Reprinted from CANCER, Vol. 38, No. 7, July 1976. 

# THE EFFECT OF MEASURING ERROR ON THE RESULTS OF THERAPEUTIC TRIALS IN ADVANCED CANCER

CHARLES G. MOERTEL, MD,* AND JAMES A. HANLEY, PHD†

**In this study, 16 experienced oncologists each measured 12 simulated tumor masses employing their usual clinical methods. Unknown to the oncologists, two pairs of these tumors were identical in size. This permitted a total of 64 measurement comparisons of the same investigator measuring the same size mass and 1920 comparisons of different investigators measuring the same size mass. If a 50% reduction in the product of perpendicular diameters is accepted as a criterion, the objective response rate due to measuring error alone was 7.8% by the same investigator and 6.8% by different investigators. If a 25% reduction criterion is used, the respective "placebo" response rates were 19% and 25%. In the clinical setting it is recommended that the 50% reduction criterion be employed and that the investigator should anticipate an objective response rate of 5 to 10% due to human error in tumor measurement.**

*Cancer* **38:388–394, 1976.**

IN RECENT DECADES, THE PRACTICE OF CANCER medicine and the technology of experimental cancer therapy have reached progressively higher levels of scientific sophistication. Preclinical screening and small—and large—animal toxicological studies provide the investigator with a great wealth of data for his clinical trials. Remarkable measuring equipment and refined methodology can dissect the pharmacokinetics and the physiologic and immunologic effects of a new treatment modality with almost infinite detail.

Similarly, the laudable new trends towards more objective and more controlled clinical investigation in cancer therapy have led to the elucidation of the many principles of proper scientific experimentation. Emphasis has been placed, and rightly so, on careful study planning, clarity in the statement of study objectives, thoughtful case selection, meticulous treatment methodology, and the efficient use and correct interpretation of study results.

Work conducted as an activity of the Eastern Cooperative Oncology Group.

Supported by NIH Grants CA 13650 and CA 12721.

* Mayo Clinic and Mayo Foundation, Rochester, Minnesota.

† State University of New York at Buffalo, Amherst, New York.

Address for reprints: Charles G. Moertel, MD, Mayo Clinic, Rochester, MN 55901.

Received for publication June 9, 1975.

There is, however, one essential element of this experimentation which is perhaps too frequently forgotten amidst such technical sophistry, namely, the actual measurement of the study end point. The culmination of most experimental therapeutic trials for solid tumors occurs when a man places a ruler or caliper over a lump and attempts to estimate its size. With this is introduced the inevitable factor of human error. Although the ultimate aim of therapy is increased survival, only few of our current approaches achieve that goal. To search for antitumor activity in a new modality by using survival as an endpoint is a far too complex and time-consuming effort and one that is frequently confused by the multiple therapies that may be attempted in any single patient.

Apart from some attempts to examine optimum methods of measuring lesions visible on roentgenograms,[1] this vital area of cancer treatment methodology has been largely ignored in oncology writings. Since at present we cannot eliminate the human error factor, it would seem advisable to make an effort to understand it, to learn how great an effect it can have on the validity of our results, and to consider how to defend against it.

This report deals with an investigation of the sources and magnitude of measurement variations which arise in evaluating anticancer activity of modalities used in the treatment

of solid tumors. The results emphasize the serious effects that such measurement errors can have on reported results, and suggest that a recognition of the limitation of measurement accuracy should be a major consideration in the formulation of criteria for clinical antitumor activity.

## MATERIALS AND METHODS

This study was designed as a simulation of the conditions when a clinician measures tumors under the circumstances usually encountered in oncologic practice.

Twelve solid spheres were selected, measuring from 1.8 to 14.5 cm in diameter. It was assumed that this size range would cover the sizes usually encountered in measurable clinical masses such as subcutaneous, lymph node, and intra-abdominal tumors. These masses were then arranged in random size order on a soft mattress and covered with a layer of foam rubber. This layer measured 0.5 in. in thickness for the six smaller masses to approximate skin and subcutaneous tissue and 1.5 in. for the six larger masses to approximate abdominal wall. Each of 16 experienced physicians practicing in oncology was then asked to measure the diameter of each sphere using the usual technique and equipment (ruler or caliper) he employed in clinical practice.

