MOUSE TO MAN: STATISTICAL PROBLEMS IN BRINGING A DRUG TO CLINICAL TRIAL

MARVIN A. SCHNEIDERMAN
NATIONAL CANCER INSTITUTE

1. Introduction

Problems of inductive inference and the ethics of the clinical trial of new drugs have led statisticians to pay much attention to the stopping rules in experiment designs [1], [2], [3]. I suggest that for clinical trials the statisticians should probably be paying as much attention to the problem of starting rules. When should a new material go into clinical trial? At what dose? At what frequency? By what route?

"Starting" problems also reflect problems of inductive inference, but of a different order of magnitude, or of a different qualitative nature than the problems of inductive inference following clinical trials. There the question is how to extend the findings on a small group of humans, selected in some special way, treated under some special circumstances, by special physicians, to the much larger group of humans suffering from the same illness, but to be treated under less specialized circumstances by less specialized physicians in less specialized institutions. In the preclinical trial phase, the earliest work is done on nonhuman species, almost always with different metabolic schemes and systems than man. The jump from these results to man is sometimes a large leap, made mostly in the dark.

If inferences within a species are difficult, as almost all writers on clinical trials attest by their urging of cautious generalizations and limited extensions of the trial results [4], [5], then what must be the difficulties of making the species to species jump that the preclinical trial work requires? I see three major pretrial problems for which a body of statistical theory may need to be built. Two of the three derive from the species to species problem.

2. Predictability for man of screening in lower animals

The first problem is, why try a given new material in man? Either a strong biochemical rationale has been developed for the action of the material, or work in lower animals has shown activity against the illness in which we are interested. Signs of this activity are often found through a screening process, and the statisticians have attended to the development of efficient screening procedures

[6], [7], [8]. They have done much less, however, in developing techniques for sorting out whether the screens are effective predictors for the same disease in man. For example, in the early work of screening for antimalarials, the avian screens used were not good predictors for activity in man. A primate screen using monkeys as the test animal eventually led to the development of useful substitutes for quinine [9], [10]. Today's anticancer screens also may be unsuitable.

Evaluation of the effectiveness of the animal screens in the anticancer area is beset by equal, or perhaps greater, difficulties. The animal screens have to be looked upon not as predicting devices but rather as techniques for concentrating or enriching the mix from which one selects the materials to try in man. A good screen is a device for increasing the probability that a material when tested in man will yield positive responses. Within the framework it would be meaningful to find which screen(s) gives the maximum concentration, or whether it might be possible to combine the results from several screens in some rational way to give a higher concentration. Since almost only those materials showing substantial activity in the animal screens are ever tested in man, suitable data to develop correlations as the statistician understands them are not available. Correlations computed from truncated samples, as these are, substantially underestimate the true correlations [11].

There are some problems of costs that need to be considered. The problem may need to be formulated several different ways. The first, and obvious cost, is the cost of false positives. A screen calling a treatment positive which later proves to be useless in man carries with it the ethical cost of testing an ineffective drug in man—possibly even depriving the patient of a better treatment. In this formulation, the measure of worth of a screen would be the "concentration" of true positives achieved. The higher the concentration, the smaller the proportion of false positives, and hence the lower the ethical costs. However, a pure concentration technique runs the risk of losing potentially useful materials (that is, false negatives) and a way to allow for the cost of discarding useful materials, depriving patients of their potential, must be developed. Dunnett [12] proposed that the cost of one false negative was what it cost to produce one true positive—because for each material lost, one new one would have to be found to replace it.

Relating "activity" in animals to activity in man assumes that there are good and suitable measures of activity in man. It is toward this latter goal that the controlled clinical trial, with a placebo as one of the first agents to be entered into comparison, is directed. There is evidence, in cancer at least, that measured responses to the same material differ substantially from time to time and place to place [13], [14]. What is active in one form of the disease is not active in others. Several people have suggested that these differences are a reflection of patient differences more than anything else [15], [16]. The statistical problem here lies in the design of trials, and in the discovery of predicting variables (or covariates) that would tend to minimize these patient differences, or at least permit the experimenter to take them into account when analyzing his results.

