EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes - Screening

Mat Reeves BVSc, PhD

Outline:

I. Introduction

II. Characteristics of disease
   A. Pre-clinical phase

III. Concept of lead time

IV. Characteristics of screening tests
   A. Sensitivity
   B. Specificity
   C. Yield

V. Evaluation of the outcomes of screening
   A. Study designs (methods)
   B. Measures of effect
   C. Biases

VI. Pseudo-disease (Over diagnosis)

VII. Feasibility of screening (criteria for implementation)

I. Introduction
The goals of screening are to reduce mortality and morbidity (and/or avoiding expensive or toxic treatments). Screening is a form of secondary prevention. Screening is designed to detect disease early in its asymptomatic phase whereby early treatment then either slows the progression of disease or provides a cure. The premise of screening is based on the concept that early treatment will stop or retard progression of disease. Screening therefore has both diagnostic and therapeutic components.

Screening involves the examination of asymptomatic people who are then classified as either:

- Unlikely to have disease (TN or FN), or

- Likely to have disease and therefore require further diagnostic evaluation.
Screening is very different from diagnostic testing:

<table>
<thead>
<tr>
<th>Testing</th>
<th>Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sick patients are tested</td>
<td>Healthy, non-patients are screened</td>
</tr>
<tr>
<td>Diagnostic intent</td>
<td>No diagnostic intent</td>
</tr>
<tr>
<td>Low to high disease prevalence</td>
<td>Very low to low disease prevalence</td>
</tr>
</tbody>
</table>

There are two fundamentally different types of screening:

*Mass or population-based screening* is the application of screening tests to large, unselected populations e.g., mammography screening for breast cancer in women < 40 yrs of age.

*Case finding* is the use of screening by clinicians to identify disease in patients who present for other unrelated problems e.g., blood pressure measurements.

The format, organization, and intent of these two types of screening are fundamentally different. Mass screening requires a completely different organizational approach to successfully implement - involving policy makers, government, the medical community, and public health on a national basis.

II. Characteristics of Disease

For a disease to be a suitable candidate for screening it must have a sufficiently long *pre-clinical phase*. Pre-clinical phase is defined as the period between when early detection by screening is possible and when the clinical diagnosis would usually have been made.

A. *Pre-clinical phase (PCP):*

<table>
<thead>
<tr>
<th>Pathology begins</th>
<th>Disease detectable</th>
<th>Normal Clinical Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The point that a typical person seeks medical attention depends upon availability of medical care, as well as the level of medical awareness in the population. An example of disease with a long pre-clinical phase that suggests screening might be useful is colorectal cancer (i.e., PCP = 7-10 years). Diseases with a short pre-clinical phase are unlikely to be good candidates for screening e.g., childhood diabetes (weeks to a few months).

The prevalence of detectable pre-clinical disease in a population (and NOT the prevalence of disease itself) is a critical determinant of the potential utility of screening. The prevalence of pre-clinical disease is dependent upon:
i. incidence rate of disease  
ii. average length (duration) of pre-clinical phase  
iii. recent screening (decreases prevalence)  
iv. detection capabilities of the test (greater sensitivity results in higher prevalence)  

### III. Lead time

Lead time is the interval from detection by screening to the time at which diagnosis would have been made without screening. Lead time is the central rational of screening since it equals the amount of time by which treatment is advanced or made "early". Lead time results in the longer awareness of the disease and does not necessarily imply any improved outcome (since after lead time has occurred early treatment must then be effective for screening to be beneficial). Figure 1 illustrates this concept, in panel a (no screening), there are three cases at time zero who are already in the PCP i.e., in whom pathology has already started. In panel b, screening is conducted at time 0 and ‘converts’ the PCP in these three subjects into lead time. Thus the disease is advanced 3 years, 2 years, and 1 year, respectively, in these 3 cases.

