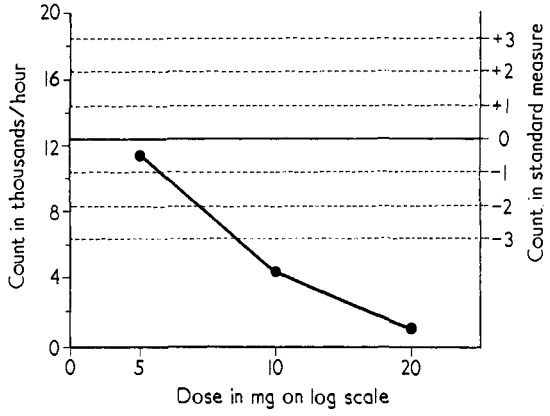

Fig. 3. The action of chlorpromazine on caffeine-induced hyperactivity of mice. The count of jiggle cage movements at several doses of chlorpromazine is plotted on a chart which also indicates the effect of isotonic saline both in counts and in units of standard measure.

Certain well established principles which have broad application are being followed in using this approach. These might be stated as follows.

1. All compounds are considered to be at least as active as the saline, glucose, water, or food used in control animals.
2. Only compounds having activity which is beyond some predetermined value are considered to be interesting.
3. The predetermined value for interesting activity should be substantially different from the mean effect of saline, etc.
4. The test made should answer only one question, 'interesting or not?'
5. Any determination of 'how interesting' will be a completely separate undertaking.

The third and fourth points are of great practical importance.

Unfortunately, to the chemist who made the compound, these concepts may be particularly repulsive. For example, he may wish to know whether his compound possesses even the slightest degree of superiority over saline, or in another case he may wish to ascertain, with considerable accuracy, how its effect, even if slight, compares to that of another known active compound. However, from the practical viewpoint, if the predetermined value is only slightly different from that of saline, most of the available physical capacity for running the test and probably all of the intellectual capacity for coping with the results can be consumed in testing just a few compounds, because large numbers of animals will be required. It is important to recognize that in searching blindly it is usually necessary to investigate a very large number of compounds before finding one which is interesting and therefore the method of investigation must be efficient and economical.

Several considerations are unusual in connection with screening. A test may be designed so that it will be quite sensitive in accepting compounds whose activity is just at the level defined as being of interest. Another design for the test may be less sensitive in this respect, but, paradoxically, more likely to find interesting compounds in the long run. This is shown in Table III for a hypothetical test.

Table III. Paradoxical effect of test sensitivity on yield of active compounds

(1) Type of sequential test	(2) No. of compounds screened	(3) Ability of test to find 'interesting' compounds	1% of (2) × (3) No. of 'interesting' compounds found
2+2 mice	4900	0·62	30
2+4 mice	4810	0·72	35
4+4 mice	2410	0·92	22

Regardless of the test design, a constant number of animals is available for use; let us assume 10,000 per year in this case. The frequency of interesting compounds in the general population is fixed and independent of the test design; in this case it is assumed for convenience to be 1 in 100. The criterion for interesting

activity is defined as 50 per cent protection or more in a test where saline can protect only 1 per cent of the animals. Three two-stage sequential designs for the test are considered, but in every case at least two mice must be protected to accept a compound as interesting. If two mice are protected at the first stage the compound is accepted, while if none are protected the compound is rejected, and in both these cases there is no second stage. Therefore, there will be a second stage only when a single mouse is protected at the first stage. In the first line of Table III, $2+2$ means two mice tested first and if a second stage is run it will also be on two mice.

Analysis of the performance of these three designs shows that as more mice are used in the design, fewer compounds are screened per year (column 2), but the sensitivity of the test or its ability to detect interesting compounds rises (column 3). However, the number of interesting compounds available for discovery is only 1 per cent of column 2. The total number of interesting compounds found by the end of the year in each case is, therefore, 1 per cent of the product of the second and third columns. A very high price can be paid for increasing the sensitivity of the test.

