Methodologic Overview of Screening Studies

Diana L. Miglioretti, PhD
University of California Davis

Thanks to William Black, MD for many of these slides!
Learning Objectives

• Understand the difference between screening and diagnostic tests.
• Appreciate the balance between the harms and benefits of screening.
• Understand when and how lead time bias, length bias, and overdiagnosis can influence screening studies.
• Compare and contrast potential endpoints in screening trials.
Overview

• Screening vs. diagnosis
• Importance of natural history of disease
• Benefits and harms of screening
• Bias in screening studies
• Study designs
What is Screening?

• ACR Task Force on Screening Technologies:
  “Systematic testing of asymptomatic individuals for some target disease.”

Purpose:
• Prevent, interrupt, or delay the development of advanced disease in individuals with preclinical disease through early detection (or prevention).
• Reduce morbidity &/or mortality due to the target disease.
Screening vs. Diagnosis

**Screening**
- Healthy individuals
- Asymptomatic
- Low prevalence of disease, many people tested
- Test non-diagnostic: separates groups into high/low risk of disease
- Test is noninvasive, low risk, not time consuming, inexpensive

**Diagnosis**
- Patients, ill individuals
- Symptomatic
- High prevalence of disease, few people tested
- Test diagnostic
- Test may be invasive, higher risk, time less of a consideration, costly

**Burden of proof for effectiveness is higher for screening interventions than for diagnostic & treatment interventions**
SCREENING & THE NATURAL HISTORY OF DISEASE
Natural History of Disease

- Preclinical Phase of Disease
  - Detectable Preclinical Phase
    - Onset of Disease
    - Detectable by Test
    - Signs or Symptoms
  - Clinical Phase of Disease
    - Death from Disease or Other causes

Critical Point

The point in the natural history of disease before which therapy is more effective.

For screening to be effective, the critical point must occur within the detectable preclinical phase.
Natural History of Disease

Onset of Disease  Detectable by Test  Signs or Symptoms  Death from Disease or Other causes

Detectable Preclinical Phase

Critical Point

Screening may be effective
Natural History of Disease

Onset of Disease

Critical Point

Detectable by Test

Detectable Preclinical Phase

Signs or Symptoms

Death from Disease or Other causes

Screening futile
Natural History of Disease

- Onset of Disease
- Detectable by Test
- Detectable Preclinical Phase
- Signs or Symptoms
- Death from Disease or Other causes

Critical Point

Screening unnecessary
All screening programs do harm; some also do good, and of these, some do more good than harm.”

Gray et al. BMJ 2008. 336(7642) 480-483
Screening Cascade

From: The Harms of Screening: A Proposed Taxonomy and Application to Lung Cancer Screening

Harris, et al. JAMA Intern Med. Published online December 09, 2013
Potential Benefits from Screening

• Decreased morbidity & mortality from disease
• Decreased morbidity & mortality from treatment
  – Less morbid treatments if disease identified earlier
• Decreased anxiety about disease (true negatives)
• Increased awareness about disease
Harms of Screening

• Direct effect of test (radiation, reaction to contrast)
• Morbidity & mortality from work-up (FP)
• Overdiagnosis / overtreatment
  ⇒ Morbidity & mortality from treatment
• Anxiety, stress, worry (from test & FP)
• Financial strain
• Opportunity costs
• Screening fatigue if start too young
## Possible Results of Screen

<table>
<thead>
<tr>
<th>Result</th>
<th>Benefit?</th>
<th>Harm?</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Negative</td>
<td>No?</td>
<td>Yes</td>
</tr>
<tr>
<td>False Negative</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>False Positive</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>True Positive (TP), overdiagnosis</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>TP, no change in outcome</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>TP, improved outcome</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
POTENTIAL BIASES IN SCREENING STUDIES
Potential Biases in Screening Studies

• Lead time bias
  – Survival time increased by lead time, even if screening ineffective

• Length bias
  – Less aggressive tumors more likely to be screen detected

• Overdiagnosis bias
  – Diagnosis of disease that would never harm an individual

