Methodologic Overview of Screening Studies

Janie M Lee, MD, MSc

University of Washington

With thanks to Diana Miglioretti

Fred Hutch Cancer Center

UW Medicine DEPARTMENT OF RADIOLOGY

1

Financial Disclosure

- Research grant: GE Healthcare

2

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.

3

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.

4

Summary

- Screening differs from diagnostic testing
- Potential effectiveness depends on the natural history of disease and treatment effectiveness
- RCT is most valid design, but has limitations
- Survival statistics are inappropriate and biased
- 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

5

5

Screening vs. Diagnosis

Screening	Diagnosis
• Healthy individuals	• Patients, ill individuals
• Asymptomatic	• Symptomatic
• Low prevalence of disease, many people tested	• High prevalence of disease, few people tested
• Test non-diagnostic: separates groups into high/low risk of disease	• Test diagnostic
• Test is noninvasive, low risk, not time consuming, inexpensive	• 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

6

6

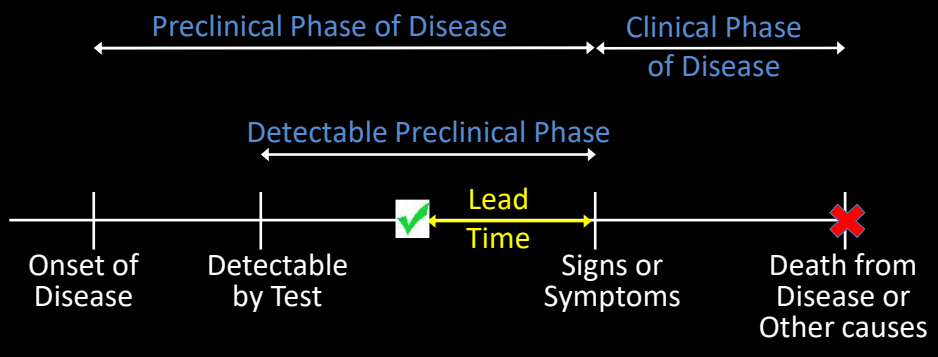
Summary

- Screening differs from diagnostic testing
- Potential effectiveness depends on the natural history of disease and treatment effectiveness
- RCT is most valid design, but has limitations
- Survival statistics are inappropriate and biased
- 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

7

7

Natural History of Disease



ADAPTED FROM: Black WC. J Clin Oncol 2006;24:3252. © Am Soc Clin Onc

8

8

Critical Point

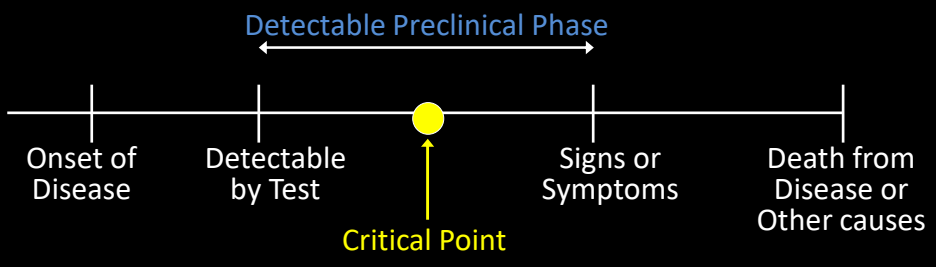
The point in the natural history of disease before which therapy is more effective.

