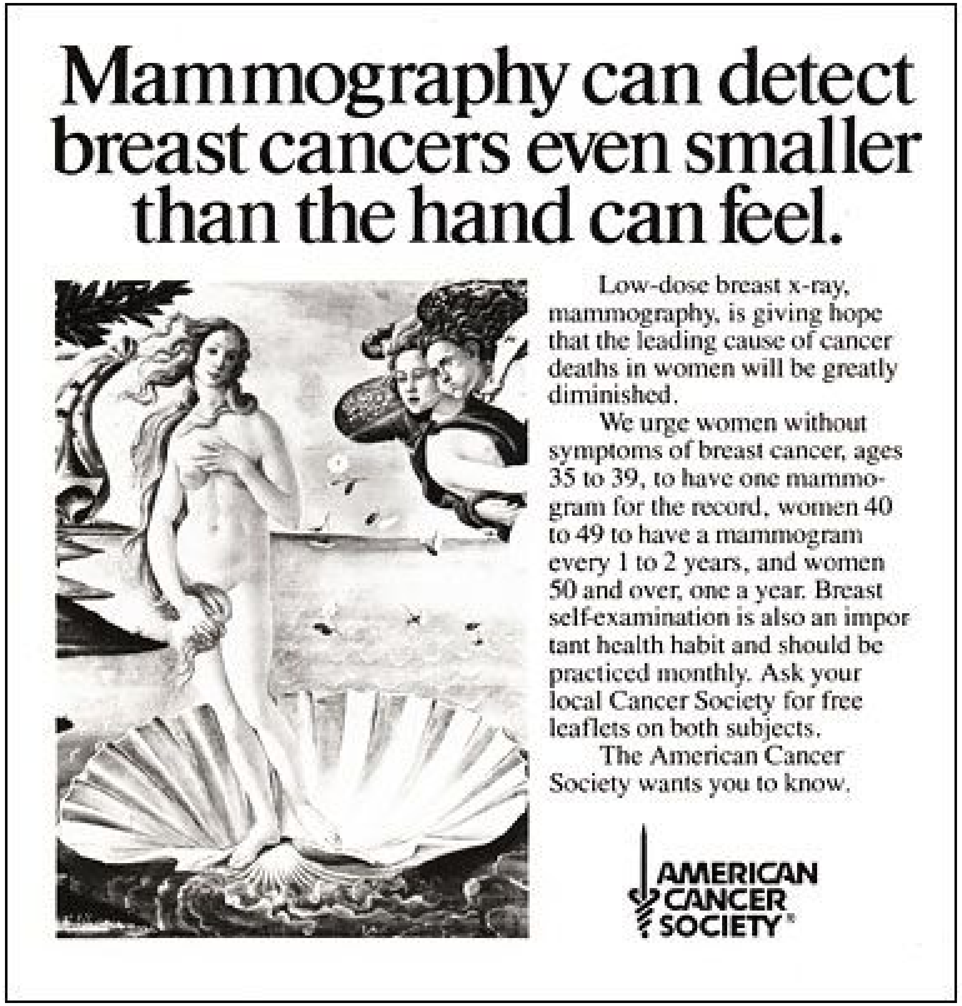
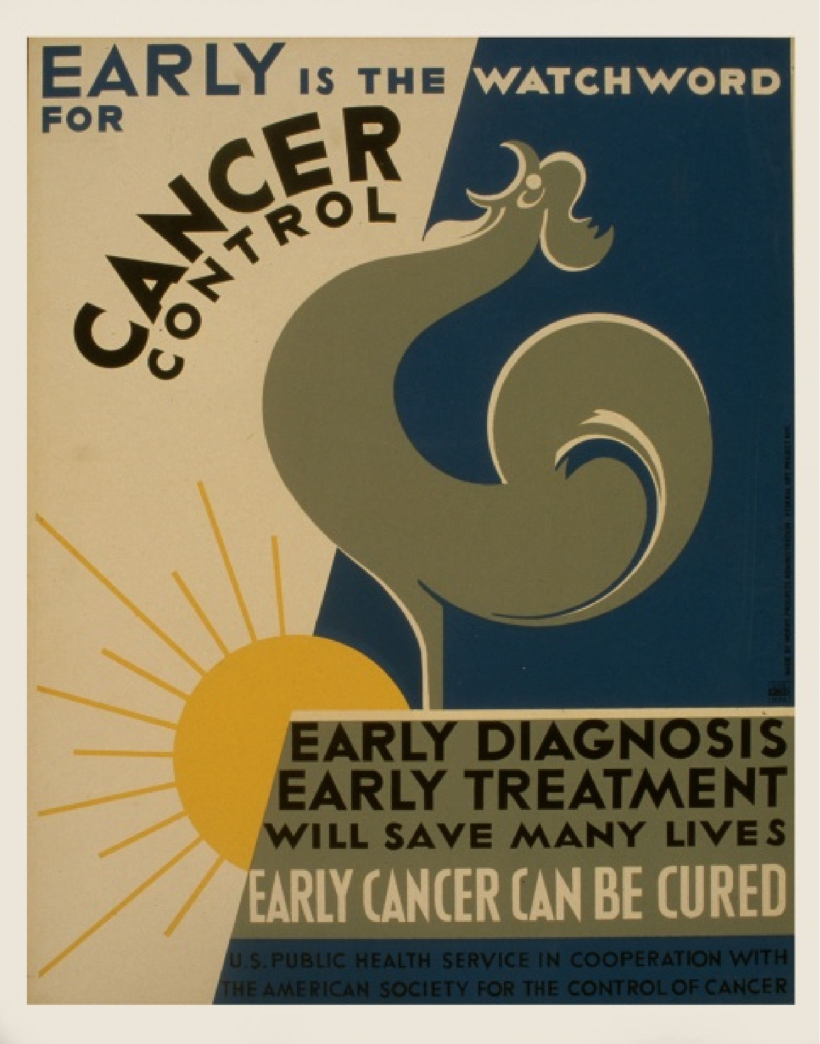