Figure 1. Relationship Between Screening, Preclinical Phase, Clinical Phase and Lead Time

![Figure 1](image_url)

Lead time is not a theory or statistical artifact, it is what would be expected with early diagnosis and what must occur if screening is to be worthwhile (it is therefore a necessary but not sufficient condition for screening to be effective in reducing mortality).
Knowledge of the distribution of lead times is useful because it indicates the length of time by which detection and treatment must be advanced in order to achieve a level of improved mortality. It can also help suggest how often you should screen - for example, the estimated lead time for invasive colo-rectal cancer is 7-10 years. Thus guidelines suggest that an appropriate screening interval for sigmoidoscopy is every 5 years.

IV. Characteristics of screening tests

A. **Sensitivity**: the proportion of cases with a positive screening test among all cases of pre-clinical disease.

The target disorder is the pre-clinical lesion, not clinically evident disease. Test operating characteristics maybe very different between the two. Se is often first determined by applying tests to symptomatic patients, but screening Se is likely to be lower.

For sensitivity to be accurately defined, all individuals who have the pre-clinical disease must be identified using an acceptable "gold standard" diagnostic test. However, the true disease status of individuals who have a negative screening test is impossible to verify, since there is no justification to do a full diagnostic work-up (this is an excellent example of verification bias). Usually, Se can only be estimated in screening studies by counting the number of interval cases that occur over a specified period (e.g., 12 months) in persons who tested negative to the screening test. These interval cases are regarded as false negatives (FNs).

B. **Specificity**: ability of screening test to designate as negative people who do not have pre-clinical disease.

There is always an inherent trade-off between Se and Sp - as one increases the other must decline. Note also that an imperfect Se affects a few (the cases), whereas an imperfect Sp affects many (the healthy!). The FP rate (1 - Sp) needs to be sufficiently low for screening to be feasible, because the prevalence of pre-clinical disease is always low, thus the predictive value positive (PVP) will be low in most screening programs. (PVP is defined as the proportion of screen positive subjects who have pre-clinical disease i.e., TP/(TP + FP)).

PVP can be improved by screening only high risk populations or using a lower frequency of screening (which increases prevalence of pre-clinical disease). Because the prevalence of pre-clinical disease will fall in populations that are repeatedly screened, PVP will be expected to decline in a successful screening program making it increasingly inefficient.

C. **Yield**: the amount of previously unrecognized disease that is diagnosed and brought to treatment as a result of screening.

Yield is affected by the Se of the screening test (a lower Se means that a smaller fraction of diseased individuals are detected at any screening), and the prevalence of pre-clinical disease in the population. A higher prevalence will increase the yield,
thus aiming screening programs at high risk populations will increase its efficiency.

V. Evaluation of screening outcomes

A. Methods:

Experimental: Conduct a RCT of the screening modality and compare the disease-specific cumulative mortality rate between the groups randomized to screening or usual care (control). The randomized design is critical in eliminating confounding due to unknown and known factors, and in allowing a valid comparison (unaffected by lead time bias) between the two groups. The RCT also allows one to study the effects of early treatment, to estimate the distribution of lead times, and identify prognostic factors.

Problems: Expense, time (many years before results are available by which time the screening technology has often changed logistical problems, ethical concerns.

Non-experimental:

I. Cohort - comparison of advanced illness or death rate between people who chose to be screened and those that do not.
II. CCS - comparison of screening history between people with advanced disease (or death) and those unaffected (healthy).
III. Ecological - correlation of screening patterns and disease experience of several populations.

Problems: Confounding due to "health awareness" (people who choose to get screened are more health conscious and have lower mortality). Poor quality, often retrospective data. Difficult to distinguish screening from diagnostic examinations.