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

9

9

Natural History of Disease

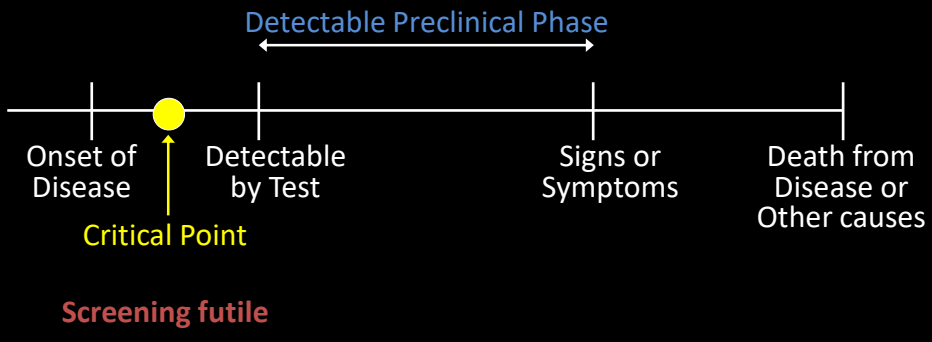


Screening may be effective

10

10

Natural History of Disease

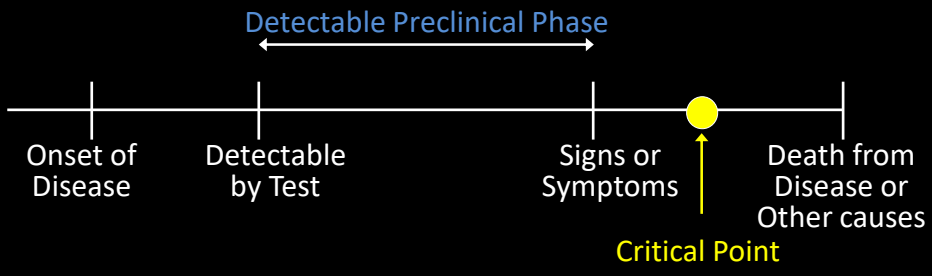


Screening futile

11

11

Natural History of Disease



Screening unnecessary

12

12

Potential Biases in Screening Studies

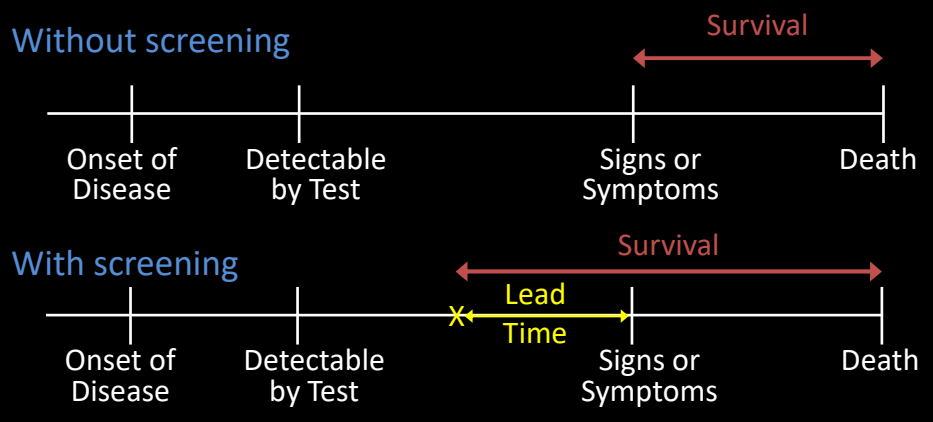
- Lead time bias
  - Survival time increased by lead time, even if screening ineffective
- Length bias
  - Less aggressive tumors more likely to be screen detected
- Overdiagnosis bias
  - Diagnosis of disease that would never harm an individual

All favor screening!