There are other special requirements in the case of screening, as the term is being used here. First, there must be a capacity for testing simultaneously a rather large number of substances. Secondly, there must be available a very large number of substances to be tested and these should represent a diversity of chemical structures. It is admirable if a chemist recognizes that his reasoning may have a flaw in it and that a compound may not perform for the intended purpose as expected. Then, logically, he will make an ample supply for trial in a variety of tests. Such a practice results in a file of compounds available for many kinds of studies. Of course, the ideal is for the supply of every compound to be inexhaustible. This impossible goal can be approached if in practice it is a strict rule that not more than one-half of the available sample can be used for any given test. Although this leads ultimately to the practical exhaustion of the compound, it should still be possible to perform a mixture melting point determination on the original with a replenishment sample for the purpose of establishing identity.

The third and final requirement is most important; there must

be maximal economy in both usage of compound and of animals in the test. As Table IV shows, a conflict in views may arise, but

Table IV. Inter-relationship of compound and animal usage.

Compound sample size per test	Number of animals per compound test
High Usage	High Usage
Every compound should be tested at maximal tolerated dose.	The test should be as sensitive as possible.
Low Usage	Low Usage
The least possible amount of compound should be used in the test.	The largest possible number of compounds should be tested.

in the case of screening this must be resolved in terms of low usage. As an additional argument, the cost of laboratory synthesis of chemicals is much more than most biologists realize and probably is about $2,000 per sample in the American pharmaceutical industry. Similarly, although animals are relatively cheap individually, their costs become an important factor collectively in a screening programme. Always to be remembered is that the more compound required for a test, the fewer compounds available; and the more animals used per compound, the fewer compounds tested per year.

The Use of the Sequential Method in Screening

It may well appear that there is no real difference between the sequential method and conventional procedures for screening. In principle this is correct. The two kinds of error are inherent in either method, the criteria of interest and no interest being chosen by judgment in both; when activity of a low order is encountered, the same uncertainty about the decision is present with either approach, and the course of action in these cases, namely, to repeat the test, is identical. In practice there are some important differences. With the sequential method, the magnitude of the risks of the two kinds of error to be accepted and the levels of activity which are of interest and of no interest are specified. Having done this, the test can be designed so that the number of animals

needed to reach a decision will always be at or near a minimum. This is an advantage of paramount importance in screening in view of the fact that it maximizes the capacity for testing and minimizes the cost. This alone justifies giving the sequential method the most careful consideration.

There are a number of commonly raised objections to the use of the sequential method. First, a definite risk of rejecting an interesting substance must be accepted. With regard to this objection, one must ask whether it is worse to know the risk than to ignore it; such a risk is inherent in any form of test. A second objection is that considerable time may be required to reach a decision if the test period is relatively long and there are multiple stages to the test. Sequential testing gives quick answers only for the very interesting or uninteresting compounds; anything intermediate will require extended testing. However, if increased test capacity is gained by sequential testing, a larger number of final decisions will be made at each cycle of testing than by more conventional methods. A third objection is that the test is rigid and inflexible. Actually this is essential for any repetitive test if it is to be successful, economical and under control.

A fourth objection is to the limited information yielded by a sequential test; namely, the compound was either interesting or not. If the difference between screening for activity and evaluating it, if it exists, is borne in mind, this is not a serious objection. In random screening one must expect that large numbers of compounds will be of no interest in the particular test. It is true that a skilled observer might glean some useful information unrelated to the test itself by making additional observations on the animals. In this fashion a tranquillizer or antidepressant might be discovered in the course of a test to determine antibacterial activity. Practically speaking, this is an objective which cannot be realized because the necessary observation would probably place a limit on the number of samples which could be both tested and observed. Because of the economy in animal and compound usage by sequential procedures, separate tests for other interesting properties in a compound are possible and more practical.