Whether the techniques of the psychologists such as factor analysis and related methods have any place here, has not been determined. Some workers have tried regression techniques [17]. At least one approach using a reliability type attack seems to have yielded a meaningful result [18]. Some quite simple classification schemes have proved to be highly discriminatory. One, involving only the physician's subjective estimates of expected survival before treatment [19], was able to segregate patients into a group which had a 30 per cent response rate in contrast to another group which had a 70 per cent response rate. The other, utilizing two or three blood count measures [16] had similar success in classification.

On the other side of the coin is the toxicity of the material. Almost all therapeutic agents have some toxicity for some men. The problem here is the same problem as the problem of screening for activity. If one screens for toxicity, how much faith, emphasis, reliance does one place on signs of toxicity shown in animals as an indicator of toxicity in man? This, of course, is the problem facing any regulatory agency. If aspirin leads to stomach wall erosion in cats, does it mean that it is unsuitable as a drug to be used in man? If quinine devastates dogs, should men not use it? If thalidomide appears to produce little in animals except, perhaps, some minor signs of central nervous system toxicity, is it a suitable material to give to pregnant women for control of their "morning sickness"? The regulatory agencies may be in a "no false negatives" situation. It may be necessary for them to attempt to permit no possibility of ever saying, "This is a nontoxic material"—and then have it prove to have some toxicity on (even a small number of) humans. This certainty is incapable of achievement. In the treatment of very serious diseases, the limitation of no toxicity whatever is not required. It is recognized that activity is bought at a price of some toxicity. The disease and its natural consequences determine the maximum toxicity permissible. As one approach, attempts are made to produce the greatest yield for a permissible toxicity.

How useful then are the animal screens in foreshadowing toxicity in man? What can the statistician do to help interpret the animal data and make them more meaningful to the physician who will eventually treat men—even at a risk of some toxicity? In the history of the anticancer drugs, some kinds of toxicity in man seem to be well forecast by the toxicity seen in experimental animals—and others are missed quite frequently [20]. Litchfield [21] in speaking of drugs in general remarked that: "Many of the most serious side effects that can result when a drug is given to man were not predictable from observations on dogs or rats. . . . [But] effects on man could be predicted better from observations on dogs than from those on rats."

Bound up in this problem of what the pharmacologists like to call "qualitative predictability," is the problem of quantitative predictability. If one is unable to predict what *kind* of toxicity will appear in man, how is one able to suggest a dose of drug that will produce only minimal toxicity? Though this is a dilemma, it is possible to escape between the horns. This requires a redefinition. If toxicity

is defined as "any bad effects that would cause us to limit or reduce the amount of drug administered," rather than a given amount bone marrow damage, or liver damage, or central nervous system derangement, then it may be possible to find doses that do not produce dose limiting toxicity—in spite of the poor specific systems prediction from animals. Perhaps an answer can be gotten to the question: "Is there limiting toxicity in man that is not predicted by the animal systems?" For this, one cannot assume that toxicity to each major system develops independently of the others, and that the only limitation is that excessive toxicity is reached in a single system. Clearly, sublimiting toxicity in several systems may leave a patient sufficiently ill that we would call him "toxic" and stop treatment. However, as a first approximation, these systems might be thought of as independent and the system which shows the greatest toxicity is the system that limits treatment. A review of the data is now in order to see which combination of animal systems is capable of predicting "limiting" toxicity. This approach will be of less help in those illnesses for which limiting toxicity must be very slight, because the illness, itself, is very minor. In these diseases the pressure to find new materials is probably also slight.

3. Starting dose in man

The next operational stage after deciding that there is enough information from the animal work to justify going to man, is to go to man. But at what dose does one start? And how does one increase (decrease) it? Since the essential requirement in working in man is to not harm him ([22], p. 102), the initial exploratory doses of a drug must be low with respect to the dose expected to cause toxicity, yet high enough to show some positive activity. Such a dose may be unattainable if the drug is incapable of producing such effects. Usually, worrying about toxicity first, the drug is started at some presumed very low dose and then escalated in the hope of reaching acceptable activity before reaching unacceptable toxicity.

