The Usefulness of Laboratory Tests in Health-Screening Programs

David L. Sackett

I discuss pitfalls in laboratory-screening programs: regression toward the mean on repeated biochemical determinations; the problem of defining normalcy in the interpretation of laboratory test results; and a remarkable professional myopia in which clinical chemists have, with rare exception, failed to accept responsibility for evaluating whether the programs in which they are engaged are of benefit to patients.

Additional Keyphrases: assessment of multiphasic testing • normal ranges • regression toward the mean • applicability of standard deviation • clinical interpretation of laboratory results

The last two decades have seen a rapid expansion of multiphasic screening and periodic health examination programs directed toward the earliest possible detection of diabetes, various cancers, hypertension, glaucoma, and other chronic diseases (1-3). Such programs are enormously attractive to clinical chemists, other health professionals, and the general public as these groups join forces to reduce the awesome toll of disability and untimely death resulting from these disorders. The basis for this commitment is the quite reasonable assumption that early diagnosis, particularly when it is achieved before symptoms appear, will result in a more favorable disease course and outcome than if one waited until a patient with symptoms sought health care. While this assumption seems valid for many acute and chronic infectious diseases, it does not appear to be so for the noninfectious chronic and degenerative diseases that constitute the major targets of most current programs of health screening by laboratory testing.

Experience with multiphasic screening and periodic health examination programs shows that they are generally well accepted by the public—although frequently not by those at highest risk of disease. The feasibility of examining quite large numbers of asymptomatic individuals rapidly and efficiently has also been demonstrated. Similarly, there appears to be little question that diagnoses can be made months—or in some cases years—before the development of symptoms prompts the patient to seek medical care. However, a preoccupation with the process of biochemical screening and periodic health examinations has obscured a consideration of the basic assumption that early diagnosis of noninfectious disease leads to an improvement in outcome.

The purpose of this presentation is to discuss three of the several issues that must be confronted by clinical chemists engaged in health screening programs. These pitfalls in laboratory-screening programs consist of:

- Regression toward the mean on repeated biochemical determinations.
- The problem of defining normalcy in the interpretation of laboratory test results.
- A remarkable professional myopia in which clinical chemists have, with rare exception, failed to accept responsibility for evaluating whether the programs in which they are engaged are of benefit to patients.

Regression toward the mean. The first pitfall is concerned with the biologic phenomenon known as "regression toward the mean" (4). It is best introduced through an example and, in order to avoid clouding the issue, I will draw an example that will be free of emotional overtones for an audience of clinical chemists, and consider the measurement of blood pressure.

In the programs that are currently being conducted in North America for the detection of hypertension in free-living populations, one observes a distribution of blood pressure values such as that shown in Figure 1. Like most anthropometric, physiologic, and biochemical attributes of man, blood pressure is neither normally nor symmetrically distributed, but tends to trail off toward one end of the scale of blood-pressure values. In this series of blood-pressure determinations, we can set an upper limit of normal, beyond which individuals are classified as abnormal or, in this case, hypertensive—and this is shown in
the shaded portion of this distribution. In other words, we have decided that these individuals are abnormal because their blood pressure exceeds some critical value.

The seeming paradox of regression toward the mean would be observed if we went back to this same total population of individuals and re-measured their blood pressure a few days later. Upon re-measurement, we would find once again that the blood pressure values would be distributed as shown in Figure 1, and that once again, a similar proportion of individuals would exceed this critical level and would be called hypertensive. To our surprise, however, many of the individuals classified as hypertensive on the first occasion would have blood pressures below the critical level at the time of re-measurement—and they would have been replaced in the hypertensive category by individuals whose blood pressure was within the normal range on the first measurement, yet rose into the hypertensive category at the time of the second recording of blood pressure.

What we are observing here are the combined effects of variation in measurement and the underlying biologic variation in blood pressure. While we tend to acknowledge the occurrence of biologic variation, the resulting regression toward the mean of outlying values is almost never referred to in the literature of clinical chemistry, and its effects are usually ignored.