One other objection is that very few types of activity are suitable for the sequential approach. For those who question the

versatility of the sequential method, Table V presents some kinds of drug activity for which satisfactory sequential tests have been developed.[12–15]

Table V. Sequential screening methods for various activities of drugs

Anticonvulsant	Hypoglycemic
Hypotensive	Antineoplastic
Diuretic	Antibacterial
Analgesic	Antiviral
Motor	Antiparasitic

It may be helpful to illustrate in some detail a particular sequential test and analyse its performance. The example selected is screening of substances for antineoplastic activity against the transplanted mammary adenocarcinoma 72j. The tumour is implanted into C3H mice 14 days before the first treatment. On the thirteenth day the mice are sorted into three tumour size groups by palpation. Each size group is used separately. The first point to examine is the effect of intra-abdominal administration of a fixed volume of 1 per cent buffered starch solution (the vehicle) daily for six days on the size of tumours harvested the day after the last injection. The mice used were drawn from the medium sized pool on day 14 after implantation. Fig. 4 shows that a size range of tumours resulted on the twenty-first day, which centred around a mean weight of 0·8 g and ranged from 0·2 to 2·0 g.

It is evident that treatment with a vehicle containing no drug at all can be associated with considerable variation in the size of the harvested tumour. It will be useful to consider this variation in the form of a ratio and in the case of Fig. 4 this can be done by dividing each tumour weight class into the average weight (0·8 *g*) of all 60 tumours. Some of the resulting classes are shown as the lower abscissal scale which is labelled C/T. This demonstrates that the average tumour was about four times larger than the smallest while being only about one-half the size of the largest. Clearly the size distribution is not symmetrical and this will require special consideration.

If this system is to be used for judging activity of drugs, it becomes necessary to convert the data to a form in which statistical methods based on a normal distribution can be applied. This is achieved for all practical purposes if the data are converted to logarithms and used in this form for computing the design of the test. However, once this is done, all subsequent decisions can be made using the data in its arithmetic form.

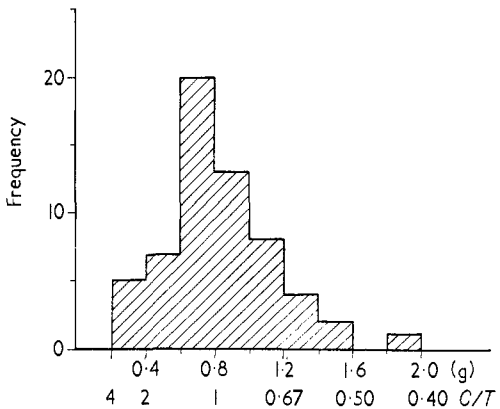

Fig. 4. Frequency distribution of weights of adenocarcinoma 72j tumours removed from 60 mice 21 days after implantation. The tumour was implanted in a large group of mice. Thirteen days later the mice were sorted into three classes (small, medium, and large tumour bearing). Sixty mice from the medium sized class were treated with a daily injection of 1 per cent starch solution from day 14 to day 20. The upper scale of the abscissa shows actual weight of tumour in grams. The lower scale shows the ratio—mean weight of 60 tumours/weight class, e.g., 0·8 g/0·2 g = 4·0, etc.

Within a given run or cycle comprising animals which receive only the vehicle and animals receiving various drugs to be examined for possible activity, the ratio of average tumour weight of controls to average tumour weight of a group of drugged animals can be found. This ratio will be referred to hereafter as the C/T ratio, i.e. control/treated. It has three advantages. First, it meets the requirements of the statistical design in which we are interested, i.e. in logarithmic form it is normally distributed. Secondly, from run to run, the results are converted to a standard

base, i.e. no effect is represented by a C/T value of 1·0 even though the average control tumour weight will vary between runs. Thirdly, the C/T value is informative, i.e. the larger the value, the more probable the drug is active as an antineoplastic agent in mice. There remains, however, the need to specify the values of C/T ratios which represent the criteria of interest.