In research on new anticancer drugs, the advice is often given [20] to give the first dose in man at 1/10 the maximum tolerated dose in the most sensitive species of lower animal in which the material has been tested. If dose is computed on a mg/kg basis, it is no great surprise to find that very often the "most sensitive species" is the largest (heaviest) species on which the material has been tested. This has led to considerable suspicion among pharmacologists that not all drugs are metabolized in some way proportional to the weight of the organism [23]. At least one instance has been recorded in which a drug was given to an elephant, basing the dose on what would be suitable for man on a mg/kg basis. The result was an elephantine disaster [24]. A suitable starting dose in man might better be based on a mg/m^2 basis, that is, on a surface, rather than a volume, or weight basis [25]. This observation has yielded some interesting results [26]. Some proponents of the mg/m^2 school (in the anticancer drug area) propose a starting dose at 1/3 the maximum tolerated dose in the

animal work. There is less talk of a "most sensitive species," since the surface measure as a metameter brings the species closer together. This result in the antitumor agents may not apply to other classes of materials. Brodie, Cosmides, and Rall [27] caution: "Most anti-tumor agents should be considered apart from agents which affect complex homeostatic systems." This would seem to mean that whatever generalizing schemes one develops from observation of the antitumor drugs may not apply to other drugs. But even within the limited context of the antitumor drugs, two questions still remain. Why 1/10, or 1/3 (or any other fraction) of the animal dose? And having started at this level, how does one proceed to raise the dose, if one must, in a cautious, reasonable fashion?

Clearly the 1/10, or 1/3, or $1/n$, derives from personal experience. In spite of the vagueness of the term, "maximum tolerated dose," people who have worked with these drugs seem to have found that most of the time a dose at 1/10 (or 1/3) the most sensitive animal dose produced no trouble in man. The fraction chosen is a function of the slope of the dose response curve. If 1/10 the dose computed on a mg/kg scale gives about the same toxicity as does 1/3 of the dose computed on a mg/m^2 scale, then some assumptions have been made about the slopes of response curves and the acceptable starting level of toxicity in man. An infinite number of combinations of these assumptions will satisfy these conditions. The slope and toxicity level should be stated explicitly, so that when working with new materials, for which slope information can be developed, starting levels other than 1/10 or 1/3 will be used if indicated, to achieve the desired low levels of toxicity.

Having started at some "safe" dose, the problem is still to move from this level to an "effective" dose. In some ways, this movement resembles the sequential procedures suggested by many statisticians [28], [29], [30] for developing a dose level for an LD50 (or some other response level). But the up and down procedures, and the usual "overshooting" in animal experiments are not ethically acceptable in an experiment on man. We must be able to come up to the proper dose from below, slowly working up to the higher level at which toxicity is more likely to occur. Several suggestions have been made for edging up to the proper dose—but none are fully satisfactory. Probably the least satisfactory was a suggestion [31] for an escalation of dose, beginning at a fixed level d, choosing some arbitrary, but small, dose interval Δ. If the first trials at dose d show no toxicity, the next trials are carried out at dose $d + \Delta$. If there is still no toxicity the following dose will be at dose $d + 2\Delta$; and subsequent doses at $d + 4\Delta$, $d + 8\Delta$, and so forth. The increasingly larger steps between successive doses have made this suggestion one which many physicians feel embodies too much risk of jumping to a very toxic dose.

A decreasing step suggestion also has been made. This is due to Bellman ([32], p. 342) in another context, and I have not seen it in any published account of preliminary dose finding. The Bellman suggestion is a form of Fibonacci search. Three decisions have to be made here: the initial dose d, the maximum possible dose d', and N, the number of steps allowable in moving upward from

dose d to dose d'. By taking a Fibonacci series of length $N + 1$, inverting the order, and spacing the doses in proportion to the N intervals in the series, one would take smaller and smaller steps in moving from d to d'. This cautious approach has considerable appeal.

Several problems remain unsolved even with the Fibonacci search. If d', the "highest possible" dose, is reached and still no toxicity is seen, should a new "highest" d'' be set, and a new N, say N', and a new search sequence followed from d' to d''? The first jump in a new search sequence will almost always be larger than the last jump in going from d to d'. How does one justify this in a program dedicated to cautious "creeping up"? Several suggestions have been made, including that each dose above d' be made larger by an amount equal to the last step in the prior search that brought one to d'. This, of course, loses the efficiency of the Fibonacci search. No guide seems to exist for the choice of the number of steps, or terms in the basic series—and, of course, all the problems inherent in any other system requiring a decision about a proper "lowest" dose d, exist here, too. A problem of sample sizes exists here, too. Should there be equal sample sizes (how many?) at every step, or should the number of patients at a step increase (decrease) upon approaching the high dose?

