The Development and Evaluation Of Cancer Diagnostic Tests

By JOHN E. DUNN, Jr., M.D., and SAMUEL W. GREENHOUSE

ANY ATTEMPT to bring the majority of human cancer cases to clinical recognition in a curative stage, at least until new therapeutic methods are established, involves the ability to recognize the disease in an asymptomatic individual. Searching the general population for unsuspected cancer using clinical procedures has been explored through cancer detection centers, but has been found to be impractical because of the inadequate capacity of such facilities and the relatively high cost per examinee.

The need for a procedure to indicate the existence of unsuspected cancer has led to many attempts over the years to devise a laboratory procedure that would show whether or not an individual is harboring a cancer. These procedures have usually taken the form of chemical, biological, physical, or immunological measurements on readily available human materials such as blood, urine, and exudates, or of skin tests.

A priori it can be said that the possibility of developing such a diagnostic test for a disease is dependent on unique and specific substances produced by or as a result of the disease, which can be measured by laboratory procedures; or by quantitative changes in normal body con-

Dr. Dunn directed the field investigations program of the control branch, National Cancer Institute, National Institutes of Health, Public Health Service, from 1948 until his present assignment to the California Department of Public Health at Berkeley. Mr. Greenhouse is a statistician in the biometrics section of the technical services branch, National Cancer Institute.

Public Health MONOGRAPH

No. 12

The accompanying article discusses the principal findings presented in Public Health Monograph No. 12, published concurrently with this issue of Public Health Reports. These papers, by authors from the Medical College of Alabama, Tufts College Medical School, the Schools of Medicine of the Universities of Washington, Tennessee, and Kansas, and the Research Laboratory of the Jewish Memorial Hospital, Roxbury, Mass., were assembled by the National Cancer Institute, National Institutes of Health, Public Health Service.

Readers wishing the data in full may purchase copies of the monograph from the Superintendent of Documents, United States Government Printing Office, Washington 25, D. C. A limited number of free copies are available to official agencies and others directly concerned on specific request to the Public Inquiries Branch of the Public Health Service. Copies will be found also in the libraries of professional schools and the major universities, and in selected public libraries.

• • •

Evaluation of cancer diagnostic tests. Public Health Monograph No. 12 (Public Health Service Publication No. 275). U. S. Government Printing Office, Washington, 1953. Price 30 cents.

stituents that are more or less uniquely associated with a specific disease. Various immunological tests for acute infectious diseases are classical examples of the former, and the glycosuria associated with diabetes mellitus, of the latter. Unfortunately, cancer research has not as yet demonstrated well-established qualitative changes, either in cancer tissue as such, or in the host organism supporting the cancerous growth. Quantitative changes are known to occur in cancer tissue as compared to the corresponding normal tissue. Also, there are quantitative changes in the host, but these changes are not uniquely associated with cancer. The question, then, becomes one of whether a proposed diagnostic procedure based on empirical observation is perhaps founded on a mechanism that is yet unknown or not fully elucidated in the mass of cancer research knowledge, or whether the known quantitative host changes, singly or in combination, might not serve as a means of distinguishing cancerous from normal individuals and from those with other diseases.

In general, attempts to find a diagnostic test for cancer have been met with an attitude of pessimism since the body of cancer research knowledge has apparently not yet established a firm basis for development of such a test. On the other hand, those faced with the urgent demand that something be accomplished now to reduce human cancer mortality are confronted with the necessity of taking calculated risks.

In 1948, several university groups indicated an interest in exploring proposed cancer diagnostic tests to determine their usefulness by requesting grant funds from the National Cancer Institute. These projects were approved by the National Advisory Cancer Council and a loosely coordinated program was developed in which the five university groups looked to the cancer control branch of the National Cancer Institute for liaison and technical assistance in the analysis of data. The five groups carrying on this work are under the direction of:

Dr. Stuart W. Lippincott, professor of pathology, University of Washington Medical School, Seattle, Wash.

Dr. F. Homburger, director, cancer research and cancer control units, department of surgery, Tufts College Medical School, Boston, Mass.

Dr. J. K. Cline, chief, cancer research department, Medical College of Alabama, University of Alabama, Birmingham, Ala.

Dr. Douglas H. Sprunt, director, Institute of Pathology, University of Tennessee Medical School, Memphis, Tenn.

Dr. Robert E. Stowell, professor of pathology and oncology, University of Kansas Medical Center, School of Medicine, Kansas City, Kans.

The principal aims of this program were:

1. To determine whether any of the many diagnostic tests for cancer proposed in the past meet the original claims made for them by their developers when carefully evaluated by another laboratory.

2. To follow up any leads in basic biological, chemical, or biochemical research bearing on the diagnostic problem and possibly leading to the development of a test.

Additional purposes served by this program, purposes that were made apparent only after the program had been under way, were:

1. To provide a much needed statistical methodology in order to unify and make comparable different evaluations of the same test and also evaluations of different tests.

2. To utilize the experience and facilities of the participating groups to evaluate tests developed currently.

3. To obtain leads meriting further investigation resulting from the analysis of data collected by the various groups.

The purpose of this paper is to describe briefly the accomplishments of this program in fulfilling these aims. Much of what follows has already been said or reported elsewhere; the remainder is new.

Methodology

In order to evaluate the practical usefulness of cancer diagnostic tests, the performance requirements for a useful test must first be decided upon. Since the primary objective was a test that would indicate the probability of unsuspected cancer, the requirements of a case-finding or general population screening test were given primary consideration. Criteria were proposed and a statistical method for analyzing data on the basis of these criteria was developed (*1*). It was realized that a test would have other uses as well, such as for differential diagnosis in a diseased individual. Evaluation studies, therefore, included patients with

diseases other than cancer. However, it appeared reasonable that a test that would not distinguish satisfactorily between individuals apparently free of disease and those with cancer could not distinguish the latter from other diseased individuals.

Most of the general tests proposed to detect the presence of cancer are blood tests, based on the principle that some factor, for example, the blood proteins, enzymes, or an immunological agent capable of reacting with an antigen, appear in the blood serums of cancerous individuals and either are lacking or are quantitatively different from that in the blood serums of normal individuals. Measurement of this factor in a group of normal individuals and in a group of individuals with known cancer, in most cases, gives rise to a continuous variable more or less symmetrically distributed about a modal value which differs in the two groups. To make this process a diagnostic procedure it is necessary to select one value of the variable, the so-called critical value, to serve as the dividing point for future tests. If a person's measurement, say, exceeds the critical point, he would be classified as positive; if it falls below, he would be called negative. (In the past we have stressed the fact that no single test can do more than result in these designations and only provides evidence that a person called positive has some probability of having the disease. It is only after a person so classified has undergone clinical study that cancer can be diagnosed.) Once the critical point has been selected for any set of data, it becomes possible to refer to two measures inherent in the test, namely, the proportion of false negatives (sensitivity) and the proportion of false positives (specificity). These measures are completely dependent upon the choice of critical value and, in fact, vary as the critical point varies. We illustrate this in the following tabulation, based on an evaluation of the Huggins iodoacetate index by Homburger and associates (*2*).

Critical value:	*Percent false positives*	*Percent false negatives*
5.3	5	55
5.9	10	43
6.8	20	24
7.8	40	17
8.6	60	9

The implications in analyzing evaluations of diagnostic tests are clear. Contrary to past experience, an evaluation must not adhere blindly to the same critical value reported by the originator. The investigator must find the critical value that will give either the same specificity or sensitivity as that obtained by the originator and then compare the remaining measure. For example, if, from the previous table, one were to advocate the Huggins iodoacetate index because it gave 24 percent false negatives for 20 percent false positives, then the purpose of an evaluation is to confirm that this procedure gets 24 percent false negatives for 20 percent false positives. The investigator evaluating the test collects his own data and attempts to reproduce the biological and chemical aspects of the procedure as carefully as he can. Often biases exist but, even if they did not, characteristics of the distribution, such as the critical point giving 20 percent false positives, are subject to sampling variation. (Our experience has been that consciously or unconsciously the investigator makes some modification in techniques and that sampling variation is small compared to these biases.) Thus, if the investigator were to seek out the value 6.8 as his critical point because the originator used this value, he might find 35 percent false negatives and 15 percent false positives or any other proportions. On this basis he would conclude the original report has not been verified. What, in fact, he should do is find the critical value giving him, say, 20 percent false positives and then determine the percent of false negatives for this critical point. If this turns out close to 24 percent or less, the evaluation confirms the original report; if it is considerably higher, then the investigator rejects this test based on the way he performed it.

The two essential points in an evaluation program are, therefore:

1. Two sensitivity measures must be compared where each is obtained for a fixed specificity (or the converse).

2. The critical value to be found in determining the sensitivity is itself determined by the given specificity.

Costs of examination and incidence of cancer are such that it was reasoned no screening program for cancer could tolerate more than 10

percent false negatives for at most 5 percent false positives. These criteria need not be followed if one searches for a diagnostic test to be used for other purposes, such as differential diagnosis.

Evaluation of Tests

A considerable number of reports have thus far been published as a result of the cancer diagnostic test evaluation program. A list of these reports, classified according to type of substance being measured, appears as an appendix to the most recent publication, a monographic collection appearing as Evaluation of Cancer Diagnostic Tests, Public Health Monograph No. 12. In addition to tests on which reports have been published, several tests were evaluated by the various groups at the request of the National Cancer Institute and other institutions. These represented tests being developed currently and for which no large-scale evaluation was necessary in order to reject them. On the other hand, several tests announced in the last 5 years were evaluated fully and reports were published.

Unfortunately, as reference to these publications will show, none of the procedures evaluated has been judged capable of discriminating between individuals with cancer and those without cancer to any reasonably high degree. For a cost of 5 percent false positives among presumably normal, healthy individuals, these tests, as evaluated, detected as positive from 10 percent to about 75 percent of known cancer patients, with the majority ranging from 40 percent to 60 percent. For the most part, these tests also found as positive from 25 percent to 50 percent of patients with diseases other than cancer. But more serious from the point of view of screening is the fact that these tests gave rather poor results among known cancer patients with well-established disease. Presumably, if groups of individuals with very early cancer were available, these tests would detect as positive still smaller proportions.

Evaluation and Developmental Findings

Although results have been negative in the search for a general test for cancer, all of the participating groups are continuing in some developmental field of their own. In some cases, investigation is being made into the diagnosis of cancers of specific sites; in others, work is being continued on those general tests which a group thought promising. Every group is doing research into the development of its own procedure, both on the laboratory and clinical level.

All participants have concluded from their evaluations thus far that much has yet to be learned about the relationships to the cancer process of those factors which these tests purport to measure. The awareness and the need of a greater understanding of the effects of this process on the biochemistry of the individual are evident in the report on the Proceedings of the First Conference on Cancer Diagnostic Tests (*3*). The very purpose of this conference, sponsored by the National Cancer Institute, was that ". . . further developmental research in the cancer diagnostic test field should be stimulated."

It would appear that one reason these tests have failed is the lack of specificity in the factors assumed to be changed by cancer. Generally, these factors seem to be affected by many disease processes. In fact, they are found to vary among normal individuals. This raises some interesting questions concerning the concept of a diagnostic test (see paper by Toennies in reference 3). Given that normal individuals really differ with respect to a given factor, does the single individual's test value vary with respect to time or does it remain relatively stable? If it does change with time is this variation random around some true value and, if so, how does it compare with the variation among individuals? If it is relatively small then obviously if an individual's test value begins to increase progressively over time, he should be suspect even if his test values are not above the critical point (assuming cancer values are on the average larger than normal values). However, before such serial testing on an individual basis can be of use, much data must be gathered to answer the above questions on variation.

A start in this direction was made by one laboratory, which was able to obtain more than one blood specimen on normal persons over a

period of a year. We illustrate some of these ideas referring to the evaluation, by this laboratory, of the least coagulable protein test proposed by Huggins (*4*). With respect to this test, the values of 137 normal individuals ranged from 1.10 to 1.91 with a variance, $\sigma^2=0.0139$. The variation among individuals, measured by the variance, can be considered to be made up of three components: variation among true individual values, σ^2_{Ind}, variation between specimens from the same individual when specimens are taken over a period of time, σ^2_{sp}, and variation due to the reproducibility, or measurement error, of the technique, σ^2_m. Estimates of these components were as follows: $\sigma^2_{Ind}=0.0075$, $\sigma^2_{sp}=0.0022$, $\sigma^2_m=0.0035$. Variation due to specimens, σ^2_{sp}, represents about 16 percent of the observed variation among individuals and about 30 percent of the estimated variation of true individual test values. Consider an individual with a true value of 1.3. Assuming no improvement in technique can be accomplished to reduce measurement error, 95 percent of specimens from this person should result in values ranging from

$$1.3-2\sqrt{\sigma^2_{sp}+\sigma^2_m} \text{ to } 1.3+2\sqrt{\sigma^2_{sp}+\sigma^2_m}$$

or 1.15 to 1.45. Now, if this person gets cancer, his true value should begin to increase and hence his test values should eventually fall outside his normal range. When this occurs, his test should be considered positive even though no value is greater than the critical point (in this case, 1.63).

Considerations of this sort have come out of the analysis of the data gathered in the evaluation program. For two other tests, σ^2_{sp} represented a much greater proportion of the total variation. In fact, for one test it was almost 50 percent of the total variation and exactly equal to the estimated variation of true test values. In these instances, no serial testing on an individual basis would be meaningful.

Present and Future Developments

As indicated earlier, the various laboratories that have been engaged in diagnostic test evaluation have continued investigating certain procedures that still appear to hold some promise, and are exploring developmental possibilities that have attracted their interest. The former includes further work with a serum flocculation reaction that has undergone additional development since originally reported; exploration of fluorescence phenomena observed in the blood from cancer patients; polysaccharides of serum that are augmented in cancer patients; and use of several serum protein determinations in combination. Developmental investigations by these groups include investigations into a sensitive means of detecting abnormal steroid in the blood or urine; a complement fixation test; a study of the factor responsible for liver catalase reduction in cancer; and a specific measurement of prostatic acid phosphotase. This last has been developed to the point where several laboratories are evaluating it as a means of diagnosing premetastatic prostatic cancer.

REFERENCES

(*1*) Dunn, J. E., Jr., and Greenhouse, S. W.: Cancer diagnostic tests. Principles and criteria for development and evaluation. Public Health Service Publication No. 9. Washington, D. C., U. S. Government Printing Office, 1951.

(*2*) Homburger, F., Pfeiffer, P. H., Page, O., Rizzone, G. P., and Benotti, J.: Evaluation of diagnostic tests for cancer. III. Inhibition of thermal coagulation of serum for iodoacetic acid (the Huggins-Miller Jensen test). Cancer **3**: 15–25 (1950).

(*3*) Proceedings of the First Conference on Cancer Diagnostic Tests, 1950. Public Health Service Publication No. 96. Washington, D. C., U. S. Government Printing Office, 1951.

(*4*) Ellerbrook, L. D., Meek, E. C., and Lippincott, S. W.: Tests for the least coagulable serum protein and the iodoacetate index. J. Nat. Cancer Inst. **12**: 49–89 (1951).