The actual "tumor" diameters are shown in Table 1. The participants were unaware that "tumors" 5 and 6 were designed to have the same diameter and so to provide an estimate of the reproducibility of each physician's measurements of tumor size. Tumors 7 and 8 were also designed for this purpose (the slight difference in true diameters 5 and 6 and in 7 and 8 reflect variations in the manufacturing process).

These "tumors" were palpably very similar to those encountered in clinical practice. The setting, however, was a very idealized representation of practice conditions where tumors tend to be less accessible, nonspherical, and of varying texture, and where the bearer of the tumor is not often as immobile and compliant as a mattress. The results which follow can thus be viewed as conservative estimates of measurement variations occurring under real-life clinical conditions.

TABLE 1. Tumor Diameters, Actual and as Measured (in Centimeters) by 16 Investigators

| | Tumor number and diameter (cm) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Investigator | 1.8 | 2.3 | 2.7 | 3.5 | 4.3 | 4.4 | 5.3 | 5.5 | 6.4 | 7.5 | 8.6 | 14.5 |
| 1 | 1.3 | 1.8 | 2.5 | 3.0 | 4.0 | 4.0 | 5.0 | 6.5 | 5.0 | 7.0 | 7.5 | 15.0 |
| 2 | 1.5 | 2.0 | 3.0 | 4.0 | 4.5 | 3.0 | 5.0 | 4.5 | 5.0 | 6.5 | 7.5 | 13.0 |
| 3 | 1.5 | 2.0 | 2.5 | 3.5 | 4.0 | 4.0 | 5.5 | 5.0 | 6.0 | 7.0 | 8.0 | 15.0 |
| 4 | 1.5 | 1.8 | 2.5 | 3.3 | 4.0 | 3.8 | 4.7 | 4.5 | 5.0 | 6.4 | 7.0 | 13.0 |
| 5 | 1.8 | 2.0 | 3.0 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 6.0 | 6.5 | 9.0 | 15.0 |
| 6 | 1.8 | 2.3 | 3.0 | 4.0 | 4.5 | 5.0 | 5.5 | 5.5 | 6.0 | 7.5 | 7.5 | 15.0 |
| 7 | 1.0 | 2.0 | 2.5 | 3.0 | 4.0 | 3.5 | 4.0 | 5.0 | 4.0 | 6.0 | 6.0 | 15.0 |
| 8 | 1.5 | 2.0 | 3.0 | 4.0 | 5.0 | 3.5 | 7.0 | 5.5 | 6.5 | 6.5 | 4.0 | 19.0 |
| 9 | 1.6 | 2.0 | 2.5 | 3.5 | 4.0 | 4.5 | 5.5 | 5.5 | 5.5 | 7.5 | 9.0 | 14.0 |
| 10 | 1.2 | 1.5 | 2.0 | 3.0 | 4.0 | 5.4 | 4.5 | 6.0 | 6.0 | 7.0 | 7.5 | 15.0 |
| 11 | 1.5 | 2.2 | 3.0 | 4.3 | 4.2 | 5.2 | 5.5 | 5.4 | 6.0 | 7.5 | 10.5 | 18.0 |
| 12 | 1.5 | 1.5 | 2.0 | 3.0 | 4.5 | 3.0 | 6.0 | 4.0 | 4.5 | 4.5 | 8.5 | 14.0 |
| 13 | 2.0 | 2.0 | 3.0 | 3.5 | 4.0 | 4.5 | 5.5 | 5.5 | 6.0 | 7.0 | 8.0 | 15.0 |
| 14 | 1.6 | 2.2 | 2.8 | 3.8 | 4.5 | 5.6 | 5.3 | 5.4 | 6.2 | 6.8 | 8.1 | 14.0 |
| 15 | 0.8 | 1.0 | 1.5 | 2.0 | 3.0 | 2.0 | 4.0 | 3.0 | 4.0 | 5.0 | 6.0 | 12.0 |
| 16 | 1.7 | 1.8 | 2.2 | 3.5 | 3.8 | 3.8 | 4.8 | 4.6 | 5.0 | 6.0 | 6.0 | 13.0 |
| Group mean | 1.49 | 1.88 | 2.56 | 3.46 | 4.16 | 4.06 | 5.21 | 5.12 | 5.42 | 6.61 | 7.5 | 14.69 |
| Group bias | −.31 | −.42 | −.14 | −.04 | −.14 | −.34 | −.09 | −.38 | −.98 | −.89 | −1.09 | .19 |