There are complications even within a species when searching for an optimal dose. Probably it is best to find a possible starting dose, and then work from there to an optimal dose—within the same species. Extrapolation between species of an optimal dose at this point seems unwise, and unattainable. As an example of some of the difficulties, an attempt was made to use toxicity data on normal mice to indicate the best routes and dose levels (that is, the routes and level which would give longest median survival), for two drugs, given by two routes [33]. Table I shows that in none of the four cases did the normal animal "maximum tolerated dose" coincide with the dose that produced the longest survival in tumor bearing animals, although in no case was the difference more than two dose steps away.

TABLE I

"MAXIMUM TOLERATED" DOSE VERSUS "LONGEST SURVIVAL" DOSE

Maximum dose always is highest dose with no diminution in median survival, or dose 1 step below lowest dose which produced decreased survival in normals. Part of the problem here may be attempting to equate a rather vaguely defined "maximum tolerated" dose, with a more sharply realizable "optimal" dose.

Drug:	MTX		DCMTX	
Route:	Subcutaneous	Oral	Subcutaneous	Oral
Max. dose in normals	1.3	2.1	45	27
Dose giving longest survival in L-1210 bearing mice	.76	3.5	75	75
Dose steps difference	−1	+1	+1	+2

Other procedures for starting dose finding have been attempted in man with varying degrees of success. When some preliminary "safe" dose data are known, a two stage procedure has been used. The first stage tries several doses as a preliminary, to verify the safety of the doses. Toxicity is observed at each of these doses, and those that produce too high toxicity are ruled out at once. The second stage of the search then narrows to one or two doses, apparently closer to the level that gives acceptable toxicity. This approach was used by DeVita and his colleagues [34] in the comparison of three different dose schedules of the same drug.

In the first stage of the DeVita study this was the plan of attack:

single dose given at	300, 400 mg/m^2 once;
daily dose given at	100, 150, 200 mg/m^2 once a day for 3 days;
weekly dose given at	25, 50, 75, 100, 125 mg/m^2 once a week for 6 weeks.

Toxicity began to appear at some of these dose levels, and levels for the second stage were chosen:

single dose	250 mg/m^2;
daily for 3 days	125 mg/m^2;
weekly for 6 weeks	90 mg/m^2.

However, unforeseen troubles arose, and the authors' remarks are worth quoting: "The results indicated a discrepancy between toxicity in the preliminary (1st stage) and subsequent (2nd stage) trials. For instance, in this study one would not have expected to see a difference in toxicity at the 90 and 100 mg/sq m weekly dose schedules. Yet the data indicate that in the 2nd stage of the study a dose of 90 mg/sq m/week was less toxic in terms of leukopenia and thrombocytopenia. . . . This discrepancy also applies to the 125 and 150 mg/sq m dose levels. One possible explanation is that a certain amount of bias exists in that patients with a poorer risk are selected for the early clinical trials. As information becomes available, the investigators obtain experience in using the drug, and patients with less advanced disease are probably selected."

What does this imply for the statistician (and physician)? It implies that experiment designs have to be developed that can reduce the patient selection bias that seems to have entered this otherwise very well designed study. For single, and isolated studies this may not be possible. But where a research group is involved in a continuing program of drug evaluation, it may be attainable. An enlargement of statistical techniques to encompass research programs rather than one at a time studies appears to be in order. A randomization scheme between studies, as well as within studies, might achieve it. Patients would be randomized to one of several studies, and then to the specific treatment and dose within each separate study. Thus, if a research group were working with several drugs at once the first randomization would assign the patient to drug X and the second randomization would assign him to treatment α within drug X. The difficulties of changing quality of patients would be largely avoided by

having all patients eligible for all studies going on at any time. This might be administratively difficult to work out, but the administrative price would be worth paying to avoid the scientific or inferential error that DeVita suggests: "This [bias] . . . may result in the selection of less effective dose levels and contribute to lower response rates."

DeVita touches on an important philosophic problem here, too. If patient selection bias could lead to a less effective dose, it may be important to specify that early trials be conducted on the same kind of patients that one hopes to eventually use the drugs on. On the other end of the scale perhaps we may spare future generations of medical students if it is demonstrated that finding toxicity levels in "normals" does not tell too much about proper dose levels in sick people.