B. Measures of effect:

1) Comparison of survival experience (or duration)

Important!!! The efficacy of a screening program cannot be assessed by comparing the duration of survival of screen detected cases and cases diagnosed clinically. Despite the fact that this is commonly done, such analyses over-estimate the effect of screening because of the following three factors:

i) Selection bias - patients who choose to get screened are more health conscious, better educated and have an inherently better prognosis. Selection bias can also occur when subjects decide to get screened because they have symptoms.

ii) Lead-time Screen-detected cases will survive longer even without benefit of early treatment, simply because they are detected earlier! This is shown in Figure 2 - the average survival duration is increased from 1.3 years (with no screening - see panel a.) to 2.5 years (with screening - see panel b.) purely due to the fact that the 3 subjects
with disease were identified earlier (i.e., survival is increased due to the lead time).

Figure 2. Effect of screening and effective treatment on survival duration and mortality rate

iii) Length-biased sampling - screen detected cases are not simply a sample of all cases in a population but represent a sample of cases prevalent in the asymptomatic (pre-clinical) phase. Screening preferentially identifies slow growing, indolent cases that
have a long pre-clinical phase. Slow growing tumours will obviously have a better prognosis because they have both a long pre-clinical phase and a long clinical phase, as illustrated in Figure 3.

Figure 3. Length-biased sampling (from Fletcher et al., 1997)

2) Disease-specific mortality rate (DSMR)
The only truly valid measure of the efficacy of a screening program is to conduct a randomized screening trial where the DSMR in the group assigned to screening is compared to the group assigned to no screening. Unlike the survival duration, the DSMR will not be changed by early diagnosis (i.e., lead time). This concept is illustrated in Figure 2. In panel c, screening has resulted in a mortality reduction after 5 years, since the first subject no longer dies at year 5, (and so continues to live as indicated by the arrow). The average survival duration calculated at the end of year 5 is still 2.5 years. However, since there are now only 4 deaths (as opposed to 5 deaths previously) the DSMR drops from 100 per 100,000 person years to 80 per 100,000 (equivalent to a 20% reduction in mortality). Thus, it is the DSMR and not the survival duration, that accurately reflects the benefit of screening.

There is one caveat about the DSMR however; within the confines of a screening trial the specific cause of death is usually assigned by an adjudication committee. Ideally this assignment is done without knowledge of the screening group that a particularly subject was assigned to. However, maintain blinding to this fact is often difficult, especially given that there is usually a detailed examination of the specific circumstances around each death. If the original random assignment becomes unblinded there is the real potential for bias to be introduced into the process. For example, in a breast cancer trial, there might be a tendency to call deaths that occurred in the mammography group not breast cancer related, while in the control group there might be a tendency to overdiagnose breast cancer as a cause of death.
Because of these difficulties there is now considerable debate in the screening community that the ideal measure of screening efficacy should be all-cause mortality rather than the DSMR, because all-cause mortality is clearly not subject to these same biases (the subject is either dead or alive) (see Black WC, JNCI, 2002).

VI. Pseudo-disease or Over-diagnosis

One potential negative side-effect of screening is pseudo-disease or over-diagnosis which is the identification of disease that would not have become clinically apparent in the absence of screening. This can involve three forms:

i) Over-diagnosis - cases detected that would never have progressed to a clinical state – i.e., cancer cases with limited malignant potential. This is in fact an extreme form of length-biased sampling. A classic example is pap testing which despite reducing the incidence of invasive cervical cancer results in a large increase in the overall incidence of cervical cancer because of the “over-diagnosis’ of carcinoma in situ (See Table below). Other examples include PSA testing and low-grade prostate cancer, and mammography and ductal carcinoma in situ (see Ernster et al, 1996).

