PHC 6000: Epidemiology Methods 1

Lesson 11 – Screening

Lusine Yaghjyan
Department of Epidemiology

Lesson Objectives

- Communicate the pros and cons of screening
- Explain the importance of selection bias, lead-time and length-time bias in screening programs
- Understand how to evaluate screening efforts

Prevention

- Primary:
  - To prevent new cases of disease occurring and therefore reduce the incidence of disease
- Secondary:
  - To reduce the consequences of disease (death or morbidity) by *screening* asymptomatic patients to identify disease in its early stages and intervening with a treatment which is more effective because it is being applied earlier.
  - It cannot reduce disease incidence
- Tertiary:
  - To reduce the consequences of disease (esp. complications and suffering) by treating disease and/or its direct complications in symptomatic patients.

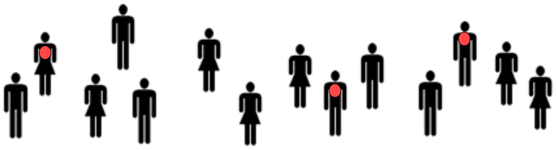
Example - Fire Prevention

- Primary (prevent fires from starting)
  - Education
  - Outside fire bans (drought)
- Secondary (early detection)
  - Smoke detectors
  - Lookout towers
- Tertiary (reduce consequences)
  - Fire brigades & smoke jumpers

Screening

Purpose: to prevent, interrupt, or delay the development of advanced disease in the subset with a pre-clinical form of the target disease through early detection and treatment.

Identifies asymptomatic individuals who may have the disease



Screening

- "...the identification of unrecognized disease or defect by the application of tests, examinations or other procedures..."
- "...sort out apparently well persons who probably have disease from those who probably do not."
- "...not intended to be diagnostic..."

Screening

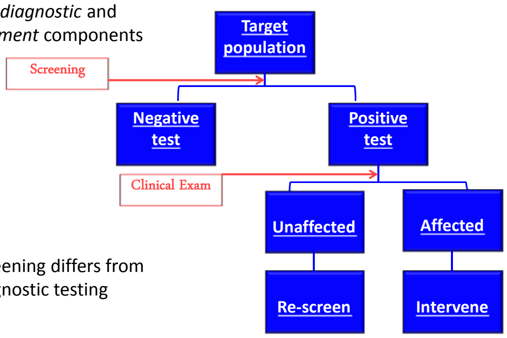
- Objective: to reduce mortality and/or morbidity by early detection and treatment.
- *Secondary prevention*.
- Asymptomatic individuals are classified as either unlikely or possibly having disease.

Types of Screening

- Population-level screening
  - National level policy decision to offer mass screening to a whole sub-group of a population
    - e.g., mammography screening (women 40+)
    - e.g., Vision and hearing screening of school-age children
- Individual-level screening
  - Occurs at the individual patient-physician level
  - Also refereed to case finding
    - e.g., BP screening every time you visit MD

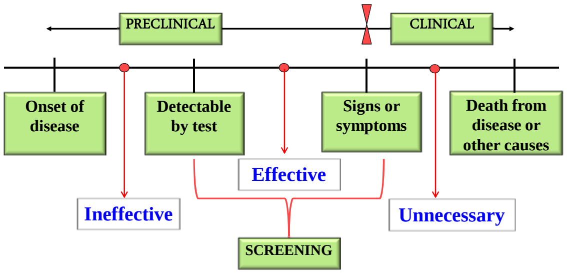
Screening Process

Effective screening involves both *diagnostic* and *treatment* components

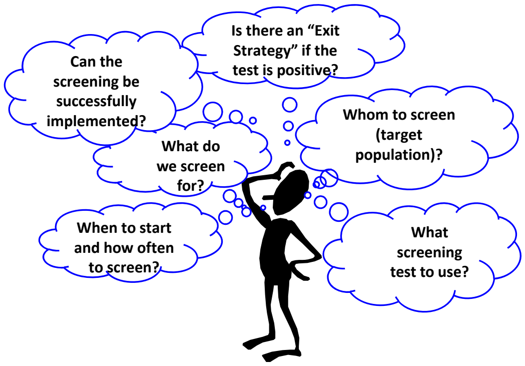


Screening differs from diagnostic testing

Natural History of Disease



Important Considerations


Is there an "Exit Strategy" if the test is positive?
Can the screening be successfully implemented?
Whom to screen (target population)?
What do we screen for?
When to start and how often to screen?
What screening test to use?