Considering the scatter in the starch-treated control tumours, it was decided in the case of this test that activity of drugs which resulted in ratios of control to treated tumour weights of 1·43 or less (that is, tumours 70 per cent or more of the size of control) was of no further interest, but ratios of 5·0 or greater (that is, tumours 20 per cent or less of the size of controls) represented interesting activity. It was further decided that not more than one out of twenty uninteresting compounds should be accepted erroneously, and not more than one out of twenty interesting compounds tested should be rejected erroneously. By sequential analysis it was calculated that a decision to accept or reject would be reached with an average usage of 4 to 6 mice per compound. The unit test group was accordingly made up of 3 mice. Each run or cycle would consist of one group on a known active substance, two groups on a 1 per cent buffered starch solution, and the remaining groups on the compounds to be tested. THIO-TEPA was selected for use as the known active substance. The complete cycle for one stage would span 21 days; 14 for implantation and growth to palpability, and 7 days for treatment, removal and weighing of tumours. Even though this is an inconveniently long cycle, each week a new cycle starts and an older cycle ends. At the end of a cycle, all drugs tested would be classified as those of interest, those of no further interest, and those which could not be classified. Each of the latter would be subjected to trial in an additional three-mouse group. It was decided also that no compound would go through more than three cycles without a decision. Instead, at the end of the third stage on such a compound, it would be classified interesting or not. With a maximum of three stages of testing it was then possible to calculate that about 79 per cent of the compounds ought to be classified as interesting or not at the first stage, 18 per cent ought to require two stages, and 3 per cent ought to go on to three stages. After several successive runs, a levelling-off of final decisions at 81 per

cent of the compounds in any run would be expected. One additional decision was made; namely, that the maximum daily dosage to be used would be 250 mg/kg. Note that, throughout, the basic principles outlined earlier have been followed.

Having carried out this testing scheme for a considerable time, it is now possible to present data illustrating how the test has operated. In Table VI is presented a summary of 12 months' operation of this particular screening programme. Eighteen hundred and thirty-eight compounds required a total of 2,922 tests to reach a decision. This is 1·6 tests per compound and

Table VI. Twelve months' screening with mammary adenocarcinoma 72j

Kind of item	No. of items	No. of tests[a]	Tests per item	Mice per item
Unknowns	1838	2922	1·59	4·8
THIO-TEPA	1	136	—	—
1% Buffered starch	1	286	—	—
	1840	3344	1·82	5·5

[a] 3 Mice per test.

4·8 mice on the average. The standard drug and starch controls required a total of 422 additional 3-mouse groups, so that in processing the 1,838 compounds an average of 1·8 tests or 5·5 mice per compound were used altogether on this tumour.

During the several years this programme has been in operation, data have been collected on compounds which appear to have marginal activity. The practice was adopted that when a compound was accepted as interesting at any one of the three stages, it would always be processed through a screen a second time exactly as if it had not been tested before. This was done, not from lack of faith in the test, but because any interesting qualitative observation should be repeated before embarking on more elaborate studies.

Table VII presents the results of almost four years of operation. During this period, out of more than 4,900 different compounds tested, 211 were accepted at either the first, second, or third stage

as active. All of these were then reprocessed with the result that 67 were rejected as being of no interest. These 67 are believed to represent examples of calling compounds interesting when in reality their activity is very probably less than the level of $C/T = 5 \cdot 0$ which was defined as interesting, but more than the level of $C/T = 1 \cdot 43$ which was specified as being of no interest.

In these terms, the incidence of accepting marginally active compounds was about 14 out of 1,000 compounds screened.

Table VII. 3 years 9 months' screening with mammary adenocarcinoma 72j

Class of result	First test no. of items	Results of second test[a]
Interesting	211	144
Not interesting	4715	67
Total	4926	211
Unconfirmed/total tested = 67/4926 = 0·0136 = 14/1000		

[a] There was a second test only on those compounds which were interesting in the first test.

So far the results from operation of this procedure have been examined in terms of quantities of compounds processed. It will be worth examining the performance of the test to determine if it has been in accordance with its design. The importance of gauging the actual performance in these terms should not be underestimated. Many kinds of factors can invalidate what appears to be a satisfactory procedure. These range from unknowingly designing the procedure on an unsound premise to discovering that a systematic bias has inadvertently crept into the operation. In the design of this procedure, the risk for either kind of error was set at 1 in 20, or $P = 0 \cdot 05$. In using sequential analysis for determining the operating constants of the procedure, certain approximations are made in the calculations. These alter the probabilities of error in the conservative direction, to values less than 1 in 20. In addition, in the calculations performed, the variability of the response of the test animals in the form of the standard deviation is used. If later this standard deviation decreases during successive runs, the risks become smaller and the

converse is true. It is essential, therefore, to be able to detect any major shift in the variability of the test system. Since this variability must include also the response to an active drug, it is necessary to test a known active material in each run.