13

13

Lead Time Bias – Effect on Survival

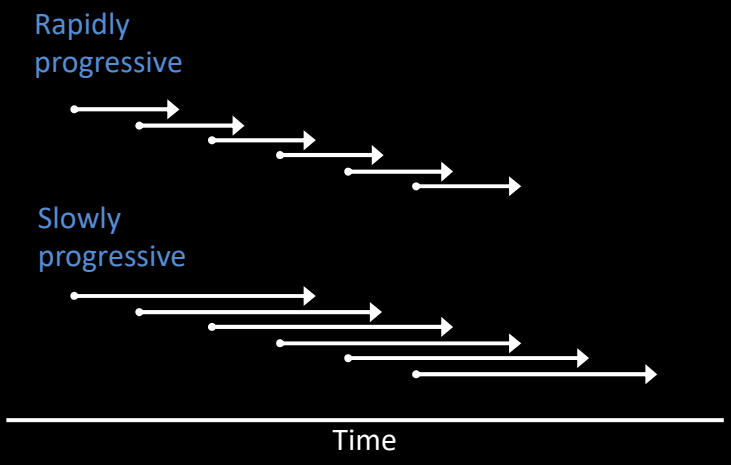


ADAPTED FROM: Black WC. J Clin Oncol 2006;24:3252. © Am Soc Clin Onc

14

14

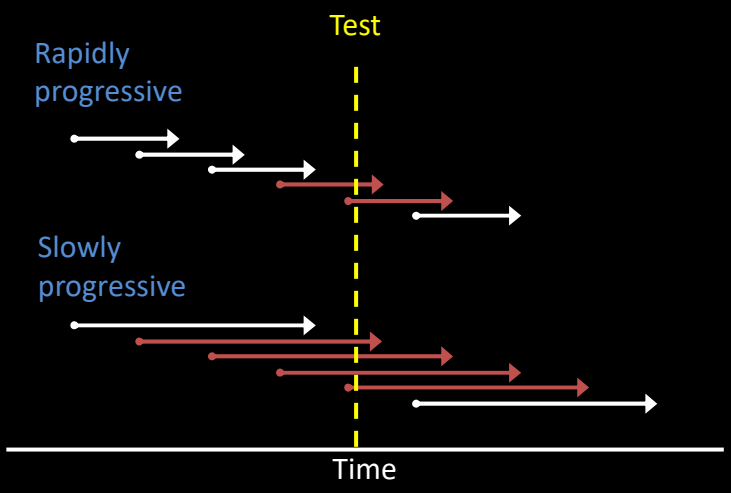
Length Bias



ADAPTED FROM: Black WC. J Clin Oncol 2006;24:3252. © Am Soc Clin Onc

15

Length Bias



ADAPTED FROM: Black WC. J Clin Oncol 2006;24:3252. © Am Soc Clin Onc

16

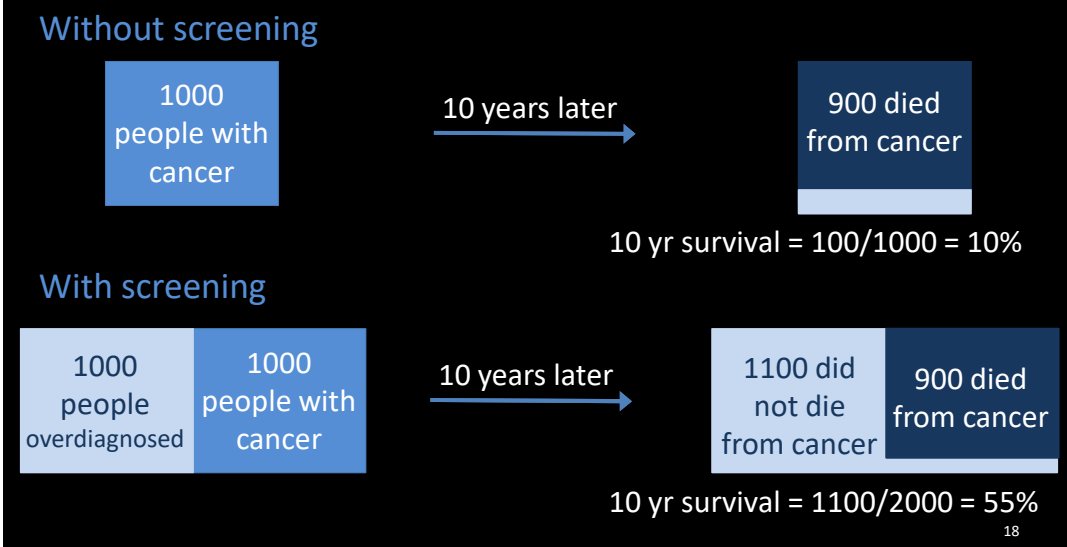
Overdiagnosis

- The diagnosis of disease that will never cause symptoms or death during the person's lifetime
- A harm of screening
  - Leads to treatments that don't benefit the person and may do harm → "Overtreatment"
  - Once a disease is screen-detected, it is typically impossible to know if it was "overdiagnosed"