Table. Example of over-diagnosis: increase in carcinoma in situ of the cervix following introduction of mass screening (pap testing) in Connecticut (Laskey et al 1976)

<table>
<thead>
<tr>
<th>Year</th>
<th>Carcinoma in situ</th>
<th>Invasive carcinoma</th>
<th>Total</th>
<th>% in situ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950-1954</td>
<td>3.8</td>
<td>18.1</td>
<td>21.9</td>
<td>17</td>
</tr>
<tr>
<td>1955-1959</td>
<td>9.7</td>
<td>17.1</td>
<td>26.8</td>
<td>36</td>
</tr>
<tr>
<td>1960-1964</td>
<td>18.8</td>
<td>13.6</td>
<td>32.4</td>
<td>58</td>
</tr>
<tr>
<td>1965-1969</td>
<td>28.6</td>
<td>11.6</td>
<td>40.2</td>
<td>71</td>
</tr>
<tr>
<td>1970-1973</td>
<td>32.8</td>
<td>10.9</td>
<td>43.7</td>
<td>75</td>
</tr>
</tbody>
</table>

ii) Competing risks - cases are identified that would have been interrupted by an unrelated death. An example would be the identification of prostate cancer in an 85 year old man who would have died of stroke.

iii) Serendipity - the identification of disease due to diagnostic testing that comes about for another reason. Example, chest x-ray for TB screening that identifies lung cancer, or colonoscopy detection of colorectal cancer following a positive FOBT (Ederer et al, 1997).
VII. Feasibility and Need for Screening

There are several other important issues beyond the demonstration that screening leads to decreased mortality and/or morbidity that need to be addressed before deciding to invoke a screening program. These include:

a) **Acceptability**: The program should be convenient, free of discomfort, efficient and economical.

b) **Efficiency**: A low PVP indicates a wasteful program, as most of the test positive individuals worked up will not have disease. A high PVP can still be associated with only a few cases detected and a small reduction in overall mortality. If mortality from the disease is normally low or if the risk of death from other causes is high (for example in the very aged) then the screening program will not reduce mortality very much.

c) **Cost-effectiveness**: Should these health care dollars be spent on this program? Most population based screening programs run about $30-50,000 per year of life saved (or higher).

Another way of evaluating the need and feasibility of screening is to place all the subjects who would develop the condition you are trying to help (e.g., lung cancer or prostate cancer) into one of the following three groups:

1. **A cure is necessary but not possible (Nec,NotPos)**. In other words, if the target condition is death from lung cancer, these subjects are going to die of lung cancer regardless, and so would not be helped by a screening program.

2. **Cure is possible but not necessary (Pos,NotNec)**. This group includes subjects who develop lung cancer but will not die of it – this is an example of over-diagnosis (cases die of something else before dying of lung cancer). Again, a screening program will not be helpful to this group.

3. **Cure is necessary and maybe possible (Nec,Pos)** This is the only group that can benefit from screening! They represent cases of lung cancer who would have died of the disease had it not been for the effect of the screening program (this of course assumes that the screening program is effective in reducing the risk of death).

In terms of feasibility of screening it is helpful to consider the relative sizes of these three groups. While it is not possible to be absolutely sure of the sizes of these three groups, a reasonable estimate can be made based on knowledge of the natural history of disease, the potential of the intervention to identify the condition early, and potential effect of treatment to impact the outcome, and finally the potential to identify undiagnosed but benign disease.

The charts below give a hypothetical example of this sort of assessment for two cancers. For
Lung Cancer, the size of group 3 (Nec,Pos) is maybe 10%, but 80% are in group 1 (Nec, Not Pos) and hence can’t be helped at all, while a further 10% are in group 2 (Pos, Not Nec) and don’t need to be helped. For prostate cancer, the size of group 3 (Nec,Pos) is maybe 20%, but now we have a lot more men, say 60%, who are in group 2 (Pos, Not Nec) – these represent men with low grade or “benign” acting prostate cancer that will not kill them. Finally, there is another 20% in group 1 (Nec, Not Pos) who will die of prostate cancer regardless of any screening program. Obviously you would like most people to be in group 3 because this represent the positive (worthwhile) effects of screening. You would also like to keep group 2 as small as possible, since this represent the negative (or wasteful) effects of a screening program.

Lung Cancer:

Prostate Cancer: