Screening for Disease

Stanley H. Weiss, MD, FACP, FACE
Professor of Medicine at Rutgers New Jersey Medical School
Professor of Epidemiology at Rutgers School of Public Health
PI & Director, Essex-Passaic Wellness Coalition
Immediate Past Chair, International Joint Policy Committee of the Societies of Epidemiology
Cancer Liaison Physician and Oncology Committee Vice-chair, University Hospital, Newark, NJ
Executive Board, New Jersey Public Health Association
Due Diligence

In medicine as in life, it is essential to critically assess what you are told. Statements that on their face may seem reasonable, and indeed even may be true, can be misleading...

Those who know me, are well aware that I am quite allergic to furry animals.

This led to a problem with having pets in our home, despite my wife and children being animal lovers.

As a child, my daughter was very much in love with the following wonderful and adorable creature, to which I have never demonstrated any allergic reaction...
So, let me re-state this quite plainly:

A live unicorn has never led me to sneeze.

I think we can all concur on the (likely) truth of this statement.
Why not just test/screen everyone for everything?

1. The test must provide useful information, that is actionable.
2. Harm may be caused, not just benefit.
3. There are costs, even for something that appears trivial.
4. Society makes decisions on resource allocation.
5. Individuals may have different preferences, and come to different conclusions – esp. wrt what is worthwhile for them (such as when they must pay out-of-pocket for non-covered tests).

Thus, today we’ll review these issues, one by one, to see how they relate to medical screening.
• Most office visits include testing for temperature – with a thermometer.
• Thermometers vary in range and accuracy.
• There are also inappropriate methods.

(Cartoon showing inappropriate method deleted due to copyright.)
What is screening?

Screening is a test for *likely presence* of a disease in *asymptomatic* persons that can be applied to a *large population*.

- **Likely presence** — Screening is usually *not* definitive; suspicious or “positive” results must be examined further.
- **Asymptomatic** — If someone has signs or symptoms of a disease, testing for that disease is *diagnosis*, *not* screening.
- **Large population** — Assessment of the usefulness of a screening test depends on its being applied on a population basis.
It is TESTING, *not* screening...
When there is any of the following:

- History of family disease so that the individual is no longer “representative” of the general population or the “to be screened” population. Examples:
  - Parent with known genetic mutation that predisposes to the disease (*e.g.*, BRCA1 or BRCA2);
  - Other significant family history.
- The testing is **DIAGNOSTIC**, not screening, when a patient has:
  - Symptoms or
  - Signs.
Some Possible Problems with Screening

• A **negative** screening test is sometimes wrong, called a “false negative.”
  • If it is relied upon, disease may be missed by the physician and patient being lulled by the result.

• A **positive** test may require follow-ups by expensive and/or invasive tests that can cause harm – including iatrogenic harm in those without the disease/condition.
Flow diagram for a mass screening test

(Image deleted due to copyright.)

Examples

• Congenital metabolic diseases of the newborn (*e.g.*, phenylketonuria)
• Blood pressure
• Cholesterol
• Colorectal cancer
• Diabetes
• Breast cancer (*mammograms*)
• Cervical cancer (*Pap smears, HPV*)
• Prostate cancer
• Sexually transmitted infections
Some definitions

- **Positive** test result — result of a screening test that indicates that the patient *is* likely to have the disease.
- **Negative** test result — result of a screening test that indicates that the patient *is not* likely to have the disease.
- **False negative** test result — an (erroneous) **negative** test result in a patient who really **does** have the disease.
- **False positive** result — an (erroneous) **positive** test result in a patient who really **does not** have the disease.
Validity of screening tests — Characteristics of the test only (irrespective of the population)

- **Sensitivity** = ability of a test to correctly identify those who have disease
  - **Sensitivity** = \( \frac{A}{A+C} \)

- **Specificity** = ability of a test to correctly identify those who do *not* have disease
  - **Specificity** = \( \frac{D}{D+B} \)