All favor screening!
Lead Time Bias – Effect on Survival

Without screening

Onset of Disease | Detectable by Test | Signs or Symptoms | Death

With screening

Onset of Disease | Detectable by Test | Signs or Symptoms | Death

Length Bias

Rapidly progressive

Slowly progressive

Test

Time

Length Bias

Rapidly progressive

Slowly progressive

Test

Time

Length Bias Examples

• Over-representation of ER+ vs. triple negative invasive breast cancer in mammography screening

• Over-representation of brochioalveolar carcinomas versus small-cell carcinomas in CT lung cancer screening trials.

Screening more likely to detect slower growing tumors than aggressive tumors.
Overdiagnosis

• The diagnosis of disease that will never cause symptoms or death during the person’s lifetime
• A **harm** of screening
  – May turn people into **patients** unnecessarily
  – May lead to **treatments** that don’t benefit the person and may do **harm** ➔ **Overtreatment**
Overdiagnosis

• **Type I: non-progressive (possibly regressive)**
  – *e.g.*, some DCIS

• **Type II: progressive**
  – *e.g.*, early stage cancer, prostate cancer in older men
  – More common in slow growing tumors, older individuals, and those with comorbidities

• Once a disease is screen-detected, it is typically impossible to know if it was “overdiagnosed”
Figure 3. Screen Detection Capability Based on Tumor Biology and Growth Rates

Stage Distributions as Rates and Percentages

Caution: percentages (among diseased) misleading if overdiagnosis present!

A=Optimal case
B=Worst case
C=Intermediate case

Effect of Overdiagnosis on Case Survival

Without screening

1000 patients with cancer

10 years later

900 died from cancer

10 yr survival = 100/1000 = 10%

With screening

1000 people overdiagnosed

1000 patients with cancer

10 years later

1100 did not die from cancer

900 died from cancer

10 yr survival = 1100/2000 = 55%
# Effects of Overdiagnosis on Screening Performance

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Underlying Truth</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease</strong></td>
<td><strong>No Disease</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Test +</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td></td>
<td>(TP)</td>
<td>(FP)</td>
</tr>
<tr>
<td><strong>Test -</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
<tr>
<td></td>
<td>(FN)</td>
<td>(TN)</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{TP+O}{TP+O+FN} \)  
PPV = \( \frac{TP+O}{TP+O+FP} \)  
Specificity = \( \frac{TN}{TN+FP-O} \)  
NPV = \( \frac{TN}{TN+FN} \)  
Detection rate = \( \frac{TP+O}{N} \)
Effects of Overdiagnosis on Outcomes

- Falsely increases detection rate and incidence
- Falsely increases sensitivity of test
- Falsely increases PPV of test
- Falsely improves stage distribution (as a percentage)
- Falsely improves case survival
- Does not decrease population mortality
SCREENING STUDY DESIGNS
Randomized Controlled Trial

• Strongest study design
• Randomization evenly distributes the known and unknown confounders
• Groups similar except for screening test under study
• Removes selection bias (except volunteer bias)
Screening RCT

Enroll screen eligible subjects

Randomize

Screen Arm

Control Arm

Assess Endpoints

Assess Endpoints
Screening Protocol

- Study population
- Screening exam: method, frequency, interval
- Equipment specifications
- Interpretation, reader experience/training
- Follow-up after screening assessment
  - Work-up criteria
  - Treatment plan (standard of care?)
- Quality control at every step
- Endpoints/outcomes/reference standard
- Sample size/power calculation
Study Population

- High risk for preclinical disease?
  - Limits generalizability to average-risk individuals
- No clinical signs or symptoms of disease
- Willing and able to:
  - undergo screening or not, for all rounds and duration of study
  - undergo workup and treatment
  - to be followed for outcomes
Screening Strategy

- Number of screens
- Time between screens
- Include first/prevalent screen or require prior screening?
- Follow-up time after last screen
- Equipment specifications
- Interpretation specifications
- Work-up of positive assessments
Endpoints/Outcomes