When Should We Screen?

When:

- It is an important health problem
- There is an accepted and effective treatment
- Disease has a recognizable latent or early symptomatic stage
- There are adequate facilities for diagnosis and treatment
- There is an accurate screening test
- There is agreement as whom to consider as cases

Screening Examples

- Hypertension
- Cancer screening (breast, cervical cancer, colorectal)
- Tuberculosis
- STDs

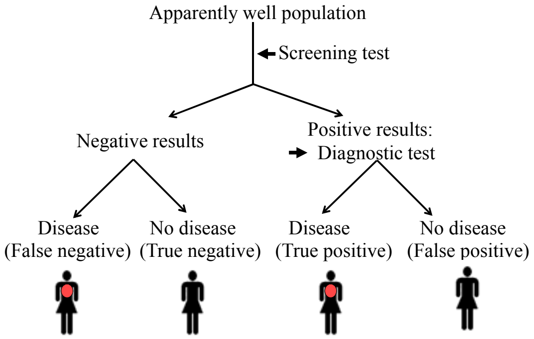
Characteristics of a Good Screening Test

- Validity (accuracy, e.g., good sensitivity and specificity)
- Reliable (repeatability, e.g. consistent results)
- Yield (number of cases identified per thousand screened)
- Cost – benefit (compare costs avoided due to early detection of the disease against cost of the screening. Does the test merely uncover more disease that is expensive to treat without appreciable advantage?)
- Acceptable (discomfort, hassle, cost of obtaining test)
- Follow-up services (plan needed to deal with positive results)

Gold Standard

- The gold standard is the best single test (or a combination of tests) that is considered the current preferred method of diagnosing a particular disease X
- All other methods of diagnosing X, including any new test, need to be compared against this 'gold' standard
- The gold standard is different for different diseases
- The gold standard for X may be considered outdated or inadequate, but any new test designed to replace the gold standard *has* to be initially validated against the gold standard. If the new test is indeed better and there are ways to prove that then the new test may become the new gold standard

Logic of Screening



Apparently well population

Screening test

Negative results

Positive results: Diagnostic test

Disease (False negative)

No disease (True negative)

Disease (True positive)

No disease (False positive)

	True Disease	
Test Results	Present	Absent
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Characteristics of Screening Tests

How good is the test at identifying people with and without the disease?

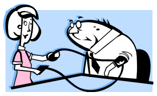
In other words:

If we screen a population, what proportion of people will be correctly classified with regards to the disease status?

- **Sensitivity:** the proportion of cases with a positive screening test among all individuals with pre-clinical disease, i.e. ability of a test to identify those who have the disease
- **Specificity:** the proportion of individuals with a negative screening test result among all individuals with no pre-clinical disease, i.e. ability of a test to exclude those who don't have the disease

Related Issues

- Tests with dichotomous results – tests that give either positive or negative results
- Tests of continuous variables – tests that do not yield obvious “positive” or “negative” results, but require a cutoff level to be established as criteria for distinguishing between “positive” and “negative” groups



How Good is the Test?

		Disease present?	
		Yes	No
Test result	Positive	True positive	False positive
	Negative	False negative	True negative

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive + False negatives}}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative + False positives}}$$

Predictive Values

In the clinical setting, a more important question is:

If the test results are positive (or negative) in a given patient, what is the probability that this patient has (or does not have) the disease?

In other words:

What proportion of patients who test positive (or negative) actually have (or do not have) the disease in question?

Predictive Value

		Disease present?	
		Yes	No
Test result	Positive	True positive	False positive
	Negative	False negative	True negative

$$\textbf{Positive predictive value (PPV)} = \frac{\text{True Positives}}{\text{True Positive + False Positive}}$$

$$\textbf{Negative predictive value (NPV)} = \frac{\text{True Negative}}{\text{True Negative + False Negative}}$$

Prevalence and Predictive Value

Hypothetical Example 1 - Screening Test A

100 people are tested for disease. 15 people have the disease; 85 people are not diseased. So, prevalence is 15%.

Test sensitivity is 67% and specificity 53%

		Disease present?		
		Yes	No	Total
Test result	Positive	10	40	50
	Negative	5	45	50
	Total	15	85	100

Positive Predictive Value:
10/50 × 100 = 20%

Negative Predictive Value:
45/50 × 100 = 90%

Prevalence and Predictive Value

Hypothetical Example 2 - Screening Test A

100 people are tested for disease. 30 people have the disease; 70 people are not diseased. So, prevalence is 30%.

Test sensitivity is 67% and specificity 53%

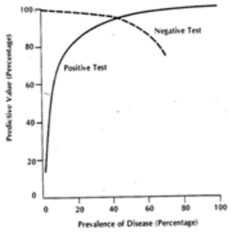
		Disease present?		
		Yes	No	Total
Test result	Positive	20	33	53
	Negative	10	37	47
	Total	30	70	100

Positive Predictive Value:
20/53 × 100 = 38%

Negative Predictive Value:
37/47 × 100 = 79%

Prevalence and Predictive Value

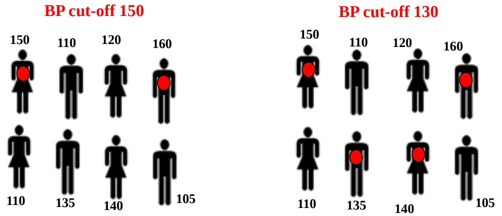
- Using the same test in a population with higher prevalence increases positive predictive value
- Increased prevalence results in decreased negative predictive value



Relationship between disease prevalence and predictive value in a test with 95% sensitivity and 85% specificity.
(Mausner, 1985)

Prevalence and Predictive Value

- *When considering predictive values of diagnostic or screening tests, recognize the influence of the prevalence of disease*
- Prevalence will be affected by changing cut offs for continuous outcomes



Specificity & Predictive Value

- As specificity increases, positive predictive value increases
- As sensitivity increases, positive predictive value also increases, but to a much lesser extent

Disease present?

Test result

	Yes	No	
+	100	400	500
-	100	400	500
	200	800	1000

Prevalence =20%
Sensitivity = 50%
Specificity = 50%

PPV=100/500=20%

Disease present?

Test result

	Yes	No	
+	180	400	580
-	20	400	420
	200	800	1000

Prevalence =20%
Sensitivity = 90%
Specificity = 50%

PPV=180/580=31%

Disease present?

Test result

	Yes	No	
+	100	80	180
-	100	720	820
	200	800	1000

Prevalence =20%
Sensitivity = 50%
Specificity = 90%

PPV=100/180=56%

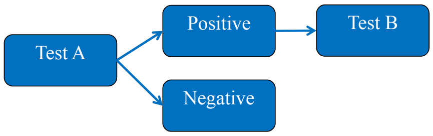
Use of Multiple Screening Tests

- Sequential (Two-stage) Testing
- Simultaneous Testing

Sequential Testing
(Two-Stage Screening)

After the first (screening) test was conducted, those who tested **positive** were brought back for the second test to further reduce false positives

The overall process will have increased specificity but reduced sensitivity


Test A

Positive

Test B

Negative

Hypothetical Two-Stage Screening

Test 1 (non-fasting blood sugar)

Assume:
- Disease prevalence = 5%, population = 10,000
- Sensitivity = 70%, specificity = 80%
- Screen **positives** from the first test

Disease present?

Test 1

	Yes	No	Total
Positive	350	1900	2250
Negative	150	7600	7750
Total	500	9500	10000

Only Pos. Test 1 are given Test 2

Hypothetical Two-Stage Screening

Disease present?

Test 1

	Yes	No	Total
Positive	350	1900	2250
Negative	150	7600	7750
Total	500	9500	10000

Disease present?

Test 2 (Glucose Tolerance Test)

	Yes	No	Total
Positive	315	190	505
Negative	35	1710	1745
Total	350	1900	2250

sensitivity 90%, specificity 90%

Net Sensitivity = 315/500 = 63%

Net Specificity = (7600 + 1710)/9500= 98%

Hypothetical Two-Stage Screening

Subject is disease positive when **both** tests are positive

Subject is disease negative when **either** of the tests is negative

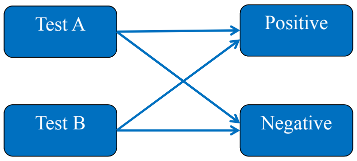
Net sensitivity = Sensitivity 1 x Sensitivity 2

Net specificity = Spec1 + Spec2 –(Spec1 x Spec2)

Simultaneous Screening

The goal is to maximize the probability that subjects with the disease (true positives) are identified (increase sensitivity)

Consequently, more false positives are also identified (decrease specificity)



Simultaneous Screening

Disease positives are defined as those who test positive **by either one test or by both tests (i.e. at least one test is positive)**

Net sensitivity = sens 1 + sens 2 –(sens 1 x sens 2)

Disease negatives are defined as those who test negative **by both** tests

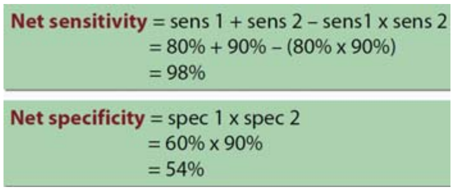
Net specificity = specificity test 1 x specificity test 2

Example of a Simultaneous Testing

In a population of 1000, the prevalence of disease is 20%

- Two tests (A and B) are used at the same time
- Test A has sensitivity of 80% and specificity of 60%
- Test B has sensitivity of 90% and specificity of 90%

Calculate net sensitivity and net specificity from using Test A and Test B simultaneously



Net sensitivity = sens 1 + sens 2 – sens1 x sens 2
= 80% + 90% – (80% x 90%)
= 98%

Net specificity = spec 1 x spec 2
= 60% x 90%
= 54%

Net Gain and Net Loss

In simultaneous testing, there is a net gain in sensitivity but a net loss in specificity, when compared to either of the tests used

In sequential testing when positives from the first test are re-tested, there is a net loss in sensitivity but a net gain in specificity, compared to either of the tests used

Trade off Between Sensitivity and Specificity

• Screening for PKU in newborns places a premium on sensitivity rather than on specificity

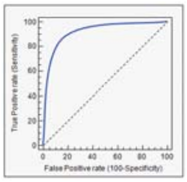
– The cost of missing a case is high
– Effective treatment exists
– Downside is a large number of false-positive tests

• Screening for breast cancer should favor specificity

– Further assessment of those tested will result in biopsies that are invasive and costly

Receiver-operator Characteristic (ROC curve)

- Sensitivity and specificity are inversely proportional, meaning that as the sensitivity increases, the specificity decreases and vice versa
- An ROC curve is a line graph that plots the probability of a true positive result against the probability of a false positive result for a range of cutoff points



- ROC curves are particularly valuable ways of comparing alternative tests for the same diagnosis. The overall accuracy of a test can be described as the area under the ROC curve; the larger the area, the better the test
- Tests that perform less well have curves that fall closer to the diagonal running from lower left to upper right

Reproducibility (reliability)

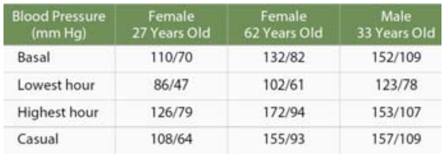
Means that the results of a test or measure are identical or closely similar each time it is conducted

- Because of variation in laboratory procedures, observers, or changing conditions of test subjects (such as time, location), a test may not consistently yield the same result when repeated
- Different types of variation
  - Intra-subject variation
  - Intra-observer variation
  - Inter-observer variation

Intra-subject variation

- Variation in the results of a test conducted over (a short period of) time on the same individual
- The difference is due to the changes (such as physiological, environmental, etc.) occurring to that individual over that time period

Variation in Blood Pressure Readings: A 24-Hour Period



Blood Pressure (mm Hg)	Female 27 Years Old	Female 62 Years Old	Male 33 Years Old
Basal	110/70	132/82	152/109
Lowest hour	86/47	102/61	123/78
Highest hour	126/79	172/94	153/107
Casual	108/64	155/93	157/109