Data such as those in Fig. 4 establish the variability of the test system with respect to uninteresting drugs. However, this variability must remain reasonably constant from run to run if the decisions are to be made in accordance with the desired risks of error. This point is examined by keeping a control chart as

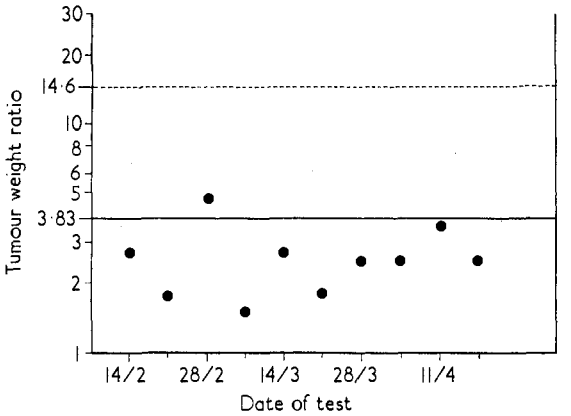

Fig. 5. Control chart for variability of tumour weights in starch-treated mice. In each run there are six mice which receive only 1 per cent starch treatment. The ratio of the largest to smallest tumour weight within the six mice for each run is plotted. The solid line represents the historical value for this ratio. The dashed line represents the historical value increased by three standard deviations.

shown in Fig. 5. Here the ratio of largest to smallest starch-treated tumour weight is plotted for each successive run. The solid line represents the original mean ratio of 3·83 on which the test was designed, and the upper dashed line represents a ratio of 14·6 which is a departure of three standard deviations from the mean. The ordinate represents increasing variability from 1 which is no variability, to 30 which would represent finding a control tumour 30 times heavier than the smallest control tumour in that run. So long as any run gives a ratio falling below the

dashed line and in the general area of the original mean variability, that run is acceptable in terms of the test design. This chart is important also for indicating the development of a trend with time so that appropriate action may be taken. In the case shown here, the ratios are so consistently below the historical mean that it is certain the selected risks have been altered in a conservative direction.

Dr. Charles Dunnett of these Laboratories has developed a method of calculating the exact probabilities associated with a particular design and Mr. Richard Lamm has applied this method to the 72j tumour test system. Table VIII gives the calculated probabilities of making a Type I error (accepting an uninteresting

Table VIII. Compounds which are of no interest ($C/T = 1 \cdot 43$)

Decision	Probability at stage			Sum
	1	2	3	
Accept[a]	0·0054	0·0007	0·0010	0·0071
Reject	0·7839	0·1801	0·0289	0·9929
Sum	0·7893	0·1808	0·0299	1·0000

[a] Type I error.

compound) at each of the three stages as well as the total probability of this kind of error. The probability of a Type II error—the rejection of an interesting compound—is identical because symmetrical risks were specified. The risk is greatest at the first stage and diminishes rapidly at the second and third stages. However, the total risk of 0·0071 is only 1/7 as great as the original specified value 0·05. The procedure is theoretically much more conservative than intended. The sum of probabilities at each stage indicates the proportion of decisions expected, e.g. for stage 1, 0·7893 indicates that 79 per cent of the substances in a run require only one stage of testing.

However, Fig. 5 indicated that over a period of time the variability of control response had decreased. It seemed desirable, therefore, to recompute the probabilities of error in order to find

out the extent to which a reduction in variability had altered the situation. This reduction represented a change in the coefficient of variation of mean control tumour size from 46 per cent to 36 per cent. Table IX shows the recomputed values and again these apply to either type of error. An important point must be

Table IX. Compounds which are of no interest ($C/T = 1 \cdot 43$)

Decision	Probability at stage			Sum
	1	2	3	
Accept[a]	0·0011	0·0001	0·0002	0·0014
Reject	0·8285	0·1557	0·0144	0·9986
Sum	0·8296	0·1558	0·0146	1·0000

[a] Type I error.

mentioned here, namely, these estimates of risk in Tables VIII and IX are calculated from the within-run-variability using data from untreated controls and THIO-TEPA treated animals. In other words, between-run-variation has been excluded from the calculation.

It is one thing to calculate these risks and quite another to attempt to find out whether, in actual practice, the errors made are in agreement with theory. It should be noted that the mistaken acceptance of an uninteresting compound is easily discovered by the procedure of repeating the test on every compound accepted. However, the rejection of an interesting compound will not be discovered because, once a compound is rejected, ordinarily there is no further test. However, there are several ways to assess whether the procedure is providing data in accordance with theory. First, the results obtained with the standard drug THIO-TEPA, which is run in every cycle, can be examined. These ought to relate to the total probability of 0·0014 (Table IX) for erroneously rejecting an interesting compound. Secondly, since in each cycle there are two control groups, one of these can be retained as a control and the other treated as a placebo or a simulated drug which is truly of no interest and should be rejected. This approach attempts to gauge the acceptance of uninteresting

compounds. Thirdly, according to the sums of the first three columns of Table IX, only one stage of testing would be needed to reach a decision on 83 per cent of the compounds, two stages for 16 per cent, and three stages for only 1 per cent of the compounds. These proportions apply either to interesting or uninteresting compounds and since the latter predominate, they can be used as a test.

To take the first possibility—namely, erroneously rejecting an interesting compound—THIO-TEPA has given the results shown in Fig. 6. This shows the results of 89 successive tests on this

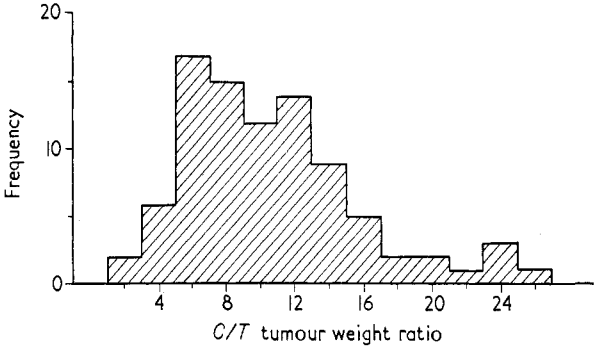

Fig. 6. Frequency distribution of 89 successive C/T values for THIO-TEPA obtained using the medium sized 72j tumour and a dosage of 5 mg/kg day for 6 days.

compound using the medium sized 72j tumour. The ratios of control to THIO-TEPA tumour weight have been grouped from small to large. It is evident that in successive three-mouse groups this compound gave C/T values covering a wide range. The average value was around 9 which is well above the level of 5 defined as being of interest. However, once out of the 89 runs, a ratio was obtained which indicated that this known active compound should be rejected. THIO-TEPA is also used as a positive control in tests on the small and large 72j tumours and only once out of a total of 247 runs on the three size classes to date has THIO-TEPA been rejected. This is a rejection rate of 0·0040. Unfortunately, the exact frequency of rejection of THIO-TEPA will not be very certain until it has been tested about 10,000 times,

more or less. However, the real value in having THIO-TEPA in every test is that it provides a measure of run-to-run variability.

As to the second possibility, the use of one control group in each test as a placebo or simulated uninteresting compound, it can be calculated that the probability of *not rejecting* such a simulated compound is 0·034. For the purposes of this trial, 94 of 188 control groups were used to simulate compounds. Three of the 94 were *not rejected* and this gives, therefore, an observed proportion of 0·032 which is in close agreement with the theoretical value of 0·034.

The third measure of performance, the percentage of compounds requiring one, two, or three stages for a decision, was examined by drawing a random sample of 200 compounds which had been rejected and determining the number of cycles which had been used for each compound in reaching the decision to reject it. The results are shown in Table X which summarizes the three measures of performance of the test.