• Comparisons of **survival** are invalid and biased!
  • Lead time bias, length bias, overdiagnosis bias

• Disease-specific mortality
  • Most widely used & accepted
  • Assumes cause of death can be determined accurately and screening doesn’t increase risk of dying from other causes

• All cause mortality =>
All Cause Mortality

- Not affected by cause-of-death misclassification
- Insensitive measure of efficacy
  - Breast cancer screening: sample size 25–60 times larger (1.2-1.5 million per arm) if overall vs. DSM
- Still useful to measure along with DSM
  - May reveal deficiencies in randomization
  - Puts screening in perspective
    - Annual FOBT: 33% ↓ DSM ⇒ 1% ↓ overall mortality
- Helps ensure a major harm (or benefit) is not being missed.
  Important if test or treatment causes mortality.
Other Endpoints/Outcomes

• Absolute risk reduction or number needed to screen to prevent one death (reciprocal)

• Stage of target disease at diagnosis (rates, not percentages)
  – Include both screen-detected and interval cancers

• Adverse events
  – Morbidity caused or prevented by screening

• Quality of life

• Resource utilization and costs
  – Medical and nonmedical/opportunity costs
Sample Size Determination

- Disease-specific mortality
  - Or expected sensitivity and specificity if performance trial
- Effectiveness of screening in reducing disease-specific mortality
  - Or expected differential performance if performance trial
- Frequency/duration of screening intervention and follow-up
- Compliance in each arm
- Desired power and significance level
## Sample Size

alpha = 0.05 (one-sided), beta = 0.20

<table>
<thead>
<tr>
<th></th>
<th>Smokers 60-69</th>
<th>All 40-69</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screen vs. no screen</td>
<td>2,100</td>
<td>12,700</td>
</tr>
<tr>
<td>RRR 50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compliance 100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screen vs. no screen</td>
<td>44,800</td>
<td>274,000</td>
</tr>
<tr>
<td>RRR 20%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compliance 80%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RCT Limitations

• Compliance, contamination rates
• Selection bias
• Statistical power
• Generalizability
  • Participants
  • Screening tests and radiologists
  • Treatment and supportive care
Example: **National Lung Screening Trial**

- Enrolled 53,454 persons 55-74 years at high risk (30 pack years) from 8/2002 to 4/2004
  - 33 medical centers
- Randomly assigned to three annual screens with either
  - Low-dose CT
  - Single view chest x-ray
- Followed through 2009
- Power: 90% to detect 21% decrease in lung cancer mortality
## Results: National Lung Screening Trial

Adherence >90%

<table>
<thead>
<tr>
<th></th>
<th>Low Dose CT</th>
<th>X-Ray</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall rate</td>
<td>24.2%</td>
<td>6.9%</td>
</tr>
<tr>
<td>False-positive rate</td>
<td>23.3%</td>
<td>6.5%</td>
</tr>
<tr>
<td>Cancer rate (per 100,000 PY)</td>
<td>645</td>
<td>572</td>
</tr>
<tr>
<td>Lung cancer deaths (per 100,000 PY)</td>
<td>247</td>
<td>309</td>
</tr>
<tr>
<td>Risk reduction (RR)</td>
<td>20% (95% CI 6.8% to 26.7%, p=0.004)</td>
<td></td>
</tr>
<tr>
<td>All cause mortality RR</td>
<td>6.7% (95% CI 1.2% to 13.6%, p=0.02)</td>
<td></td>
</tr>
<tr>
<td>Other cause mortality RR</td>
<td>3.2% (p=0.28)</td>
<td></td>
</tr>
</tbody>
</table>
B  Death from Lung Cancer