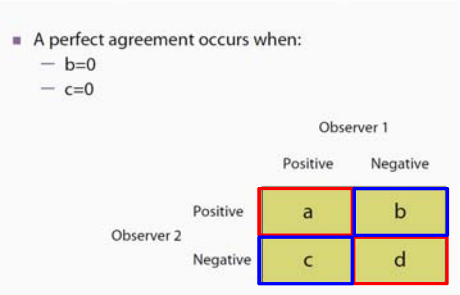
Inter-Observer and Intra-Observer Variation

Inter-observer variation is a variation in the result of a test due to multiple observers examining the result (inter = between)

Intra-observer variation is a variation in the result of a test due to the same observer examining the result at different times (intra = within)

The difference is due to the extent to which observer(s) agree or disagree when interpreting the same test result

Agreement between Two Observers



- A perfect agreement occurs when:
  - b=0
  - c=0

		Observer 1	
		Positive	Negative
Observer 2	Positive	a	b
	Negative	c	d

Concordant readings

Discordant readings

Percent Agreement (Concordance)

$$\text{Overall Percent Agreement} = \frac{a + d}{a + b + c + d} \times 100$$

$$\text{Percent Positive Agreement} = \frac{a}{a + b + c} \times 100$$

Note: This is a conditional probability

Example

		Radiologist A	
		Positive	Negative
Radiologist B	Positive	4	5
	Negative	2	6

Calculate overall and percent positive agreement

Percent Agreement (Concordance) -more than 2 categories

	Reading #1			
Reading #2	Abnormal	Suspect	Doubtful	Normal
Abnormal	A	B	C	D
Suspect	E	F	G	H
Doubtful	I	J	K	L
Normal	M	N	O	P

$$\text{Percent agreement} = \frac{A+F+K+P}{\text{Total}} \times 100$$

Kappa Statistic

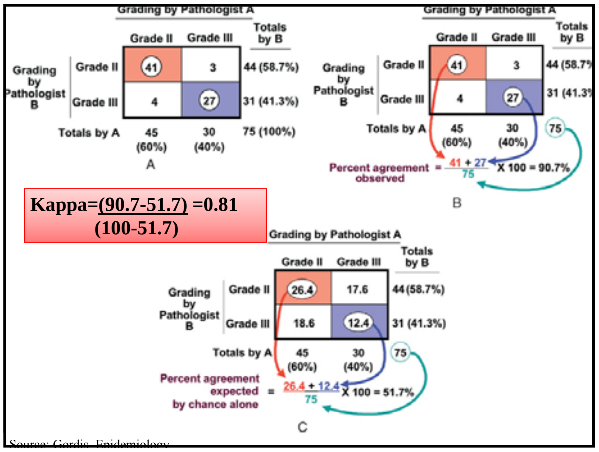
To what extend the readings agree beyond what we would expect by chance alone?

Kappa=

$$\frac{\text{P observed-P expected by chance alone}}{1 - \text{P expected by chance alone}}$$

OR

$$\frac{\text{\% agreement observed - \% agreement expected by chance alone}}{100 - \text{\% agreement expected by chance alone}}$$



Grading by Pathologist A

		Grade II	Grade III	Totals by B
Grading by Pathologist B	Grade II	41	3	44 (58.7%)
	Grade III	4	27	31 (41.3%)
	Totals by A	45 (60%)	30 (40%)	75 (100%)

A

Grading by Pathologist A

		Grade II	Grade III	Totals by B
Grading by Pathologist B	Grade II	41	3	44 (58.7%
	Grade III	4	27	31 (41.3%
	Totals by A	45 (60%)	30 (40%)	75

$$\text{Percent agreement observed} = \frac{41 + 27}{75} \times 100 = 90.7\%$$

B

$$\text{Kappa}=\frac{(90.7-51.7)}{(100-51.7)} = 0.81$$

Grading by Pathologist A

		Grade II	Grade III	Totals by B
Grading by Pathologist B	Grade II	26.4	17.6	44 (58.7%)
	Grade III	18.6	12.4	31 (41.3%)
	Totals by A	45 (60%)	30 (40%)	75

$$\text{Percent agreement expected by chance alone} = \frac{26.4 + 12.4}{75} \times 100 = 51.7\%$$