These are inherent characteristics of *the test*, *not* influenced by prevalence of the disease. But these values may differ for special populations.
Sensitivity & specificity *both* matter

A perfectly sensitive test for prostate cancer:
A man walks into the office. If he is always told (without any examination or questioning) that he has prostate cancer:
What’s the sensitivity?
  100%
What’s the specificity?
  0%
How useful is this? NOT AT ALL!
Validity of screening tests — How the tests operate in a given population

- **Positive predictive value** (PPV) - probability that a patient with a positive test result actually has the disease
  - PPV = A / (A + B)

- **Negative predictive value** (NPV) - probability that a person with a negative test result is truly free of disease
  - NPV = D / (D + C)

PPV and NPV are influenced by prevalence of disease, including special characteristics of the group (or person) being tested.
THESE SAME PRINCIPLES APPLY TO ALL TYPES OF TESTING OR EVALUATION, NOT JUST IN THE SCREENING SITUATION

Quick examples

- If the person being screened is from a very low-risk population,
  - Positive PV likely low,
  - Negative PV likely high.
- If the person is from a very high-risk population, it is the reverse:
  - Positive PV likely high (confirming the pre-test impression);
  - Negative PV may be limited.
PPV & NPV can be much more meaningful than sensitivity and specificity

Suppose a test has 99% sensitivity and 99% specificity. Also suppose prevalence of asymptomatic disease in population is 0.1% (1 in 1,000). Consider a population of 100,000 individuals:

<table>
<thead>
<tr>
<th></th>
<th>Number with disease</th>
<th>Number without disease</th>
<th>Totals with and without each test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number with positive test result</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number with negative test result</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals with and without disease</td>
<td></td>
<td></td>
<td>100,000</td>
</tr>
</tbody>
</table>
PPV & NPV can be much more meaningful than sensitivity and specificity

[Suppose a test has 99% sensitivity and 99% specificity. Also suppose prevalence of asymptomatic disease in population is 0.1% (1 in 1,000).]

<table>
<thead>
<tr>
<th></th>
<th>Number with disease</th>
<th>Number without disease</th>
<th>Totals with and without each test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number with positive test result</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number with negative test result</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals with and without disease</td>
<td>100</td>
<td>99,900</td>
<td>100,000</td>
</tr>
</tbody>
</table>

Among these 100,000 persons, 100 (0.1%) will have the disease, and thus the remaining 99,900 (= 100,000 – 100) will not have the disease.
PPV & NPV can be much more meaningful than sensitivity and specificity

[Suppose a test has 99% sensitivity and 99% specificity. Also suppose prevalence of asymptomatic disease in population is 0.1% (1 in 1,000).]

<table>
<thead>
<tr>
<th></th>
<th>Number with disease</th>
<th>Number without disease</th>
<th>Totals with and without each test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number with positive test result</td>
<td>99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number with negative test result</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals with and without disease</td>
<td>100</td>
<td>99,900</td>
<td>100,000</td>
</tr>
</tbody>
</table>

Because the test is 99% sensitive, 99 of the 100 with the disease will test positive, and 1 will test negative.
PPV & NPV can be much more meaningful than sensitivity and specificity

[Suppose a test has 99% sensitivity and 99% specificity. Also suppose prevalence of asymptomatic disease in population is 0.1% (1 in 1,000).]

<table>
<thead>
<tr>
<th>Number with positive test result</th>
<th>Number with disease</th>
<th>Number without disease</th>
<th>Totals with and without each test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number with positive test result</td>
<td>99</td>
<td>999</td>
<td></td>
</tr>
<tr>
<td>Number with negative test result</td>
<td>1</td>
<td>98,901</td>
<td></td>
</tr>
<tr>
<td>Totals with and without disease</td>
<td>100</td>
<td>99,900</td>
<td>100,000</td>
</tr>
</tbody>
</table>

Because the test is 99% specific, 99% (98,901) of the 99,900 without the disease will test negative, and 1% (999) will test positive.
PPV & NPV can be much more meaningful than sensitivity and specificity

[Suppose a test has 99% sensitivity and 99% specificity. Also suppose prevalence of asymptomatic disease in population is 0.1% (1 in 1,000).]