## RESULTS

### Group and Individual Measurement Bias

In Table 1, the sphere diameters recorded by each of the 16 investigators are displayed along with the actual diameters of the 12 spheres in question. The group, on the average, underestimated the diameter of 11 of the 12 tumors. This negative group bias, ranged from 1% to 18% of the actual diameter. Individual biases are also apparent, as some investigators, notably 12 and 15, consistently measured below the group means.

It is also evident that measurements are not recorded on a strictly continuous scale, but are rounded to the nearest ½ cm in a large number of instances, with (apparently) many investigators rounding to the nearest centimeter. This feature of itself may result in erroneous reporting if different investigators, using different rounding conventions, estimate tumor size in serial evaluations.

### Precision of Repeated Measurements by the Same Investigator

Each investigator's recorded measurements of tumors 5 and 6 allow the reproducibility of his estimates to be assessed since these two "tumors" are essentially the same. Based on the 16 differences (the difference between column 5 and column 6 in each of the 16 rows of Table 1), it is seen that the "within investigator" variance is $0.61^2$, which represents a coefficient of variation (standard deviation as a fraction of the mean diameter) of 14%. The corresponding measures for the repetition in measurements of tumors 7 and 8 are $0.68^2$ and 16%.

### Precision of Repeated Measurements by Different Investigators

The differences between the 16 entries in any column of Table 1 represent two sources of variation: (a) the positive or negative bias or error introduced by different investigators, and (b) the possible fluctuations which arise when any one investigator repeats a measurement, the magnitude of which have been referred to above. In columns 5 and 6, these two sources can be separated; the bias or error introduced when an investigator (chosen at random) begins to measure a "tumor" is on the average approximately 9%.

When this bias is compounded with the investigator's error on repeated measurements, the average variation between the tumor diameters recorded by two different investigators is estimated to be $(9^2 + 12^2)^{1/2} = 17\%$ of the actual diameter.

These estimates of this compound error, based solely on the 32 observations on tumors 5 and 6 are generally confirmed by the actual variation noted in tumors 1–4 and 9–12, where the two sources of variation cannot be separated. It is of interest that the variations among the 16 measurements for each of the

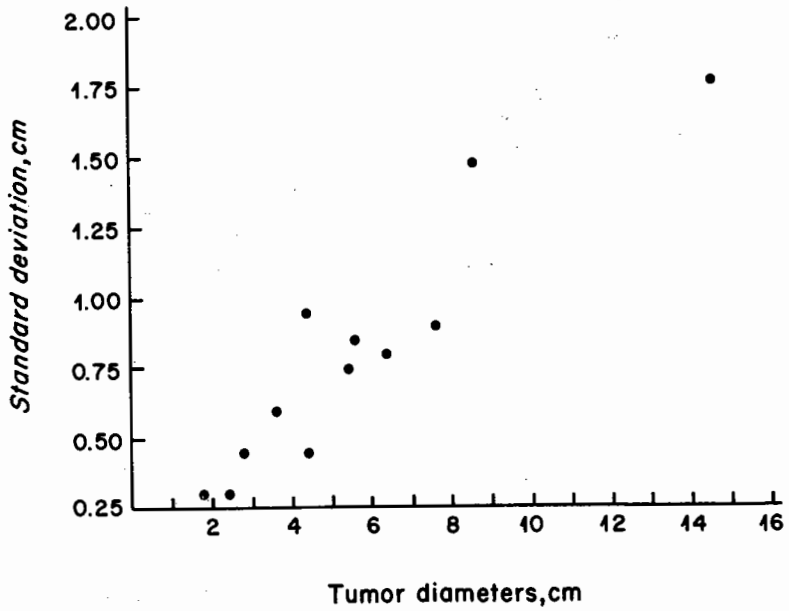


FIG. 1. Variations in tumor measurements as related to tumor size. The vertical axis represents the standard deviation of the 16 observations in each column of Table 1.

12 tumors remain on the average at approximately 20% of the tumor diameter regardless of the size of the diameter in question. This "error proportional to size" is evident in Fig. 1.

### Tumor Measurements Judged in a Clinical Setting

The effect of the sizable measurement variations are best appreciated if viewed in the context of actual clinical study conditions, using the usual clinical parameters and the commonly held criteria for tumor "regression."

Since the tumor is three-dimensional, ideally three dimensions should be measured and results expressed in terms of total tumor volume. This, however, is seldom possible because, although it is feasible to measure length and width, the parameter of depth is not accessible. A possible compromise would be to express just a single linear measurement, perhaps the longest diameter, a sum of linear measurements, or the sum of the longest perpendicular diameters. In measuring pulmonary metastasis, where lesions are usually spherical, Gurland and Johnson[1] have shown that a simple measurement of greatest diameter will suffice. Such linear measurements, however, would frequently present distortions for tumor masses in other sites which are usually of eccentric shape and may show eccentric regression in response to therapy. For example, a perineal recurrence of a rectal carcinoma will frequently present the outline of a long, narrow ellipse. Assuming such a mass originally measured 10 × 3 cm in longest perpendicular diameters and reduced after therapy to 9 × 1 cm, the longest single diameter would have shown only a 10% regression and the sum of perpendicular diameters, only a 23% regression—neither of these figures would be classified as objective responses in most studies although the actual tumor mass had reduced substantially. It is now most generally accepted that a tumor measurement should be expressed in "area" calculated as a product of the longest perpendicular diameters. For the example above, a much more appropriate 70% regression would be recorded, a very respectable objective response. For the remainder of this analysis, this parameter of "area" will be employed and the effect of measuring error will be assessed for studies that would employ either a 25% or a 50% reduction in such area as an indicator of an objective response.

By considering the 64 recordings of the diameter of "tumors" 5–8, it is possible to examine the rate of false reporting when the *same* investigator repeats an observation on a tumor which has remained the *same* size (Table 2). In the 64 possible reporting situations (16 where the investigator measured tumor 6 after tumor 5, 16 in the reverse case, together with the corresponding 32 for tumors 7 and 8), the measurement errors would lead to a 19% falsely reported "tumor shrinkage by at least 25%" or to 8% false-positives if the more stringent "50% reduction in tumor cross-sectional area" was employed.

If one regarded a tumor which grew by more than 25% in area as disease "progression," then Table 2 shows that in 12 (19%) of the 64 instances, a repeat measurement on a tumor which remained the same size would have been reported—incorrectly—as evidence of disease progression. Even if one demanded that in order to be termed a "progression," a tumor should increase by more than 50% in baseline area, then the number of false-negatives would not be reduced.

Similarly, one can evaluate the error rates if, instead, the *same* size tumor is measured by two *different* investigators. In this situation, there is a total of 1920 configurations—240 investigator-pairs measuring each of 8 tumor pairs—as is shown in Table 3. If a regression of "at least 25%" were required, the percentage of false-positives would be 24.8%; if "at least 50% shrinkage" was stipulated, this error rate would fall to 6.8%.

In reporting tumor growth, the false-negative rate, again with different investigators, would be 31.5% if the criterion was "any growth of 25% or more in tumor area" or 17.8% if only those increases of at least 50% were to be considered disease progression.

Based on the results above, it would seem that a 25% regression in tumor area is inadequate to ensure a reasonable degree of accuracy or specificity. Even if a 50% shrinkage is stipulated, variations in sequential measurements by the same or by another investigator will sometimes result in unjustified claims of tumor response. It is also clear that the error in measurement would also cause premature termination of a drug because of supposed treatment failure. A measured 50% increase in tumor area cannot always be regarded as real.