Figure 2 shows the inevitable result. We once again see, in the upper portion of this Figure, the group of values that have, upon the initial determination, been found to lie beyond the upper limit of normal. Their average value is expressed as $\mu_1$. Now, instead of blood pressure, let us assume that we are talking about serum calcium concentrations, or the concentration of glucose in blood one hour after the ingestion of 75 g of glucose, or some other chemical determination.

In the lower portion of this Figure, we see what happens in this sub-group of individuals who are classified as abnormal on the initial determination, if the determination is repeated. It is noted that these extreme values have tended to re-distribute themselves at somewhat lower levels than before, and their new average value, $\mu_2$, is somewhat closer to the overall average for the total population. This is what we mean by "regression toward the mean."

Although this phenomenon is quite easy to demonstrate, its implications are almost universally ignored by clinical chemists. First, the clinical chemist is subjected to a considerable amount of unwarranted abuse by clinicians who, upon ordering repeat determinations for patients with outlying laboratory values, find that the repeated results have returned to values within the normal range. We clinicians tend to attribute this regression toward the mean as a result of sloppy clinical laboratory quality-control, and the clinical chemist is blamed for a perfectly natural phenomenon in human biology. While a greater understanding of regression toward the mean by those who use the clinical chemistry laboratory would be beneficial to the ego of the clinical chemist, there is perhaps a more important reason for recognizing this phenomenon from the point of view of the patient and the health-care system generally.

Suppose that an aggressive clinician, after the initial laboratory determination but before the repeat determination, institutes a therapeutic regimen designed to arrest or delay the deleterious effects of the abnormal finding. If he subsequently repeats the laboratory determination he will find, on the average, as we have just seen, that it is at a lower value than previously. As a result, if he does not understand the phenomenon of regression toward the mean, he runs the risk of assuming that he has somehow done the patient a great service, thereby committing the patient to a prolonged course of expensive therapy with its attendant risks of side-effects and the inevitable expense to the health-care system.

In summary, clinical chemists often find that individuals with extremely high or low initial values for a laboratory test tend to exhibit less extreme results when re-tested. The reason for this phenomenon of regression toward the mean is inherent in biologic variation, and the failure to recognize and compensate for this phenomenon may lead to a profound misuse of the chemistry laboratory. Not only may
patients be subjected to the risks and discomfort of unnecessarily prolonged treatment; regression toward the mean may also lead to the erroneous conclusion that the therapy being applied is effective. I believe that clinical chemists must play a much more aggressive role as consultants to clinicians who use the clinical laboratory, and that this consultant role must include educating the clinician about the phenomenon of regression toward the mean.

**Normal range.** The second general issue that I wish to discuss briefly relates to the definition of the normal range in laboratory test results. One of the most common approaches to setting the normal range for a clinical laboratory test is to describe it in terms of the mean value plus or minus two standard deviations. To paraphrase a former U. S. Supreme Court justice, the use of a statistical concept such as the standard deviation to set the limits of normal for clinical laboratory tests represents the cross-sterilization of disciplines, for it represents taking a misunderstood concept from sampling statistical theory and misapplying it in an individual clinical situation (5).

There is clearly no statistical justification for using the standard deviation to set the limits of normal and, as we shall see shortly, such an approach also lacks clinical validity.

In partial response to the criticism of the standard deviation as a rational approach to deciding normalcy, the range of normal has frequently been defined in terms of percentiles, and the range of normal values is increasingly referred to in terms of 5 and 95 percentiles. In addition to lacking either statistical or clinical justification, this latter approach leads to a dilemma: if the upper limit of normal is set by the 95th percentile, it inevitably follows that the proportion of patients who will be labeled as normal after the performance of some number of mutually independent laboratory tests is this proportion, 0.95, raised to the power of the number of laboratory tests performed. The implication of this relationship is shown in the following examples:

\[
\begin{align*}
\text{n (no. of lab tests)} & \quad \text{(0.95)^n (proportion of patients labeled "normal")} \\
1 & \quad .95 \\
5 & \quad .77 \\
20 & \quad .36 \\
100 & \quad .006
\end{align*}
\]

We see that as the number of laboratory tests increases, the likelihood that the patient will continue to be classified as normal rapidly diminishes (6). At the hospitals in which I practice, the average patient undergoes about 20 laboratory determinations, and we see that, by chance alone, he has only about one chance in three of being classified as normal for all of his results even if he is perfectly healthy! As clinicians have become more aware of this phenomenon, the attention paid to the unsolicited laboratory reports has substantially diminished. Indeed, a highly respected authority in North America has indicated that the clinician may wish to adopt a policy of ignoring unsolicited laboratory information (6).