<table>
<thead>
<tr>
<th>Number with positive test result</th>
<th>Number without disease</th>
<th>Totals with and without each test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number with disease</td>
<td>99</td>
<td>1,098</td>
</tr>
<tr>
<td>Number with negative test result</td>
<td>999</td>
<td></td>
</tr>
<tr>
<td>Totals with and without disease</td>
<td>1</td>
<td>98,902</td>
</tr>
<tr>
<td></td>
<td>98,901</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100,000</td>
</tr>
<tr>
<td></td>
<td>99,900</td>
<td></td>
</tr>
</tbody>
</table>

There are thus a total of 1,098 persons with positive test results and 98,902 with negative test results.
PPV & NPV can be much more meaningful than sensitivity and specificity

[Suppose a test has 99% sensitivity and 99% specificity. Also suppose prevalence of asymptomatic disease in population is 0.1% (1 in 1,000).]

<table>
<thead>
<tr>
<th>Number with positive test result</th>
<th>Number with disease</th>
<th>Number without disease</th>
<th>Totals with and without each test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number with positive test result</td>
<td>99</td>
<td>999</td>
<td>1,098</td>
</tr>
<tr>
<td>Number with negative test result</td>
<td>1</td>
<td>98,901</td>
<td>98,902</td>
</tr>
<tr>
<td>Totals with and without disease</td>
<td>100</td>
<td>99,900</td>
<td>100,000</td>
</tr>
</tbody>
</table>

We can now calculate the predictive values in this setting. In particular: \( \text{PPV} = \frac{99}{1098} \approx 9\% \). In other words, a positive test result still means \textit{only} a 9% chance of having the disease.
When does it make sense to screen?

1. Disease being screened for is an important health problem.
2. Treatment when the disease is latent (asymptomatic) is more effective and/or less risky than treatment after the development of symptoms.
3. There is a suitable screening test:
   - Suitability criteria include: test validity (sensitivity, specificity, predictive values), low cost, ease of administration, safety, imposition of minimal discomfort on administration, acceptability to both patients and practitioners.
4. There is appropriate follow-up available for those identified with a positive test:
   - There is a treatment for the disease: In other words, making a diagnosis at this “earlier” time point may make a difference. This is called a “critical point.”
   - Facilities for diagnosis and treatment are available.
Spectrum of Infection and Disease
Stage Issues in Epidemiology

Early Diagnosis and the Natural History of Disease

STAGES
1. Biologic Onset
2. Early Diagnosis Possible
3. Usual Clinical Diagnosis
4. Outcome
Figure 3-9. Spectrum of disease.
Sackett et al. 2nd Ed.
Fig 5-1 (p. 155)

Figure 5-1 from Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine, Second Edition (Boston: Little, Brown and Company), 1991. Ch. 5: Early Diagnosis.

(Deleted due to copyright)
Early Diagnosis and the Natural History of Disease

CRITICAL POINT

in the natural history of a disease is a point before which therapy is either more effective or easier to apply than afterwards.
Figure 5-2 from Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine, Second Edition (Boston: Little, Brown and Company), 1991. Ch. 5: Early Diagnosis.

(Deleted due to copyright)
Early Diagnosis and the Natural History of Disease

TIMING OF CRITICAL POINT

• If between “biologic onset” and “early dx possible” time points, then screening or case finding is too late to help

• If between “usual diagnosis possible” and “outcome,” then early detection is a waste of time - just wait until symptomatic patients seek help
Early Diagnosis and the Natural History of Disease

TIMING OF CRITICAL POINT

- It is only when a critical point lies between “early dx possible” and “usual clinical dx” that screening and case finding hold any promise of improving the outcomes of those who have the target disorder.
When is screening beneficial?

• Treatment should be more effective and/or less risky if it is initiated when the disease is latent than if it is initiated when the disease is symptomatic.