The sensitivity (i.e., the degree to which tumor changes will be correctly detected), of

TABLE 2. Erroneously Declared Objective Response and Objective Progressions When the Same Investigator Repeats a Measurement on a Tumor Which Has Remained the Same Size

| | | Erroneous classifications if B is baseline tumor | | | | Tumor "area" measurements cm² | | Erroneously classified responses if A is baseline tumor | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Invest. | ≥50% growth | ≥25% growth | ≥25% shrink. | ≥50% shrink. | Tumor A | Tumor B | ≥50% shrink. | ≥25% shrink. | ≥25% growth | ≥50% growth |
| Tumors | 1 | | | | | 16.0 | 16.0 | | | | |
| 5 and 6 | 2 | + | + | | | 20.2 | 9.0 | + | + | | |
| | 3 | | | | | 16.0 | 16.0 | | | | |
| | 4 | | | | | 16.0 | 14.4 | | | | |
| | 5 | | | | | 20.2 | 25.0 | | | | |
| | 6 | | | | | 20.2 | 25.0 | | | | |
| | 7 | | | | | 16.0 | 12.2 | | | | |
| | 8 | + | + | | | 25.0 | 12.2 | + | + | | |
| | 9 | | | | | 16.0 | 20.2 | | | | |
| | 10 | | | | | 16.0 | 20.2 | | | | |
| | 11 | | | + | | 17.6 | 27.0 | | | + | + |
| | 12 | + | + | | | 20.2 | 9.0 | + | + | | |
| | 13 | | | | | 16.0 | 20.2 | | | | |
| | 14 | | | + | | 20.2 | 31.4 | | | + | + |
| | 15 | + | + | | | 9.0 | 4.0 | + | + | | |
| | 16 | | | | | 14.4 | 14.4 | | | | |
| Tumors | 1 | | | + | | 25.0 | 42.2 | | | + | + |
| 7 and 8 | 2 | | | | | 25.0 | 20.2 | | | | |
| | 3 | | | | | 30.2 | 25.0 | | | | |
| | 4 | | | | | 22.1 | 20.2 | | | | |
| | 5 | | | | | 30.2 | 36.0 | | | | |
| | 6 | | | | | 30.2 | 30.2 | | | | |
| | 7 | | | + | | 16.0 | 25.0 | | | + | + |
| | 8 | + | + | | | 49.0 | 30.2 | | + | | |
| | 9 | | | | | 30.2 | 30.2 | | | | |
| | 10 | | | + | | 20.2 | 36.0 | | | + | + |
| | 11 | | | | | 30.2 | 29.2 | | | | |
| | 12 | + | + | | | 36.0 | 16.0 | + | + | | |
| | 13 | | | | | 30.2 | 30.2 | | | | |
| | 14 | | | | | 28.1 | 29.2 | | | | |
| | 15 | + | + | | | 16.0 | 9.0 | | + | | |
| | 16 | | | | | 23.0 | 21.2 | | | | |

| Baseline tumor | Tumor measured subsequently | No. of same-investigator evaluations | No. of investigators who reported objective responses | | No. of pairings who report objective progression | |
|---|---|---|---|---|---|---|
| | | | ≥ 25% shrinkage | ≥ 50% shrinkage | ≥ 25% growth | ≥ 50% growth |
| #5 | #6 | 16 | 4 | 4 | 2 | 2 |
| #6 | #5 | 16 | 2 | 0 | 4 | 4 |
| #7 | #8 | 16 | 3 | 1 | 3 | 3 |
| #8 | #7 | 16 | 3 | 0 | 3 | 3 |
| TOTAL | | 64 | 12 (18.8%) | 5 (7.8%) | 12 (18.8%) | 12 (18.8%) |

the commonly used criterion "change in tumor area" can also be assessed. The data in Table 1 represent measurements over a wide range of tumor shrinkage/growth. Conceivably, the evaluation of a tumor at any two time points might be done by any one of the 16 × 16 investigator pairings. It is also conceivable that the tumor in question could have changed from any one of the 12 tumor sizes used in the simulation to any other of the 12 tumor sizes (including the cases, already discussed above where the tumor did not really change). In all therefore, the data could be viewed as 256 × 144 = 36,864 possible evaluations—256 investigator-pairs evaluating 144 tumor changes. As a particular example, consider the case where a tumor changed from an initial area of 4.3 × 4.3 = 18.5 cm² to 2.7 × 2.7 = 7.3 cm². In terms of "area" measurement, this tumor should be classified as