Table X. Actual performance compared to theory

	Probability	
	Calcd.	Found
Reject THIO-TEPA (C/T = 9·0)	< 0·0001	0·0040
Not reject 1% starch	0·034	0·032
Decisions made at: 1st stage	0·83	0·85
Decisions made at: 2nd stage	0·16	0·13
Decisions made at: 3rd stage	0·01	0·02

The possible discrepancy between performance and design with respect to THIO-TEPA led to a review of the particular run in which this compound was rejected. No explanation could be found for the occurrence of this apparently unlikely event. Next, the experience with this compound was reviewed (see Fig. 6). This suggested that the variability between runs was larger than was considered in the design of the test. Finally, the variability between runs was calculated by using all the data on THIO-TEPA but excluding two outlying values, one of which was the C/T

value indicating a rejection and the other a very high C/T value. The results of this are shown in Table XI and represent the best

Table XI. Compounds which are interesting ($C/T = 5 \cdot 0$)

Decision	Probability at stage			Sum
	1	2	3	
Reject[a]	0·0293	0·0068	0·0050	0·0411
Accept	0·7197	0·1908	0·0484	0·9589
Sum	0·7490	0·1976	0·0534	1·0000

[a] Type II error.

estimate of risk of rejecting compounds whose activity is just at the level defined as interesting. The risk of rejecting THIOTEPA when it had a C/T ratio of 9·0 was also calculated and found to be 0·0011 which is much larger than the value of $< 0 \cdot 0001$ shown in Table X.

The possibility seems to exist that the risk of accepting uninteresting compounds conforms to a theoretical risk calculated from the within run variability, while the risk of rejecting interesting compounds conforms to a theoretical risk based on variability between runs. Further study will be needed to verify this possibility. Aside from this one uncertain area, it seems clear that the degree of agreement with theory is such as to validate the test procedure from the technical standpoint.

It is a pleasure to acknowledge the interest and help of my colleagues in developing the thoughts expressed in this paper. Dr. Adolph Vogel and Mr. Jack Haynes both contributed actively to the material presented.

(*Received 17 February, 1960*)

References

[1] Meldrum, N. U. and Roughton, F. J. W. *J. Physiol.*, **75**, 15*P* (1932)
[2] Southworth, H. *Proc. Soc. exp. Biol., N.Y.*, **36**, 58 (1937)
[3] Strauss, M. B. and Southworth, H. *Johns Hopk. Hosp. Bull.*, **63**, 41 (1938)

[4] Marshall, E. K., Jr., Cutting, W. C. and Emerson, K., Jr. *J. Amer. med. Ass.*, **110**, 252 (1938)
[5] Mann, T. and Keilin, D. *Nature, Lond.*, **146**, 164 (1940)
[6] Davenport, H. W. and Wilhelmi, A. E. *Proc. Soc. exp. Biol., N.Y.*, **48**, 53 (1941)
[7] Höber, R. *Proc. Soc. exp. Biol., N.Y.*, **49**, 87 (1942)
[8] Roblin, R. O., Jr. and Clapp, J. W. *J. Amer. chem. Soc.*, **72**, 4890 (1950)
[9] Miller, W. H., Dessert, A. M. and Roblin, R. O., Jr. *J. Amer. chem. Soc.*, **72**, 4893 (1950)
[10] Cochran, W. G. and Cox, G. M. *Experimental Designs*, 2nd Edition, Chapter 2. (1957). New York; John Wiley and Sons
[11] Wald, A. *Sequential Analysis.* (1947). New York; John Wiley and Sons, or London; Chapman and Hall
[12] Osterberg, A. C., Haynes, J. D. and Rauh, C. E. *J. Pharmacol.*, **122**, 59A (1958)
[13] Dearborn, E. H. *ACTA Unio. Inter. Contra Cancrum*, 15 Suppl., **1**, 76 (1959)
[14] Cummings, J. R., Haynes, J. D., Lipchuck, L. M. and Ronsberg, M. A. *J. Pharmacol*, **128**, 414 (1960)
[15] Unpublished data from these laboratories.