C

Source: Gordis, Epidemiology

Interpretation

≤0.20 – Poor
0.21-0.40 –Fair
0.41-0.60 – Moderate
0.61-0.80 – Substantial
0.81-0.99 – Almost perfect

Weighted Kappa Statistic

- Sometimes, we are more interested in the agreement across major categories in which there is meaningful difference
- A weighted kappa assigns less weight to agreement as categories are further apart
- The determination of weights for a weighted kappa is a subjective issue on which even experts might disagree in a particular setting

		helpfulness of lectures in the college				
		very helpful	somewhat helpful	neutral	somewhat a waste	complete waste
helpfulness of lectures in the college	very helpful	🟨			More weight	
	somewhat helpful		🟨			
	neutral		Less weight	🟨		
	somewhat a waste				🟨	Less weight
	complete waste	More weight				🟨

Agreement for Continuous Variables

- Correlation coefficient is an alternative estimate of inter-observer reliability (bivariate and partial)
  - Bivariate correlation is the correlation between two variables
  - Partial correlation is the correlation between two variables when controlling the effect of one or more additional variables

< 0.20 very low
0.20-0.39 low
0.40-0.59 moderate
0.60-0.79 high
0.80-1.00 very high

Must also consider the p-value

Examples of Using Agreement

❖ Validation studies (self-reported and confirmed diagnosis, family history of cancer, use of prescription medication, etc.)

❖ Planning variable definitions/data collection in epi study

Examples:

- Agreement between different methods of breast density assessment in breast cancer epidemiology (BI-RADS, Tabar, Wolfe's , Boyd)
- Multi-center studies – do we need diagnosis confirmations by a single expert or could we show on a subset that the agreement is satisfactory

Examples of Using Agreement (continued)

❖ Planning variable definitions/data collection in epi study

Examples:

- Validating imputations for missing data:
  - Consortium data where breast cancer stage is missing for a subset of participants
  - An algorithms is used to define Stage for those with missing using tumor size and nodal involvement;
  - Near perfect agreement confirms that the algorithm could be used to impute breast cancer stage for those with missing data

Another Example: Methods Studies

50 mammograms , stratified random sampling on calendar year and degree of density from participants of a community medical surveillance program; density on original mammograms assessed using FMMP codes

Breast Density Assessment Study #1: Second radiologist reading

Study group	N	Simple kappa coefficient	Weighted kappa coefficient
Whole study group	50	0.65	0.73
High and intermediate density only	44	0.86	-
Intermediate and low density only	30	0.31	-
High and low density only	26	0.81	-
Date of Mammogram ≤1995	15	0.68	0.77
Date of Mammogram 1996-2000	18	0.64	0.69
Date of Mammogram >2000	17	0.63	0.70

Substantial to perfect agreement between density categories assigned on the basis of FMMP codes and categories assigned based on the second radiologist's reading.

Study # 2: Analysis of digital images

Study group/subset	Results adjusted for resolution		
	N	Simple Kappa coefficient	Weighted Kappa coefficient
Whole study group	50	0.19	0.27
High and intermediate density only	44	0.36	-
Intermediate and low density only	26	-0.03	-
High and low density only	30	0.24	-
Date of Mammogram ≤1995*	15	0.17	-
Date of Mammogram 1996-2000	18	0.00	0.17
Date of Mammogram >2000	17	0.34	0.44

Fair to moderate agreement between density assessment by FMMP codes vs. that from digital Images

In FMMP Cohort, using FMMP mammography codes for characterization of breast density is better than using information from digitized radiography films

Kappa: Advantages and Limitations

Advantages:

•If it shows a good or excellent agreement then we can use any of the two procedures to measure a particular characteristic based on our convenience and comfort

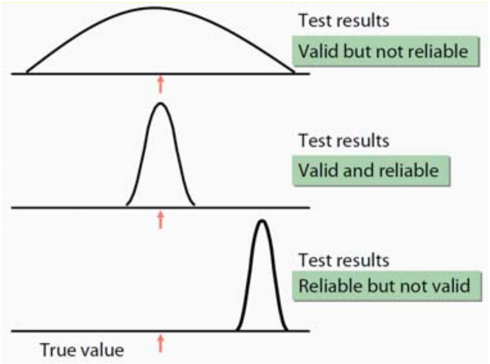
•Helps to choose cost effective procedure for any diagnosis, if there are more than one diagnostic procedures available