17

17

Effect of Overdiagnosis on Case Survival



18

18

Effects of Overdiagnosis on Screening Performance

Underlying Truth

Test Result	Disease	No Disease
Test +	True Positive (TP)	False Positive (FP)
Test -	False Negative (FN)	True Negative (TN)

Detection rate = (TP+O)/N
Sensitivity = (TP+O)/(TP+O+FN)
PPV = (TP+O)/(TP+FP-O)
Specificity = TN/(TN+FP-O)
NPV = TN/(TN+FN)

19

Effects of Overdiagnosis on Screening Performance

Underlying Truth (no overdx)

Test Result	Disease	No Disease	
Test +	90	180	270
Test -	10	720	730
	100	900	1000

Detection rate = 90/1000
Sensitivity = 90/100 = 90%
PPV = 90/270 = 33%
Specificity = 720/900 = 80%
NPV = 720/730 = 97%

20

Effects of Overdiagnosis on Screening Performance

Underlying Truth (50 cases overdx)

Test Result	Disease	No Disease	
Test +	90	50 + 130	270
Test -	10	720	730
	100	900	1000

Detection rate =

Sensitivity =

PPV =

Specificity =

NPV =

21

Effects of Overdiagnosis on Screening Performance

Underlying Truth (50 cases overdx)

Test Result	Disease	No Disease	
Test +	140	130	270
Test -	10	720	730
	150	850	1000

22

Effects of Overdiagnosis on Screening Performance

Underlying Truth (50 cases overdx)

Test Result	Disease	No Disease	
Test +	140	130	270
Test -	10	720	730
	150	850	1000

↑Detection rate = 140/1000

↑Sensitivity = 140/150 = 93%

↑PPV = 140/270 = 52%

↑Specificity = 720/850 = 85%

NPV = 720/730 = 97%

23

Effects of Overdiagnosis on Outcomes

- ↑detection rate and incidence
- ↑sensitivity of test
- ↑specificity of test
- ↑PPV of test

- Improves stage distribution (as a percentage)
  - Also related to length bias
- Improves case survival
  - Also related to length bias
- Does not decrease population (all cause) mortality

24

24

Summary

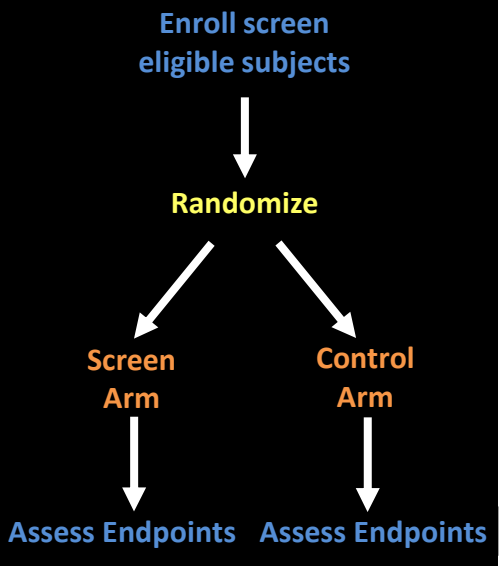
- Screening differs from diagnostic testing
- Potential effectiveness depends on the natural history of disease and treatment effectiveness
- RCT is most valid design, but has limitations
- Survival statistics are inappropriate and biased
- Once a test is shown to reduce mortality, important to measure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

25

25

Randomized Controlled Trial

- Strongest study design
- Randomization evenly distributes the known and unknown confounders
- Groups similar except for screening test under study
- Controls for most selection bias



26

Potential limitations

- No clinical signs or symptoms of disease
  - May need large sample sizes
  - Higher risk or symptomatic individuals may differentially volunteer
- Screen individuals at higher risk for disease?
  - Screening higher risk population reduces RCT sample size
  - Limits generalizability to average-risk individuals
- Willing and able to:
  - Accept randomization, for all rounds in full study
  - Undergo workup and treatment per protocol
  - Be followed for outcomes

27

27

Summary

- Screening differs from diagnostic testing
- Potential effectiveness depends on the natural history of disease and treatment effectiveness
- RCT is most valid design, but has limitations
- Survival statistics are inappropriate and biased
- 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

28

28

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 =>

29

29

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. disease specific mortality
- Still useful to measure along with Disease Specific Mortality
  - May reveal deficiencies in randomization
  - Puts screening in perspective
    - Annual FOBT: 33% ↓ DSM $\Rightarrow$ 1% ↓ overall mortality
  - Helps ensure a major harm (or benefit) is not being missed. Important if test or treatment causes mortality.

