Quantitative Approaches to the Evaluation of Screening Programs

N.E. Day, Ph.D.

MRC Biostatistics Unit, Cambridge, England, United Kingdom

The evaluation of screening programs for cancer is considered. Initial evaluation has to be in terms of mortality, but increasing importance should be attached to measures which evaluate the interaction of early detection with the disease process. These include the degree to which diagnosis is advanced, both in time and in stage of disease, and the ability of the screening test to identify presymptomatic lesions. Randomized trials are of major importance in establishing—in quantitative terms—the benefits of screening; in their absence, case-control methods can be adapted for this purpose, but cannot provide evidence of the same rigor. Currently, only screening for breast cancer and cervical cancer have been fully demonstrated to be effective.

The aim of mass screening is the detection of cancer at a presymptomatic stage. By advancing the time of diagnosis, one hopes to improve the chances of cure, and reduce the harmful consequences of advanced disease. The questions that one would ask to assess the extent to which the aim is being achieved would then be along the following lines.

1. Is mortality from the cancer in question effectively reduced in the population targeted for screening?
2. Is the overall rate of advanced cancer decreased?
3. How efficient is the screening procedure at detecting early lesions, how frequently should it be applied, and to whom?

These issues relate to the possible benefit of screening. One must also bear in mind the possibilities of harm which lead to such questions as:

1. Does the screening test itself carry a risk of increased morbidity?
2. Is there a substantial overdiagnosis of presymptomatic lesions, with consequent unnecessary and probably harmful treatment?
3. Does earlier diagnosis really help, or does it simply lengthen the time during which a person knows they have the disease?

The Role of Randomized Trials

The first issue is one of mortality. The benefit that screening should bring is a reduction in mortality from the cancer in question. By far, the most acceptable form of evidence for a reduction in mortality is that provided by a randomized trial. A population is randomly divided into 2 groups, one of which is offered screening, probably at regular intervals rather than just once, and the other offered routine medical care which is assumed to involve little or none of this form of screening. Clearly, such a trial can only be contemplated when the screening procedure is still regarded as experimental and does not form part of routine care. Unfortunately, many screening procedures are adopted clinically before their value has been established, thus preventing scientific trials of efficacy and greatly retarding correct evaluation. Radiological screening for gastric cancer and fecal blood testing for colorectal cancer are examples where, at least in some countries, very large numbers of tests are currently performed but little evidence exists on their benefits. The general acceptance of cervical cancer screening was unduly delayed for this reason.

Two malignancies for which exemplary trials of screening were performed and have been reported are breast cancer and cancer of the lung. The initial trial for breast cancer, the famous HIP study [1], compared a yearly procedure combining both mammography and a clinical examination repeated 4 times with routine care. The results after 18 years of follow-up [2] are shown in Fig. 1; the reduction in breast cancer mortality is clear. By contrast, the trials for lung cancer have not indicated a mortality reduction. One based at the Mayo clinic [3] compared sputum cytology and radiology every 4 months with yearly radiology; earlier trials had demonstrated the lack of effect for the latter but it was still accepted practice. The results after 7 years of follow-up are shown in Fig. 2. It should be noted that, in both trials, the comparison is between the group allocated to screening and the group allocated to routine care; the comparison is not confined to those in the screening group who were, in fact, screened. The HIP study provides a good example of the biases that arise when one compares the screened with the unscreened. Table 1 shows the breast cancer death rate in those who accepted the first screening invitation and those who refused, together with both rates in these 2 groups combined, i.e., the group allocated to screening, and in the control group. It is quite clear that comparison between screened and unscreened would be fraught with danger.

Large-scale randomized trials are currently in progress to assess the value of the fecal occult blood test in the early diagnosis of cancer of the colon and rectum. Although the test
Table 1. Death rates from breast cancer in the HIP study among women accepting and refusing the offer of screening, and in the control group: An example of the bias that randomization eliminates.

<table>
<thead>
<tr>
<th>Study group</th>
<th>Screened</th>
<th>Refused</th>
<th>Total</th>
<th>Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.14</td>
<td>4.75</td>
<td>6.34</td>
<td>8.96</td>
</tr>
</tbody>
</table>

Total reduction in mortality: 30%.

Fig. 1. Cumulative number of deaths due to breast cancer by interval since entry: all ages, study and control groups.

Fig. 2. Mortality from lung cancer observed in the Mayo Clinic randomized early detection trials.

is in widespread use in a number of countries, no valid evidence has, so far, been forthcoming which demonstrates an effect [4].

The Screening Process

The reduction in mortality achieved by screening programs ensues from the earlier stage at which cancer is diagnosed when detected by screening. Diagnosis is brought forward in time, and to an earlier stage of disease. The extent of these changes can be used to evaluate the efficacy of a screening program before information becomes available on mortality measures.

At the initial screening test, lesions are discovered which would otherwise have surfaced as clinical cancers subsequently. There is, therefore, a deficit of clinical cancers immediately after a screening episode among individuals classified as negative by the test. The extent and duration of this deficit provides the basic information on the capacity of the test to bring diagnosis forward in time. The incidence rate of cancers occurring after a negative test (the so-called interval cancers since they surface in the interval between 2 tests) is to be compared with the incidence rate expected in a comparable unscreened group, as shown in Fig. 3. The gap between the rate in a comparable unscreened population and the rate of interval cancers represents the cancers which were diagnosed earlier by the screening test. One can, thus, read directly from a curve such as that displayed in Fig. 3, the number of cancers in which the diagnosis was advanced by 6 months, 1 year, 2 years, and so on. In practice, of course, one does not have a smooth curve with values grouped over successive 6-month or 1-year intervals. Fig. 4 gives results from the large randomized trial of breast screening by mammography in Sweden [5], for women aged 50-69 years at entry to the study. One can see that, in the first year after a negative screening test, the rate of interval cancers was only 13% of that expected in a comparable unscreened group; in the second year after a negative test, the corresponding figure is 27%. These figures demonstrate that mammography detects 87% of the breast cancers that are destined to surface clinically in the 12 succeeding months, and 73% of the cancers that would surface 13-24 months after the test. In the third year, the figure is about 50%.

These quantities can be interpreted in terms of conventional measures of a screening procedure: sensitivity, specificity, lead time, and sojourn time [6]. The lead time is the length of time by which diagnosis has been brought forward in time by screening.