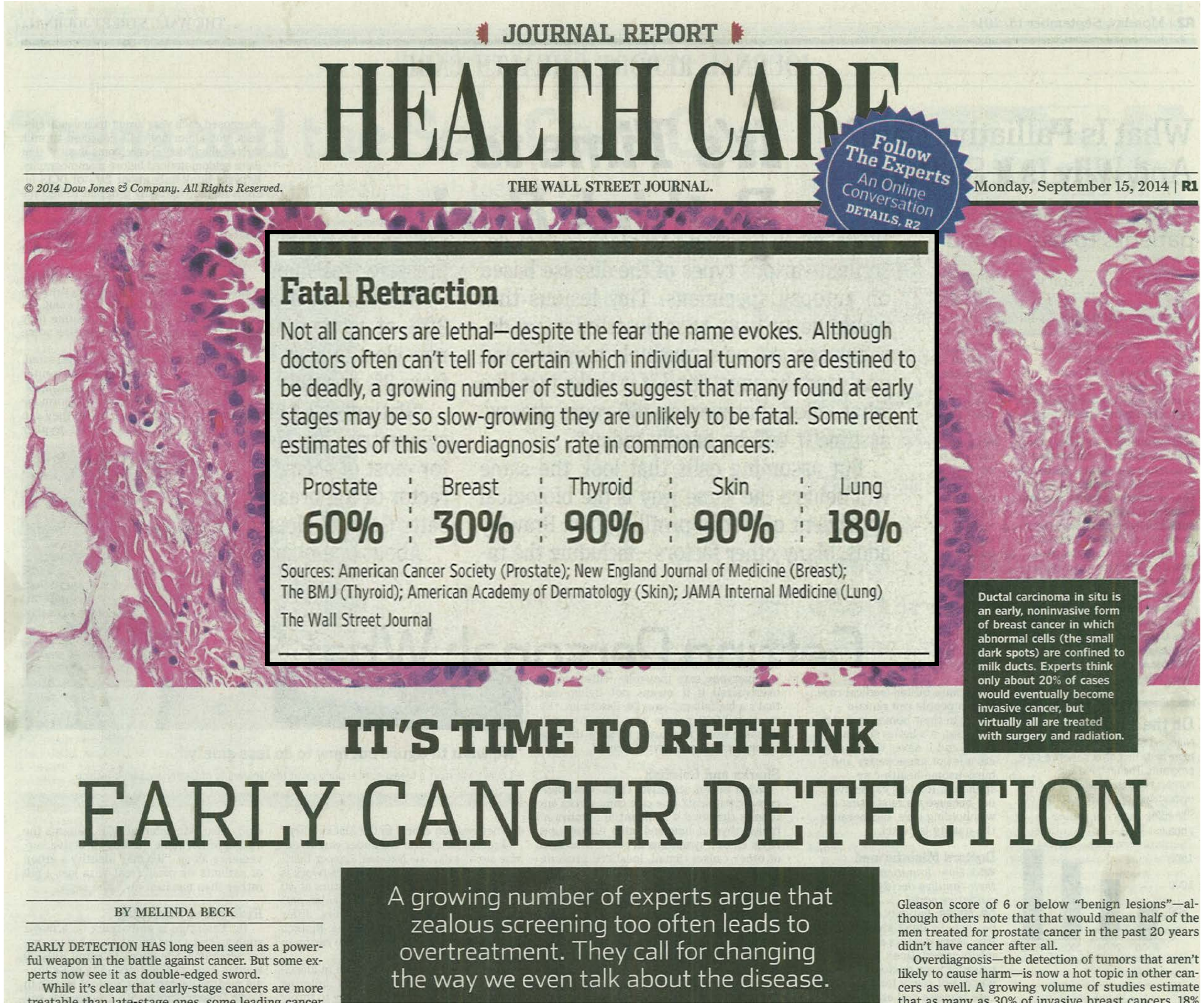
Statistical and Methodological Challenges in Cancer Screening