• The risks of subsequent diagnostic work-ups in false positives are low enough that they don’t offset the benefits of early detection in true positives.
When is screening harmful?

- False-positive results
- Overdiagnosis
- Overtreatment
Evaluation of Screening

NOTE: there will often be a transient apparent increase in incident disease when screening programs are first introduced. SO, must observe over time.

Randomized clinical trials that are carefully designed and monitored are the “gold standard” for assessing screening programs.

Screening efficacy cannot be assessed by comparing post-diagnosis survival with and without screening, because of several types of biases that can occur:

- **Lead time bias**
- **Length time bias**
- **Patient self-selection bias**
Early Diagnosis and the Natural History of Disease

• Early diagnosis will always appear to improve survival,

• Even when therapy is worthless

• Example: patients whose cancers are diagnosed early WILL have better 5-year survivals than other patients diagnosed in later, symptomatic stages.

• It is the inference that our observations prove the value of early diagnosis that is faulty!
What Is The Observed Effect of Introducing a Test that Detects a Condition Earlier in Time, but NOTHING can be done about it?

The person still dies at the exact same time – not a single day of life is gained.

BUT the time they live KNOWING they have the condition IS prolonged.  
(So, there may be adverse psychological consequences.)

This leads to “lead time” – also known as “stage-shift” bias.  THIS IS VERY IMPORTANT.
Lead time bias

With screening, the lead time in diagnosis prolongs survival even if death is not delayed.

Image from: http://www.cancer.gov/types/lung/research/nlst-qa
Figure 5-3 from Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine, Second Edition (Boston: Little, Brown and Company), 1991. Ch. 5: Early Diagnosis.

(Deleted due to copyright)
Figure 5-4 from Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine, Second Edition (Boston: Little, Brown and Company), 1991. Ch. 5: Early Diagnosis.

(Deleted due to copyright)
Sackett Figs 5-3 & 5-4 (together)


(Deleted due to copyright)
Length time bias

Image from: http://www.cancer.gov/types/lung/research/nlst-qa
Length time bias

Note that of the total of 12 cases, only 2 of the rapidly progressing cases will be identified by screening, while 4 of the slowly progressing cases will be identified.

Length-time bias refers to the fact that screening is more likely to detect slowly progressing cases. Thus, time to occurrence of symptoms (or, by similar reasoning, survival time) after diagnosis is artificially lengthened.

Length time bias

(Image deleted due to copyright.)

What Determines How Often Screening Should be Conducted

- What if fast growing cancers were to be less amenable to treatment than slower growing ones?
- Increasing the frequency of screening might just detect ONLY the faster growing ones, about which little might be done!
Figure 5-5 from Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine, Second Edition (Boston: Little, Brown and Company), 1991. Ch. 5: Early Diagnosis.

(Deleted due to copyright)
Sackett Fig 5-6 (p. 163)

*Figure 5-6 from Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine, Second Edition* (Boston: Little, Brown and Company), 1991. Ch. 5: Early Diagnosis.

(Deleted due to copyright)
Early Diagnosis and the Natural History of Disease

Three Reasons for the above Paradox:

1) **Healthy cohort effect**
   [thus; always examine whether mortality from OTHER causes has not changed]

2) **Failure to correct for “zero-time shift”**
   in assessing the value of early diagnosis
   [see figures 5-3 and 5-4]
Early Diagnosis and the Natural History of Disease

Three Reasons for the above Paradox (cont’d):

3) Relationship between **pre-clinical and clinical duration** of disease (e.g., existence of both slow & fast growing tumors) which affects likelihood of early diagnosis. (Slow growing tumors are often detectable longer than fast growing ones – and thus are more likely to be identified by any early dx strategy.)

[see figures 5-5 and 5-6]
Hypothesis testing

In statistics, there are typically two hypotheses:
• *Null* hypothesis (H₀) — whatever is being studied *has no* effect;
• *Alternate* hypothesis (H₁ or Hₐ) — whatever is being studied *has* an effect.