4. Optimal dose

Dr. DeVita's comment leads directly to the problem of optimal dose. If the candidate drug is given at a dose far from its optimal level, we could well underestimate its effectiveness. In fact, in comparative, controlled clinical trials, if there is any bias, it is most often against the new drug. Comparing a new material against an older one is comparing not only materials but also all the collective experience in handling the materials. Knowledge of how to handle the old material, how to administer it, how to treat its side effects, tends to increase the probability that a patient on the old drug will be treated better. Perhaps some organized way can be developed to give new drugs the added help they need in competing with the older ones.

First, the multiple actions of the drug must be recognized—if no more than to categorize some of these into "positive" actions, and some into "negative." Then, the competition between these two sets of actions must be taken into account. The ideas of decision theory seem to have some place here, and the introduction of cost functions gives promise of finding a route to an optimum—the properly "balanced" dose. I have suggested an approach to this optimal dose finding through minimizing a cost function which consists of two parts: the costs due to failure to produce a positive response; and the costs ascribable to producing toxicity [35]. The two are tied together by a weighting factor which attempts to assess the relative worth of the two elements. This weighting factor is a subjective function of the disease to be treated. Thus, in a self-limiting disease of minor consequence the cost of toxicity is far greater than the cost of the failure to produce a response, and the weighting factor must be large (for example, the treatment of "morning sickness" in pregnant women can permit no drug which could damage the foetus).

An attempt was made to apply this minimum cost concept in an artificial situation under some very restricting assumptions, using Monte Carlo techniques [36]. The work done implied the need for more sophisticated stopping rules, and a better model of the disease. This paper raised more questions than it answered.

Specifically, the dose finding procedure used (a hypothetical) three patients at a dose, and assumed the response indicated either a pressure to raise the dose U, or lower the dose D, for the next stage of the trial. The eight possible configurations of the U and D each led to a new dose configuration at the next stage in the trial. Some of the eight combinations of the U and D led to half step jumps in dose. Stopping rules were promulgated based on the repetition of dose patterns, much as in the up and down techniques in bioassay. Monte Carlo procedures were used to determine how effective these rules were, whether they led to unbiased estimates (they did not!) and to determine the average duration of experiments. Here are some of the deficiences of this scheme.

"First, there is the unreality of some of the assumptions. Each 'patient' is used once. . . . Our 'up-down' criteria are not good. We have limited ourselves to yes-no responses. Strong indications for movement (that is, great toxicity) should lead to larger dose steps than the steps indicated here. We have not considered speed of reaching toxicity, or recognizable activity. Our definition of optimality conceals an assumption about how we react, for example to a toxicity-with-no-response result compared to a no-toxicity-with-no-response result. We have not considered optimal spacing in time or route of administration.

"We have examined only one dose-finding configuration. Our stopping rules are arbitrary. Our movement rules contain only one shortening of the dose interval. An estimation procedure which is sensitive to the narrowing of the dose interval is needed. Perhaps the stopping rules can be tied to the estimation procedure. We have not examined whether a two-dose stage, or a four-dose stage, or more than one individual at a dose stage (for example) could produce earlier stopping, fewer way-out doses, or a more uniform average experiment size. In fact, we have not defined a measure for optimality in deciding between different dose-finding configurations. Different dose-finding configurations need to be invented, and tested for comparison with the one presented . . . and perhaps an optimal procedure for finding an optimal dose can be found."

There are other ways of trying to counter the problem of possible bias introduced by giving a drug at only one dose. For example, in a comparative clinical trial, where there is uncertainty about the best dose for the new drug, it might be possible to give the new drug at several dose levels, for comparison with the old drug, given at its optimal level. This raises experiment design problems with ethical consequences. Should the total anticipated number of patients be split between the new drug and the old, new drug patients further split, into the various dose groups? Or should we consider each dose level as a different drug, to be compared with a single control and increase the number in the control group by the square root of the number of doses, as suggested (in another context) by Dunnett [37]? Or should we take our N patients and divide them equally among the k dose groups, plus one control group, assigning each group $N/(k+1)$ patients? The Dunnett inspired suggestion appeals to me. For k different dose groups on the new drug this assigns the proportion $k^{1/2}/(k^{1/2}+1)$ of the patients to the new drug. If the largest number of doses one would use for a new drug

would be four, this would assign 2/3 of the cases to the new drug. This might be acceptable ethically.