When one views the clinical laboratory from the clinical point of view of therapeutic intervention, rather than that of diagnostic classification, the absurdity of setting of normal limits by use of percentiles or standard deviations becomes all the more apparent. To prescribe commonly accepted drugs to nonketotic adults with fasting blood sugar concentrations that are more than two standard deviations above the mean is to subject them to an increased risk of death from cardiovascular disease (7). By the same token, to withhold drugs for the treatment of diastolic blood pressures from all middle-aged males except those whose blood pressures are greater than two standard deviations above the mean is to withhold therapy from at least two-thirds of those individuals in whom randomized clinical trials have shown the clear benefit of drug therapy, in terms of the prevention of stroke and the prolongation of life (8). “Normal” values can be determined neither in the laboratory nor by the computer. They demand the close clinical study of the patient, his illness, and the natural history of treated and untreated disease. It is only through this latter approach that we can establish those laboratory values beyond which diagnosis and intervention should occur.

Indeed, it has been suggested, not at all facetiously, that the upper limit of normal is that value beyond which therapy begins to do more good than harm, and it is clear that, in these terms, the limits of normalcy have very little to do with standard deviations, percentiles, or sampling statistical theory (9). Once again, if progress is to be made in this area the clinical chemist must leave his laboratory and become an active consultant to, and collaborator with, his clinical colleagues.

**Evaluation of data.** The final topic I have selected for comment relates to the professional myopia that I believe continues to characterize the involvement of most clinical chemists in programs of health screening through the use of laboratory tests. As key participants in these screening programs, clinical chemists have, with rare exception, failed to take responsibility for evaluating whether these programs are of benefit to patients, are of no value, or are positively harmful. The points I wish to emphasize here can perhaps best be phrased in the form of five questions (2, 3).

(1) **Does laboratory screening detect important disease?**

This question asks whether, for example, laboratory screening and periodic health-examination programs have been shown to detect diseases serious enough to kill. If we look at the combined experience of several major periodic health-examination programs in North America, we see that their ability to detect lethal disease is incomplete. In reviewing these programs, it has been shown that fewer than
half of the program participants subsequently dying of cancer had their fatal disease diagnosed as a result of participation in these screening programs, and that these programs detected only 58% of subsequently fatal coronary heart disease. Indeed, the ante-mortem recognition of any and all diseases causing death was only 51% (10). If we could assume that we could totally arrest or cure these disorders once they were detected, the impact of such programs could be substantial. However, it is seen that periodic health-examination and laboratory-screening programs may clearly fail to detect as much as half of lethal diseases. Although one might argue that the program participants described here did not have their periodic health examinations at sufficiently frequent intervals, it should be pointed out that, for example, a randomized trial found that information derived from X-ray films of the chest, taken as often as every six months to detect presymptomatic lung cancer, had no detectable effect upon the mortality rate from this disease (11).

2. Will the treatment of risk factors such as abnormal laboratory values have a major impact upon the subsequent development of disease?