Hypothesis testing asks whether there is enough evidence to *reject* the *null* hypothesis. With a good statistical test, the data should lead to the right conclusion most of the time. When the data lead to the wrong conclusion, there are two ways it can be wrong:
• Type I error - *Rejecting* H₀ when it is in fact *true*
• Type II error- *Not rejecting* H₀ when it is in fact *false*

In test for disease —
• H₀: The person *does not* have the disease.
• H₁: The person *does* have the disease.
• Type I error — False positive
• Type II error — False negative
Never confuse Type I and II errors again:

Just remember that the Boy Who Cried Wolf caused both Type I & II errors, in that order.

First everyone believed there was a wolf, when there wasn't. Next they believed there was no wolf, when there was.

Substitute "effect" for "wolf" and you're done.

Kudos to @danolner for the thought. Illustration by Francis Barlow "De pastoris puero et agricolis" (1687). Public Domain. Via wikimedia.org
Why do guidelines change?

- New research refines evidence of effectiveness;
- New testing technologies;
- Improved knowledge of variation in risk factors in subpopulations allows for better design of screening programs tailored to different population-based disease risk profiles;
- Changes in treatment alter relative value of diagnosis during latent stage;
- New research may better assess harm resulting from false positives and/or overdiagnosis.
ACS Breast Cancer Screening Guidelines

Previous Guidelines
(from 2003)

• Yearly mammograms starting at age 40 and continuing for as long as a woman is in good health
• Clinical breast exam (CBE) about every 3 years for women in their 20s and 30s and every year for women 40 and older
• Women should know how their breasts normally look and feel and report any breast change promptly to their health care provider. Breast self-exam (BSE) is an option for women starting in their 20s

New October 2015 Guidelines

• Women ages 40 to 44 should have the choice to start annual breast cancer screening with mammograms if they wish to do so. The risks of screening as well as the potential benefits should be considered.
• Women age 45 to 54 should get mammograms every year.
• Women age 55 and older should switch to mammograms every 2 years, or have the choice to continue yearly screening.
• Screening should continue as long as a woman is in good health and is expected to live 10 more years or longer.
• All women should be familiar with the known benefits, limitations, and potential harms linked to breast cancer screening. They also should know how their breasts normally look and feel and report any breast changes to a health care provider right away.
Reassessment of published results

On 23 September 2010, we discussed prostate cancer screening at a quarterly meeting of the Essex County Cancer Coalition. Among the problems with major randomized clinical trials of prostate-specific antigen (PSA) testing that we discussed were:

- Inadequate follow-up time (e.g., ten years, inadequate to assess difference in outcomes given natural history of prostate cancer);
- Cross-contamination of the study arms (persons who were not supposed to receive screening did receive screening anyway);
- Failure to include adequate numbers of African-American men, who are at higher risk;
- Issues in design of screening — wrong PSA cutoff value or non-uniform PSA cutoff values in different subgroups.
- Inadequate power, i.e., sample sizes too small.
One study we discussed was a randomized clinical trial of prostate cancer screening conducted among 20,000 men in Göteborg, Sweden for 14 years of screening and follow-up. Reported results included:

- A roughly 50% decrease in mortality due to prostate cancer in the screening group vs the control group;
- Benefit greatest 10+ years from beginning of trial;
- Half the attendees who died of Pca in the screening group were diagnosed in the first round of screening, and many of those were 60+ years old at entry
- Bias towards underestimation of screening effect because number of men in the control group getting screening independently was not known.

(Next slide shows published results.)
The Göteborg Results

Cumulative Risk of Diagnosis:

Cumulative Risk of Death:

The approximate ten-year point is marked by an arrow.
Reassessment of published results

Yet there were aspects of the Göteborg study that were not published that appear to necessitate reconsideration of its apparently favorable preliminary results for PSA testing.

Critical assessment of published findings is often a complicated, challenging process.

No wonder that as more information is gathered, our interpretations and recommendations sometimes change.