30

30

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

Covered in other CTMW presentations

31

31

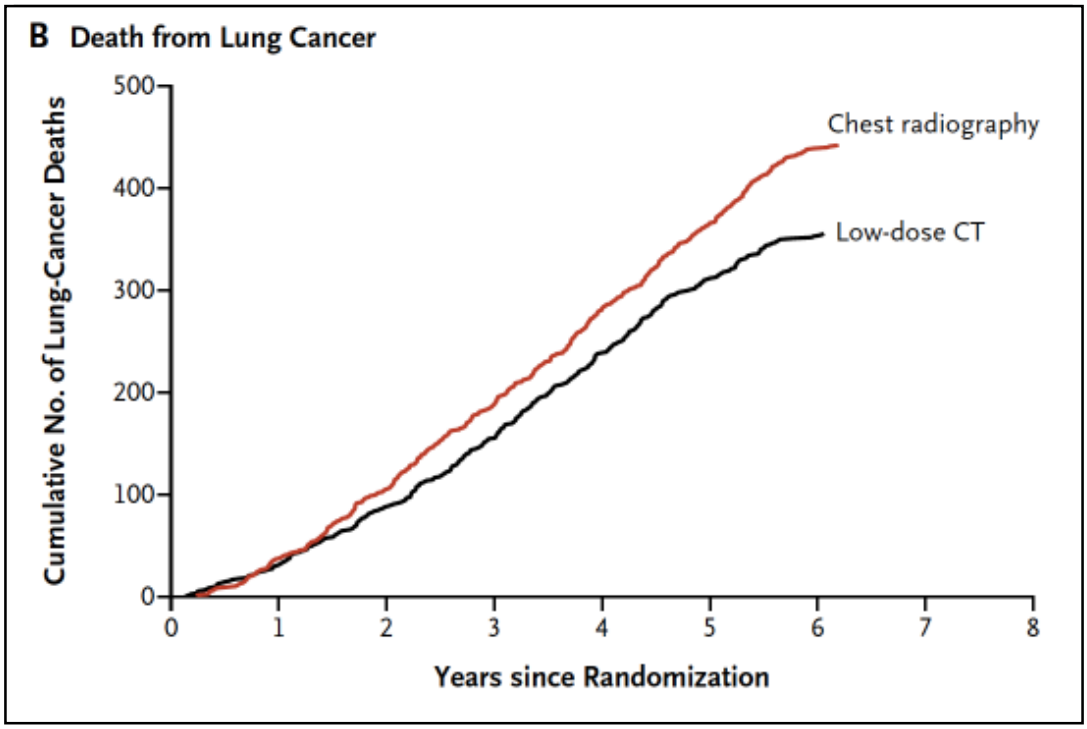
Example: National Lung Screening Trial

- 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

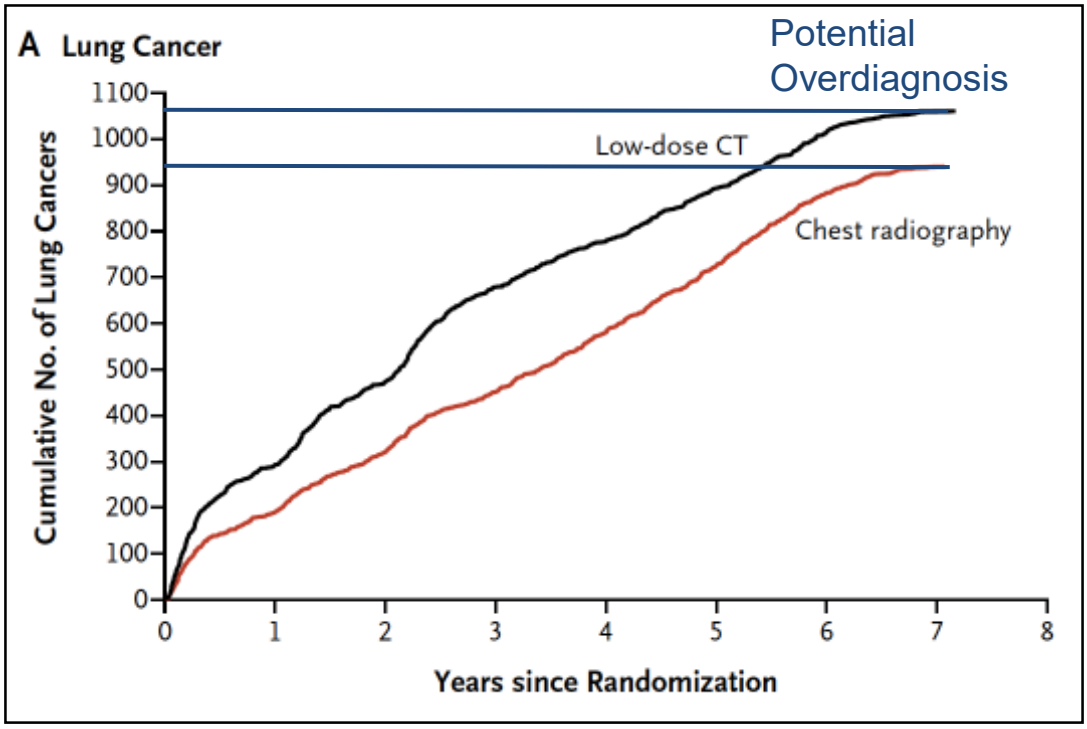
N Engl J Med 2011; 365:395-409

32

32


B Death from Lung Cancer
Cumulative No. of Lung-Cancer Deaths
500
400
300
200
100
0
Chest radiography
Low-dose CT
0
1
2
3
4
5
6
7
8
Years since Randomization

33


A Lung Cancer
Potential Overdiagnosis
Cumulative No. of Lung Cancers
1100
1000
900
800
700
600
500
400
300
200
100
0
Low-dose CT
Chest radiography
0
1
2
3
4
5
6
7
8
Years since Randomization

34

Results: National Lung Screening Trial

Adherence >90%

	Low Dose CT	X-Ray
Recall rate	24.2%	6.9%
False-positive rate	23.3%	6.5%
Cancer rate (per 100,000 PY)	645	572
Lung cancer deaths (per 100,000 PY)	247	309
Risk reduction	20% (95% CI 6.8% to 26.7%, p=0.004)	
All cause mortality risk reduction	6.7% (95% CI 1.2% to 13.6%, p=0.02)	
Other cause mortality risk reduction	3.2% (p=0.28)	

35

35

Results: National Lung Screening Trial

Adherence >90%

	Low Dose CT	X-Ray
Recall rate	24.2%	6.9%
False-positive rate	23.3%	6.5%
Cancer rate (per 100,000 PY)	645	572
Lung cancer deaths (per 100,000 PY)	247	309
Risk reduction	20% (95% CI 6.8% to 26.7%, p=0.004)	
All cause mortality risk reduction	6.7% (95% CI 1.2% to 13.6%, p=0.02)	
Other cause mortality risk reduction	3.2% (p=0.28)	

36

36

Results: National Lung Screening Trial

Adherence >90%

	Low Dose CT	X-Ray
Recall rate	24.2%	6.9%
False-positive rate	23.3%	6.5%
Cancer rate (per 100,000 PY)	645	572
Lung cancer deaths (per 100,000 PY)	247	309
Risk reduction	20% (95% CI 6.8% to 26.7%, p=0.004)	
All cause mortality risk reduction	6.7% (95% CI 1.2% to 13.6%, p=0.02)	
Other cause mortality risk reduction	3.2% (p=0.28)	