In the case of coronary heart disease, it has been repeatedly demonstrated that one’s risk of suffering, and dying from, a myocardial infarction is directly related to the concentration of one’s serum cholesterol, blood glucose, serum uric acid, and to a host of other factors. These statistical associations between predictors and subsequent disease have led to the formation of “anticoronary clubs” and to scientific and lay articles advocating nationwide programs of risk factor identification and modification in an effort to prevent or delay manifest coronary heart disease. Such programs presuppose that these statistical associations between risk factors and disease outcomes represent causal relationships. Yet, periodic health examination and laboratory screening programs based on these risk factors, or predictors, can be effective only to the extent that a significant portion of coronary heart disease can be blamed upon—or explained by—these predictors. However, it must be recognized that only a relatively small portion of coronary heart disease can be attributed to clinically abnormal values for blood glucose, serum cholesterol, and serum uric acid (12). The vast majority of victims of myocardial infarction are neither hypercholesteremic, hyperglycemic, nor hyperuricemic. Apart from this uncertainty over whether coronary risk factors actually play a part in causing the disease, it would appear that our enthusiasm for determining coronary proneness on a community-wide basis through the use of laboratory-screening programs deserves reappraisal in terms of the potentially meager prospects for any major impact through existing coronary-prevention programs.

3. What are the prospects that health behavior of our patients can be altered, once we identify an abnormal laboratory finding?

Even if a treatment program has been developed that will delay or prevent a disabling or lethal disorder, is there a reasonable expectation that individuals without symptoms will continuously take medications—or follow other health instructions—over long periods of time? You may wish to pause here in order to ponder your own success in modifying your cigarette consumption or food intake. Additional examples of the relatively low adherence to therapeutic regimens, even over quite short periods of time, are abundant. One oft-quoted study determined the extent to which 10-day courses of oral penicillin therapy given to prevent complications of streptococcal infection among children were actually completed (13). Fewer than half of the affected children were receiving the drug after three days, and fewer than one-fifth of the children completed the 10-day course of oral penicillin.

A second example is the Welsh study of a group of patients with high intra-ocular tension (14). They were instructed to the use of therapeutic eyedrops and, in addition to the physician’s usual encouragement, were visited at home regularly by nurses who reinforced the physician’s instructions and admonitions. This investigation into the effectiveness of anti-glaucoma therapy had to be abandoned because the patients who had not yet developed visual defects were simply unwilling to take their medication at regular intervals over long periods of time.

I suggest that clinical chemists, before they can advocate major programs of laboratory screening for presymptomatic disease, must have far greater assurance than is presently available that patients, particularly if they are asymptomatic, will modify their health behavior for prolonged periods of time.

4. Does action taken as a result of the screening process really alter the course of disease?

Up to this point we have merely skirted the central issue: To what extent can we point to clear-cut evidence that laboratory screening has indeed led to the prevention or postponement of disability and untimely death from chronic disease? One appropriate example for consideration here is the Papanicolaou smear for the early detection of cancer of the uterine cervix. It has long been assumed that this is an extremely worthwhile procedure, and it is widely believed that the universal application of the “Pap” smear will drastically diminish both incidence of and mortality from cervical cancer.

For a variety of reasons, properly controlled clinical trials to determine whether a Pap smear program really can lead to actions that decrease the incidence of invasive cervical cancer were not performed when the procedure was introduced more than 20 years ago. With the passing of time, the assumption of the Pap smear’s effectiveness has rendered such a trial unethical, and we are left with the mortality rate for this type of cancer as the only yardstick for measuring the effectiveness of this screening program.

A unique situation exists in Canada for determin-
ing whether the aggressive and province-wide use of the Pap smear has had an effect upon cervical cancer mortality rates. British Columbia embarked on a vigorous cervical cytology screening program more than 20 years ago, and more than half of the female population of that province had undergone at least one Pap smear by the year 1963. Investigations from the United Kingdom reviewed the published reports and available data and estimated that one might expect up to a 20% lower death rate from cancer of the cervix in British Columbia than in the rest of Canada—where no massive cervical cytology programs were in operation—by 1965. But when these investigators looked at the mortality data, they were unable to discern any differences in cervical cancer mortality rates between British Columbia, Ontario, and the rest of Canada (15). Indeed, a more recent investigation has extended this comparison through the year 1969, and all data reviewed thus far have failed to demonstrate any effect of a vigorous cervical cytologic program upon the death toll from cancer of the cervix.