TABLE 3. Erroneously Declared Objective Responses and Objective Progressions When Two *Different* Investigators Measure Tumor Which Has Remained the Same Size

| Baseline tumor | Tumor measured subsequently | Number of different investigator pairings | No. of pairings who report objective responses ≥ 25% shrinkage | ≥ 50% shrinkage | No. of pairings who report objective progression ≥ 25% growth | ≥ 50% growth |
|---|---|---|---|---|---|---|
| #5 | #5 | 240 | 29 | 6 | 64 | 23 |
| #5 | #6 | 240 | 70 | 26 | 83 | 48 |
| #6 | #5 | 240 | 60 | 8 | 94 | 58 |
| #6 | #6 | 240 | 83 | 39 | 95 | 70 |
| #7 | #7 | 240 | 57 | 7 | 67 | 37 |
| #7 | #8 | 240 | 64 | 18 | 67 | 31 |
| #8 | #7 | 240 | 51 | 7 | 66 | 38 |
| #8 | #8 | 240 | 65 | 19 | 68 | 37 |
| TOTAL | | 1920 | 479 (24.9%) | 130 (6.8%) | 604 (31.5%) | 342 (17.8%) |

"greater than 50% regression." The variations among the 256 possible evaluations in this situation are such, however that only 188/256 = 73% of the evaluations would correctly identify this tumor "response"; in the remaining 68/256 = 23%, the reduction would have been incorrectly called "no change" (16%) or "more than 50% growth" (9%). Incidentally, this 23% nondetection is due more to discrepancies between different investigators (66/240 = 27.5%) than to discrepancies between any one investigator's serial measurements (2/16 = 12.5%).

Consider first the set of six smaller tumors. Among these, the largest shrinkage would occur if a tumor changed from an area of $4.4 \times 4.4 = 19.36$ cm$^2$ to one of size $1.8 \times 1.8 = 3.24$ cm$^2$—a shrinkage to approximately 17% of its original area. Conversely, if a tumor began with an area of 3.24 cm$^2$ and grew to one of 19.36 cm$^2$, the growth would represent 595% of its original size. The frequency with which these and all intermediate changes would be correctly and incorrectly reported is summarized in Fig. 2a; the corresponding frequencies for the six larger tumors are shown in Fig. 2b.

Three points arise from Fig. 2A and B. First, as expected, as tumor changes become sizable, the number of falsely reported responses diminishes. When the actual tumor area was reduced by 75% or more, or increased by 25% or more, the rate of erroneous classification is less than 5%. Secondly, if the same investigator repeats a measurement on a tumor which has changed in size, he is more likely to correctly identify the change correctly than if a colleague performs the second measurement. This is in contrast to the approximately equal error rates which are obtained when either an investigator-pair or the same investigator measure a tumor which has not changed in size. Third, the error rates are approximately the same for the sets of larger and smaller tumors.

## DISCUSSION

From a practical clinical standpoint the primary concern in evaluating a new treatment modality is the rate at which a "response" would be recorded when a tumor mass had either remained unchanged or had actually increased in size—in essence the placebo response rate. In this study, the "placebo response rate" in unchanged masses as measured by the same investigator was 25% if a 25% regression criterion was employed and 7% if a 50% criterion was employed. When, however, the oncologist is measuring a deep-seated irregular mass through the muscle guarding of an uncomfortable patient, the errors of his measurement will undoubtedly be far greater than that recorded under the structure of this study. Certainly under real life circumstances, a 25% regression must be considered meaningless, and even a 50% regression is questionable. A third dimension, however, may be brought into clinical evaluations that can substantially increase the accuracy of response determination. This is the dimension of time. The natural course of a malignant solid tumor is to grow. This rate of growth will be variable, and in a review of published human data, Steel[2] found the volume doubling time for some solid tumors