Disadvantages:

•Influenced by prevalence of the characteristic being measured

•Affected by biases

Validity and Reliability of the Test Results



Feasibility

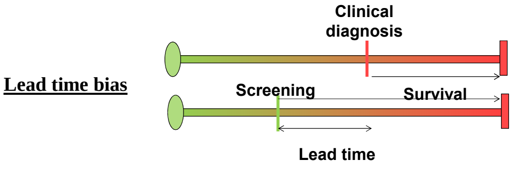
- Efficacy: Should we screen? (scientific)
- Effectiveness: Can we screen? (practical)
- Cost-effectiveness: Is it worth it? (scientific, practical, policy, political)

Screening-associated Biases

- Lead-time
- Length
- Volunteer
- Over-diagnosis and overtreatment
- Incidence-prevalence

Lead Time Bias

- Refers to a spurious increase in longevity associated with screening simply because diagnosis was made earlier in the course of the disease



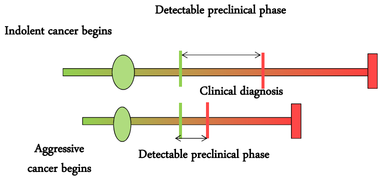
Lead Time

- Equals the amount of time by which treatment is advanced or made “early”
- Not a theory or statistical artifact but what is expected and must occur with early detection
- Does not imply improved outcome!!
- ***Necessary but not sufficient condition*** for effective screening

Length Bias

- Slowly progressing tumors have more opportunity than faster ones to be detected by screening
- Slowly progressive tumors take longer to lead to death than faster ones
- Therefore, the screen-detected cancers will appear to have an increased survival after diagnosis

Length Bias



In reality, the improved survival is a result of these cancers being more slowly progressing. Thus, the survival rate of a group of people with screen-detected cancers will be artificially increased due to length time bias compared with the survival rate of those with non screen-detected cancers.

Other Biases: Volunteer Bias

- Volunteers or compliers are better educated and more health conscious – thus they have inherently better prognosis
- Volunteers for screening programs may be healthier, on average, than persons who do not participate in screening programs
- Volunteers may also be more likely to participate and may be at overall higher risk because of family history or lifestyle

Other Biases: Over-diagnosis

May occur as the result of:

- False positives: persons who screen positive and are truly disease free, yet are erroneously diagnosed as having the disease
- Diagnosing tumors with limited malignant potential (example PSA testing, breast cancer in-situ)

Consequence:

- More favorable form of disease would result in a better prognosis and give the appearance of a very effective screening program

Screening Risks

- Physical harm from the procedure
- Over-diagnosis
- Over-treatment
- Psychological harm
  - True positives
    - ❖ "Labeling effect"(classified as diseased from the time of the test forward).
    - ❖ Fear of future tests
    - ❖ Anxiety
    - ❖ Monetary expense
  - False-negatives:
    - ❖ False sense of security
    - ❖ Delayed intervention
    - ❖ Disregard of early signs or symptoms which may lead to delayed diagnosis.

What Could Go Wrong?

- Diagnostic Test Errors
- Biologic Variability
- Measurement error
- Intra-observer variability
- Inter-observer variability

Barriers to Screening

- Lack of knowledge/awareness of symptoms of cancer
- Cost/lack of insurance
- Lack of physician recommendation
- Language barrier
- Cultural beliefs
- Psychological factors
- Fear

Evaluation of Screening Outcomes

RCT

- Compare disease-specific mortality rate (DSMR) between those randomized to screening and those not
- Eliminates all forms of bias (theoretically)
  - Example: breast cancer screening, Health Insurance Plan of New York)
- Problems:
  - Expense, time consuming, logistically difficult, contamination, non-compliance, ethical concerns, changing technology

The Only Valid Measure of Screening is…

Disease-specific Mortality Rate (DSMR)

$$\frac{\textbf{the number of deaths due to disease}}{\textbf{total number of persons with the disease}}$$

- the only gold-standard outcome measure for screening
- NOT affected by lead time
- however, there can be problems with the correct assignment of cause of death (hence some researchers advocate using only all-cause mortality as the outcome).