An example of even greater pertinence to this group concerns the results of the randomized trial of laboratory screening and periodic health examinations performed by the Kaiser–Permanente group and reported at the American Public Health Association’s Annual Meeting in 1971 (16). Of the four age and sex groups involved in the trial, a statistically significant advantage in health status was observed only among older men and this difference was, in the view of clinicians in the audience, clinically trivial. Surely, clinical chemists must confront these emerging data concerning the lack of health effectiveness of laboratory-screening programs before advocating their widespread use.

(5) Are we misled by the traditional methods used in evaluating the clinical effectiveness of screening?

The foregoing findings and conclusions contrast with most that has been written and become generally accepted about the value of laboratory screening programs. Indeed, I anticipate that many of you will recall specific patients who, diagnosed in an early stage of their disease, enjoyed prolonged survival. I would counter by suggesting that although your clinical observations are correct, your conclusions concerning them reflect an incomplete understanding of the relationship between early diagnosis, the natural history of disease, and the measurement of survival. Indeed, it can be easily shown that the usual methods of measuring survival are quite unable to answer the question of whether the treatment instituted as the result of early diagnosis through periodic health examination is effective or worthless. The following two examples will underscore this deficiency.

In this first example, shown in Figure 3, let us consider a group of patients whose disease is diagnosed at a symptomatic stage, indicated by the Dx sign (2, 3). If one looked at survival among these individuals with symptoms as the starting point in a survival measurement, we see a rather steady decline in survivors with an overall five-year survival of roughly 50%. Thus, of a group of 45-year-old patients diagnosed through routine clinical practice, about half would be alive at age 50.

Let us now suppose that, using laboratory screening techniques, we would diagnose this disorder—among patients who do not yet have symptoms—on an average of 1 year earlier than the usual time of diagnosis among symptomatic patients. In comparing a group of patients who had the disease diagnosed through laboratory screening in this earlier stage with those patients who had been diagnosed through routine clinical practice, we are in danger of making the sort of mistake shown in Figure 4.

Although this example assumes that the therapy for the disorder is no more effective when applied early—the double line—than when it is applied at the time of usual diagnosis—the single line—we note in Figure 4 that the five-year survival among the early diagnosed group is substantially better than that of the former group who were not diagnosed until they had symptoms. This gain, however, is entirely fictitious; for all we have done is to shift the starting point for the five-year survival measurement one year backward from the usual time of diagnosis—the Dx sign—to the point at which early diagnosis could be achieved—the “plus” sign. Our group of 45-year-old patients referred to earlier would simply be diagnosed one year earlier at age 44. Only 50% would be alive at age 50 as before. We would not have given them an extra year backward of disease.

---

2 Morgan, R., personal communication.
Selection of an inappropriate starting point for measuring survival, then, is one kind of mistake often made in looking at the survival rates of individuals who are diagnosed at an early stage of their disease. Their increased survival rate is guaranteed, even if the therapy instituted does nothing at all to control or reverse the natural history of their disease.

The second example of a common error in analyzing the effectiveness of laboratory screening programs arises out of the relationship between the duration of the preclinical—early or asymptomatic—and clinical—late or symptomatic—stages of disease as shown in Figure 5. Studies of various diseases such as cancers of breast, stomach, and colon have indicated that patients with these cancers who have long preclinical stages tend also to have long clinical stages of disease (17–19). Conversely, individuals with short preclinical stages for these disorders tend to have relatively short clinical stages. Although this relationship probably characterizes most diseases; its effect usually has been ignored in analyzing programs of laboratory screening. For as we can see in Figure 6, once again early diagnosis—depicted as the vertical line at which laboratory screening has occurred—will always improve survival, regardless of whether or not therapy is effective (20). This is because the laboratory screening program will be more likely to pick out those individuals with a long preclinical stage than those with a short one. As a result, individuals diagnosed through screening are guaranteed longer clinical stages of disease and better short-term survival rates than are individuals diagnosed in the usual fashion, even the therapy instituted as a result of this early diagnosis has no effect whatsoever.