37

37

Results: National Lung Screening Trial

Adherence >90%

	Low Dose CT	X-Ray
Recall rate	24.2%	6.9%
False-positive rate	23.3%	6.5%
Cancer rate (per 100,000 PY)	645	572
Lung cancer deaths (per 100,000 PY)	247	309
Risk reduction	20% (95% CI 6.8% to 26.7%, p=0.004)	
All cause mortality risk reduction	6.7% (95% CI 1.2% to 13.6%, p=0.02)	
Other cause mortality risk reduction	3.2% (p=0.28)	

38

38

Observational Studies vs RCT

- Both compare groups receiving different interventions
  - In observational studies, group assignment may be due to
    - Patient or provider factors, or policy changes
  - May be prospective or retrospective
- Strengths
  - Increased generalizability
    - Ability to enroll more diverse populations
    - Larger sample sizes potentially
    - Community and academic settings
- Multiple designs
  - Correlation/Ecological, Case-control, or Cohort

39

39

Bias Limits 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!

40

40

JAMA Internal Medicine | Original Investigation

Performance of Screening Ultrasonography as an Adjunct to Screening Mammography in Women Across the Spectrum of Breast Cancer Risk

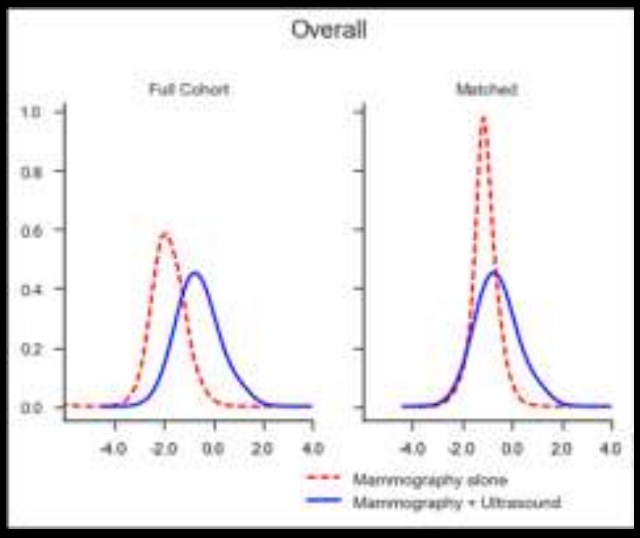
Janie M. Lee, MD, MSc; Robert F. Arao, MPH; Brian L. Sprague, PhD; Karla Kerlikowske, MD; Constance D. Lehman, MD, PhD; Robert A. Smith, PhD; Louise M. Henderson, PhD; Garth H. Rauscher, PhD; Diana L. Miglioretti, PhD

- Observational cohort study
- 6,081 Screening mammography + same day US exams compared with screening mammography alone
- But women receiving mammo + US were
  - Younger, white non-Hispanic
  - Have dense breasts, family history, higher 5-yr risk

JAMA Intern Med. 2019;179(5):658-667

41

41



- Used propensity scores to match Mammo+US exams to Mammo alone exams 1:5
- Kernel density plots provide a visual summary of propensity score distributions

42

42

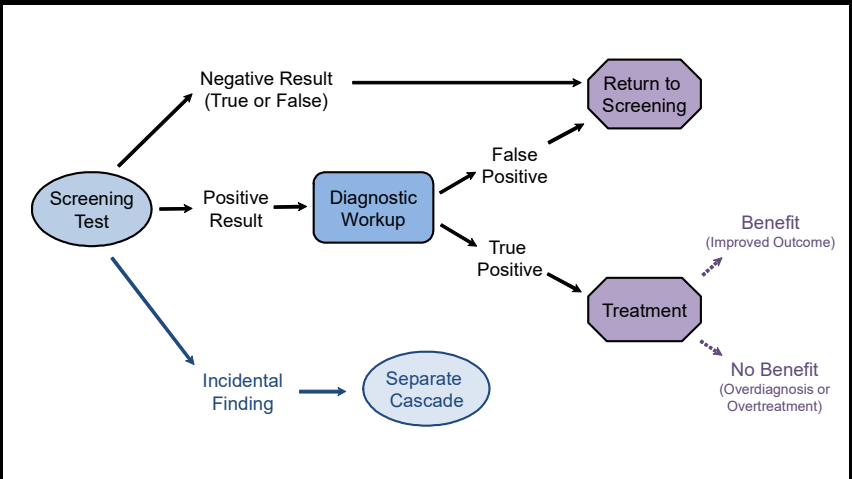
Summary

- Screening differs from diagnostic testing
- Potential effectiveness depends on the natural history of disease and treatment effectiveness
- RCT is most valid design, but has limitations
- Survival statistics are inappropriate and biased
- 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