5. Summary

Three major statistical (experiment design) problems have been described in bringing a new drug to clinical trial. Two of these derive from the difficulties of extrapolation between species. The other is a within species difficulty, and is no less for experimental animals than it is for man. Questions of ethical experimentation on man make this last problem more difficult to solve in man than in experimental animals.

The three major problems are these.

(1) What is the predictability for man of screening (both for activity and for toxicity) in lower animals? How can it be improved? Is it meaningful to talk about correlation, as the statistician uses this term, in this context? We conclude that it is not, and that some other measures of "goodness" are involved.

(2) Given animal toxicity data how does one find a safe and satisfactory starting dose in man? We suggest a combination of information on both toxicity and positive response, and a more explicit awareness (and use) of the slopes of the appropriate dose response curves. Some schemes for randomization of patients are suggested for groups studying more than one drug, to help avoid patient drift bias.

(3) How does one find an optimal dose for a new material which is to be tested in a controlled comparative clinical trial? (Anything other than optimal leads to bias against the new material.) Two not wholly satisfactory proposals are made, and reasons given for their unsatisfactoriness. These procedures are a minimum loss approach, and a dose response approach. No other explicit formal dose finding methods are known to us.

The problems of the starting rules for clinical trials designed to lead to proper inferences appear to be as difficult as the problems of stopping rules for clinical trials that recent work on sequential designs has brought to attention.

REFERENCES

[1] F. J. ANSCOMBE, "Sequential medical trials," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 365–383.

[2] T. COLTON, "A model for selecting one of two medical treatments," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 388–400.

[3] J. CORNFIELD and S. W. GREENHOUSE, "On some aspects of clinical trials," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1966, Vol. 4, pp. 813–829.

[4] L. LASAGNA, "The controlled clinical trial: theory and practice," *J. Chron. Dis.*, Vol. 1 (1955), pp. 353–367.

[5] A. B. HILL, "The clinical trial," *Brit. Med. Bull.*, Vol. 7 (1951), pp. 278–282.

[6] C. W. DUNNETT, "Approaches to some problems in drug screening and selections," Gordon Research Conference on Statistics in Chemistry and Chemical Engineering, August 1961.
[7] P. ARMITAGE and M. SCHNEIDERMAN, "Statistical problems in a mass screening program," *Ann. New York Acad. Sci.*, Vol. 76 (1958), pp. 896–908.
[8] E. P. KING, "A statistical design for drug screening," *Biometrics*, Vol. 19 (1962), pp. 429–440.
[9] G. R. COATNEY, W. C. COOPER, N. B. EDDY, and J. GREENBERG, "Survey of antimalarial agents," *Public Health Monograph*, No. 9, Washington, U. S. Government Printing Office, 1953.
[10] L. SCHMIDT, "The place of drug development and testing in the research process," *Proceedings of the Conference of Professional and Scientific Societies, Chicago*, Washington, Commission on Drug Safety, 1963.
[11] N. MANTEL, "Corrected correlation coefficients when observation on one variable is restricted," *Biometrics*, Vol. 22 (1966), pp. 182–187.
[12] C. W. DUNNETT, "The statistical theory of drug screening," *Quantitative Methods in Pharmacology*, Amsterdam, North Holland Publishing Co., 1961.
[13] I. D. J. BROSS, "Statistical analysis of clinical results from 6-mercaptopurine," *Ann. New York Acad. Sci.*, Vol. 60 (1954), pp. 369–373.
[14] M. SCHNEIDERMAN, "The clinical excursion into 5-fluorouracil," *J. Chron. Dis.*, Vol. 15 (1962), pp. 283–295.
[15] R. R. ELLISON, "Clinical applications of the fluorinated pyrimidines," *Med. Clins. N. Amer.*, Vol. 45 (1961), pp. 677–688.
[16] Y. SATHE, M. ZELEN, and J. ZWEIFEL, "The interpretation of clinical trials," *J. Chron. Dis.*, Vol. 18 (1965), pp. 385–395.
[17] E. A. GEHAN, "An application of multivariate regression analysis to the problem of predicting survival in patients with acute leukemia," *Bull. Inst. Internat. Statist.*, Vol. 39 (1961), pp. 173–179.
[18] M. MYERS, L. AXTELL, and M. ZELEN "The use of prognostic factors in predicting survival for breast cancer patients," *J. Chron. Dis.*, Vol. 19 (1966), pp. 923–933.
[19] G. L. GOLD, "A clinical comparison of three alkylating agents," in preparation.
[20] A. H. OWENS, "Predicting anti-cancer drug effects in man from laboratory animal studies," *J. Chron. Dis.*, Vol. 15 (1963), pp. 223–238.
[21] J. T. LITCHFIELD, "Forecasting drug effects in man from studies in laboratory animals," *J. Amer. Med. Assoc.*, Vol. 177 (1961), pp. 34–38.
[22] C. BERNARD, *An Introduction to the Study of Experimental Medicine*, (translated by Henry C. Greene), New York, Macmillan, 1927.
[23] P. D. HARWOOD, "Therapeutic dosage in small and large mammals," *Science*, Vol. 139 (1963), pp. 684–685 (letter).
[24] L. J. WEST, C. M. PIERCE, and W. D. THOMAS, "Lysergic acid diethylamide: its effects on a male Asiatic elephant," *Science*, Vol. 138 (1962), pp. 1100–1102.
[25] D. PINKEL, "The use of body surface area as a criterion of drug dosage in cancer chemotherapy," *Cancer Res.*, Vol. 18 (1958), pp. 853–856.
[26] E. J. FREIREICH, E. A. GEHAN, D. P. RALL, L. H. SCHMIDT, and H. E. SKIPPER, "A quantitiative comparison of toxicity data on anti-cancer agents obtained in the mouse, rat, hamster, dog, monkey, and man," *Cancer Chemother. Rep.*, Vol. 50 (1966), pp. 219–244.
[27] B. BRODIE, G. J. COSMIDES, D. P. RALL, "Toxicology and the biomedical sciences," *Science*, Vol. 148 (1965), pp. 1547–1554.
[28] K. A. BROWNLEE, J. L. HODGES, JR., and M. ROSENBLATT, "The up-and-down method with small samples," *J. Amer. Statist. Assoc.*, Vol. 48 (1953), pp. 262–277.
[29] G. B. WETHERILL, "Sequential estimation of quantal response curves," *J. Roy. Statist. Soc. Ser. B.*, Vol. 25 (1963), pp. 1–48.

[30] W. G. COCHRAN and M. DAVIS, "The Robbins-Monro method for estimating the median lethal dose," *J. Roy. Statist. Soc. Ser. B*, Vol. 27 (1965), pp. 28–44.
[31] J. LOUIS, "Coordinated Phase I studies for cooperative chemotherapy groups," *Cancer Chermother. Rep.*, Vol. 16 (1962), pp. 99–105.
[32] R. E. BELLMAN, *Dynamic Programming*, Princeton, Princeton University Press, 1957.
[33] E. FREI, III, C. L. SPURR, C. O. BRINDLEY, O. SELAWRY, J. F. HOLLAND, D. P. RALL, L. R. WASSERMAN, B. HOGSTRATEN, B. I. SHNIDER, O. R. MCINTYRE, L. B. MATTHEWS, JR., and S. P. MILLER, "Clinical studies of dichloromethotrexate (NSC-29630)," *Clin. Pharmacol. Ther.*, Vol. 6 (1965), pp. 160–171.
[34] V. T. DEVITA, P. P. CARBONE, A. H. OWENS, G. L. GOLD, M. J. KRANT, and J. EDMONSON, "Clinical trials with 1,3-Bis(2-chlorethyl)-1-nitrosourea, NSC-409962," *Cancer Res.*, Vol. 25 (1965), pp. 1876–1881.
[35] M. A. SCHNEIDERMAN, M. H. MYERS, Y. S. SATHE, and P. KOFFSKY, "Toxicity, the therapeutic index and the ranking of drugs," *Science*, Vol. 144 (1964), pp. 1212–1214. (Also see, M. A. Schneiderman, N. Brock, M. H. Myers, and B. Schneider, "Melphalan theory and exercise: assessment of drugs," *Science*, Vol. 149 (1965), pp. 1396–1398.)
[36] M. A. SCHNEIDERMAN, "How can we find an optimal dose?" *J. Toxicol. Appl. Pharmacol.*, Vol. 7, Suppl. 2 (1965), pp. 44–53.
[37] C. W. DUNNETT, "A multiple comparison procedure for comparing several treatments with a control," *J. Amer. Statist. Assoc.*, Vol. 50 (1965), pp. 1096–1121.