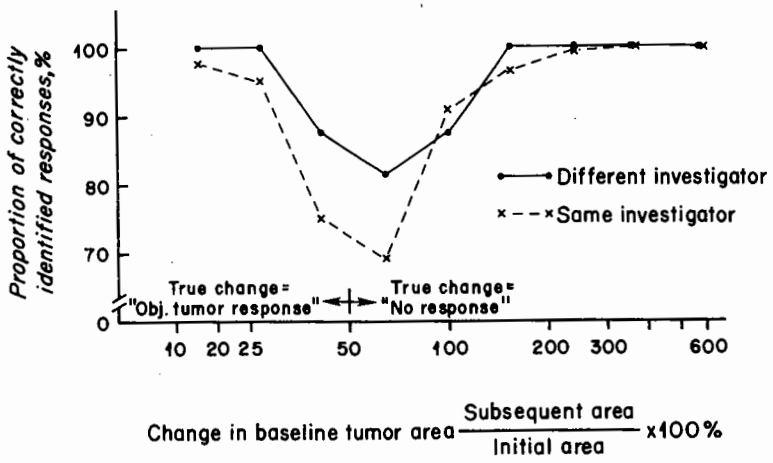


FIG. 2A. Frequency of correctly reported tumor changes (smaller tumors 1–6 only).

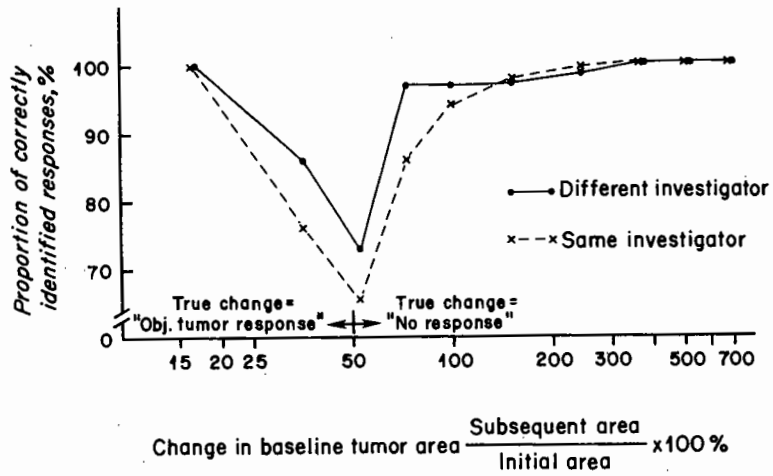


FIG. 2B. Frequency of correctly reported tumor changes (larger tumors 7–12 only).

to be as short as a week and for others over a year. A reasonable median for tumor-volume doubling time from his review could be estimated as approximately 60 days. At two months the area measurement would, therefore, have increased by 40%. A 50% regression of initial size observed at 2 months is actually a 64% regression of the size the untreated tumor would have become at that time. This natural assist to accuracy in response determination serves as counterbalance to the increased difficulties in measurement presented by true clinical circumstances.

Based on the observations of this study, it would seem reasonable to employ the following criteria for declaring an objective response:

1. A reduction in the product of the longest perpendicular diameters of the most clearly measurable tumor mass by at least 50%.

2. No increase by greater than 50% in other areas of known malignant disease.

3. No new areas of malignant disease may appear.

4. This result must be observed at least 2 months after the onset of therapy.

Utilizing these criteria an erroneous recording of objective response must be anticipated for probably between 5 and 10% of patients treated. If a criterion of regression is less than 50%, or a shorter time period is employed, a still greater "placebo" response rate must be projected.

## REFERENCES

1. Gurland, J., and Johnson, R. O.: Case for using only maximum diameter in measuring tumors. *Cancer Chemother. Rep.* 50:119–124, 1966.

2. Steel, G. G.: Cytokinetics of neoplasia. *In* Cancer Medicine, J. F. Holland, and E. Frei, Eds. Philadelphia, Lea and Febiger, 1973; pp. 125–140.