43

43

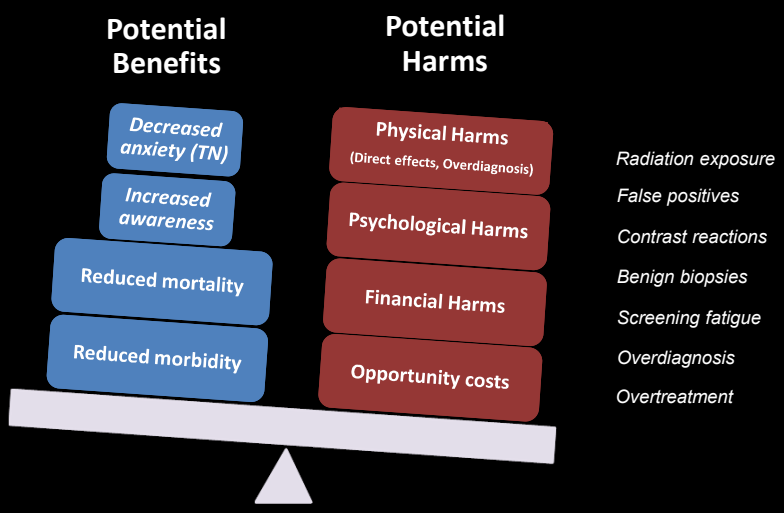
Screening Cascade



44

44

Benefit-Harm Balance of Screening



**"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

45

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

46

46

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

47

47

References

1. Black WC. Randomized clinical trials for cancer screening: rationale and design considerations for imaging tests. Journal of Clinical Oncology, 2006; 24(20): 3252-3260.
2. Black WC. Overdiagnosis: An underrecognized cause of confusion and harm in cancer screening. J Natl Cancer Inst 2000;92(16):1280-2.
3. Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. J Natl Cancer Inst 2002;94(3):167-73.
4. Black WC, Welch HG. Screening for disease. AJR. American Journal of Roentgenology 1997;168(1):3-11.
5. Esserman L, Shieh Y, Thompson I. Rethinking screening for breast cancer and prostate cancer. JAMA. 2009;302(15):1685-92.
6. Fontana RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhn JR, et al. Screening for lung cancer: a critique of the Mayo Lung Project. Cancer 1991;67(suppl):1155-1164.
7. Harris RP, Sheridan SL, Lewis CL, Barclay C, Vu MB, Kistler CE, et al. The Harms of Screening: A Proposed Taxonomy and Application to Lung Cancer Screening. JAMA internal medicine. 2013 Dec 9:-. PubMed PMID: 24322781.
8. Hillman BJ, Black WC, D'Orsi C, Hauser B, Smith R. The appropriateness of employing imaging screening technologies: Report of the methods committee of the ACR task force on screening technologies. JACR 2004;1(11):861-864.
9. Obuchowski NA, Graham RJ, Baker ME, Powell KA. Ten criteria for effective screening: their application to multislice CT screening for pulmonary and colorectal cancers. AJR Am J Roentgenol 2001;176(6):1357-62.
10. Prorok PC, Kramer BS, Gohagan JK. Screening theory and study design: the basics. In: Kramer BS, Gohagan JK, Prorok PC, editors. Cancer screening: theory and practice. New York: Marcel Dekker; 1999. p. 29-53.
11. Wilson, J. M. & Jungner, Y. G. Principles and practice of screening for disease. Public Health Pap. 34, 1–163 (1968).

48

Additional reference

- PM Marcus. Assessment of Cancer Screening: a Primer. November 2019. Available online: https://www.ncbi.nlm.nih.gov/books/NBK550212/

49

49

Thank you!

50

50