Deeply concerned over the enormous disability and loss of life associated with chronic disease, some proponents of laboratory screening programs have suggested that these programs must be instituted now, despite the absence of sound evidence for their effectiveness, under the assumption that they will either reduce the toll from chronic disease or, at worst, will have no effect. A third alternative, however, must be added to those of benefit and neutrality, and this is the possibility of frankly deleterious outcomes resulting from laboratory screening. To the social and emotional anguish that can accompany the labeling of an individual as diseased or at high risk of developing a dreaded disorder must be added the potentially harmful side-effects of long-term therapy, as, for two of many examples, in the increased risk of cardiovascular deaths following the use of either tolbutamide or phenformin in the treatment of adult onset diabetes (7), or in the use of high doses of estrogen in the treatment of coronary heart disease (21).

As key participants in health-screening programs, clinical chemists have, with rare exception, failed to take responsibility for evaluating whether these programs are of benefit to patients, are of no value, or are positively harmful. The simple alteration of risk factors and laboratory values has been erroneously accepted as evidence of effectiveness, and fundamental errors have been repeatedly committed in interpreting survival of volunteers who undergo laboratory screening. As a result, most of the clinical chemists’ efforts have been misdirected toward improving the efficiency of the process of laboratory screening.

At the present time, the vast majority of laboratory screening programs are being conducted in the absence of, or in direct contradiction to, evidence for their health effectiveness, and clinical chemists clearly must adopt a much more aggressive role in the evaluation of these programs. Indeed, at my own hospital, admission laboratory screening is limited to hemoglobin and urinalysis among adults, and no unsolicited laboratory studies are performed upon pediatric admissions.

In the possibility that they may be helpful to clinical chemists in adopting a more aggressive role in the evaluation of laboratory screening and periodic health examination programs, the criteria for screening programs recently developed at a meeting of the World Health Organization (1) are as follows:

1. Screening must lead to an improvement in end-results (defined in terms of mortality; physical, social, and emotional function; pain; and satisfaction) among those in whom early diagnosis is achieved or in the other members of the community.
The therapy for the condition must favorably alter its natural history, not simply by advancing the point in time at which diagnosis occurs, but by improving survival, function, or both. The modification of "risk factors" is not sufficient evidence of effectiveness, nor is the fact that the proposed therapy is "commonly accepted." Claims for therapeutic effectiveness must withstand rigorous methodologic scrutiny, and experimental evidence, such as controlled clinical trials, is a prerequisite. The measurement of survival and other end-results must withstand epidemiologic and biostatistical scrutiny.

Available health services must be sufficient both to ensure diagnostic confirmation among those whose screening is positive and to provide long-term care.

Compliance among asymptomatic patients in whom an early diagnosis has been achieved must be at a level demonstrated to be effective in altering the natural history of the disease in question.

The long-term beneficial effects, in terms of end-results, must outweigh the long-term detrimental effects of the therapeutic regimen utilized and the "labeling" of an individual as "diseased" or "at high risk."

The effectiveness of potential components of multiphasic screening should be demonstrated individually prior to their combination.

If the benefits of screening accrue to the community at large rather than, or in addition to, the individual identified (e.g., disease carriers, specific occupations), the community benefit claimed must withstand scientific scrutiny.

The appropriateness of the mix of screening tests to the target population must be considered, acknowledging that differences in the distributions of two diseases may render the combination of their respective screening tests inappropriate.

The cost-benefit and cost-effectiveness characteristics of mass screening and long-term therapy must be known. This knowledge is considered essential in developing an appropriate mix of diagnostic and therapeutic services in the face of finite manpower and financial resources. Therefore, a mechanism for the formal periodic weighing of costs against benefits or effectiveness should constitute a basic component of the initial screening activities.

The burden of disability for the condition in question (in terms of disease frequency, distribution, severity, and alternative approaches to its detection and control) must warrant action.

The cost, sensitivity, specificity, and acceptability of the screening test must be known, and it should lend itself to the utilization patterns of the target population.

Ideally, an estimate of the social benefit of preventing, arresting, or curing the condition in question should be known.

The preparation of this manuscript was supported in part by Grant No. 606-22-12 from the Department of National Health and Welfare of Canada.

References