Cumulative No. of Lung-Cancer Deaths

Years since Randomization

- Chest radiography
- Low-dose CT
Comparison to Test Proven to Reduce Mortality

• Screening technology evolution, compare performance of new test to current standard
• Lab studies
  – Behavior in lab may not reflect clinical practice
• Observational studies
• Prospective trials
  • Individuals may get both tests (e.g., DMIST) or may be randomized
  • Both tests – higher power but can’t evaluate interval cancer rates
• Caution: improvements in sensitivity (or cancer detection rate) may not lead to improvements in mortality and may increase overdiagnosis and overtreatment!
Example: DMIST

- Enrolled 49,528 women at 33 sites in US and Canada
- Women underwent both digital and film screening mammography, in random order
- Interpreted by two different radiologists
- Follow-up: cancer within 15 months or normal mammogram 10-15 months
- Power: 80% to detect decrease in AUC of 0.06.
Observational Studies

• Correlation / Ecological
  • Describe trends in screening frequency and disease specific mortality rates over time and/or location

• Case-control
  • Cases with adverse outcome (e.g., death from disease) matched to controls without outcome
  • Screening intensity in cases compared to intensity in controls via odds ratio

• Cohort
  • Group of screened and unscreened individuals followed over time
Observational Studies:
Correlation / Ecological Analysis

• Describe trends in screening frequency and disease specific mortality rates over time and/or location

• E.g., Harding, *JAMA Intern Med*, 2015
  • 547 SEER counties, 53,720 breast cancer diagnoses in 2000 followed for 10 years
  • Exposure: % screened in prior 2 years from NHIS and BRFSS
  • Outcomes: age-adjusted Br Ca incidence and mortality
  • 10% increase in screening ➔
    • 16% increase in cancer incidence
    • No change in breast cancer mortality
Observational Studies: Case-Control

• Cases with adverse outcome (e.g., death from disease) matched to controls without outcome
• Screening intensity in cases compared to intensity in controls via odds ratio
• Example: Elmore, JNCI, 2005
  – Women who died from breast cancer from 1983 to 1998 matched to other women on age and risk level
  – Measured and compared screening in prior 3 years
  – Non-significant reductions in mortality associated with screening
  – Limitation: not much mammography during time period
Observational Studies - Cohort

• Group of screened and unscreened individuals followed over time

• Common design in Europe with dissemination of population-based screening programs
Bias Associated with Observational Studies

• **Observer and recall bias** due to retrospective data collection
• **Selection bias**: Screened individuals at different risk than unscreened individuals
• **Confounding**: known or unknown differences b/w screened and unscreened groups also related to outcomes
  • Can only adjust for known confounders

*Can bias results in either direction!*
Simulation/Decision Modeling

• RCTs can’t answer all questions
  – Who should be screened?
  – Start and stop ages?
  – How often?
  – Impact on population incidence and mortality?
  – Cost effectiveness of different strategies?

• Decision models can extrapolate RCT results
Microsimulation Models

Life history without breast cancer

Birth

Death from other cause

Life history with breast cancer, no screening

Birth

Onset

Cancer diagnosed

Death from breast cancer

Life history with breast cancer & screening

Birth

Onset

Cancer detectable at screening

Cancer diagnosed (screen-detected)

Cancer diagnosed (in the absence of screening)

Death from other cause

Screening

Lead time

Sojourn time

Effect of screening: life-years gained and breast cancer deaths averted
Summary: Characteristics of Potential Screening Test

• Condition is important health problem
• Test is safe, noninvasive, inexpensive, and acceptable to population
• Test is able to detect preclinical disease with sufficient lead time to improve effectiveness of treatment
• Test is reasonably specific, esp. for findings requiring invasive work-up
• Overdiagnosis is rare
• Screening regimen considers the full screening process
  – Initial interpretation, work-up, follow-up, treatment
Summary

• Screening differs from diagnostic testing
• Potential effectiveness depends on the natural history of disease and treatment effectiveness
• Survival statistics are inappropriate and biased
• RCT is most valid design, but has limitations
• Once a test is shown to reduce mortality, important to measure and weigh benefits vs. harms
• Decision modeling can be used to extrapolate study results to help inform public policy
References