Ruth Etzioni PhD
Fred Hutchinson Cancer Research Center
CEDAR at the Knight Cancer Institute

Cancer screening in the distant past







JOURNAL REPORT

HEALTH CARE

Follow The Experts
An Online Conversation
DETAILS, R2

THE WALL STREET JOURNAL. Monday, September 15, 2014 | R1

Fatal Retraction

Not all cancers are lethal—despite the fear the name evokes. Although doctors often can't tell for certain which individual tumors are destined to be deadly, a growing number of studies suggest that many found at early stages may be so slow-growing they are unlikely to be fatal. Some recent estimates of this 'overdiagnosis' rate in common cancers:

Prostate	Breast	Thyroid	Skin	Lung
60%	30%	90%	90%	18%

Sources: American Cancer Society (Prostate); New England Journal of Medicine (Breast); The BMJ (Thyroid); American Academy of Dermatology (Skin); JAMA Internal Medicine (Lung)

The Wall Street Journal

Ductal carcinoma in situ is an early, noninvasive form of breast cancer in which abnormal cells (the small dark spots) are confined to milk ducts. Experts think only about 20% of cases would eventually become invasive cancer, but virtually all are treated with surgery and radiation.

IT'S TIME TO RETHINK EARLY CANCER DETECTION

A growing number of experts argue that zealous screening too often leads to overtreatment. They call for changing the way we even talk about the disease.

BY MELINDA BECK

EARLY DETECTION HAS long been seen as a powerful weapon in the battle against cancer. But some experts now see it as double-edged sword.

While it's clear that early-stage cancers are more treatable than late-stage ones, some leading cancer

Gleason score of 6 or below "benign lesions"—although others note that that would mean half of the men treated for prostate cancer in the past 20 years didn't have cancer after all.

Overdiagnosis—the detection of tumors that aren't likely to cause harm—is now a hot topic in other cancers as well. A growing volume of studies estimate that as many as 30% of invasive breast cancers, 18%

September 2014

By Mei-Sing Ong and Kenneth D. Mandl

DOI: 10.1377/hlthaff.2014.1087
HEALTH AFFAIRS 34,
NO. 4 (2015): 576–583

National Expenditure For False-Positive Mammograms And Breast Cancer Overdiagnoses Estimated At $4 Billion A Year

A recent randomized controlled trial of screening mammography in Canada reported an overdiagnosis rate of 22 percent for all screen-detected invasive breast cancer.[4] The US rate of overdiagnosis has been estimated to be 22–31 percent of all breast cancers diagnosed.[2]

MEDPAGE TODAY®

Mammography's $4-Billion Problem

— Millions of women receive false-positive results annually, and 20,000 are overtreated.

Cancer screening today

STAT
[1]

Grail, a deep-pocketed startup, shows 'impressive,' if early, results for cancer blood test

By Matthew Herper[2] @matthewherper[3]

May 31, 2019

A New $500 Blood Test Could Detect Cancer Before Symptoms Develop

Victoria Forster Contributor
Pharma & Healthcare
Cancer research scientist and childhood cancer survivor.

Cancer screening today

- Advances in technology are enabling
  - Discovery of new molecular signatures for early cancer
  - Deployment of highly sensitive methods for their detection
- Knowing that a test can detect cancer is only the first step
  - A good test should have good diagnostic performance but
  - Good performance does not necessarily make a good test!
- Ultimately, the goodness of a test rests on
  - Performance
  - Benefit
  - Harm

These are hard to measure

Statistical and Methodological Challenges

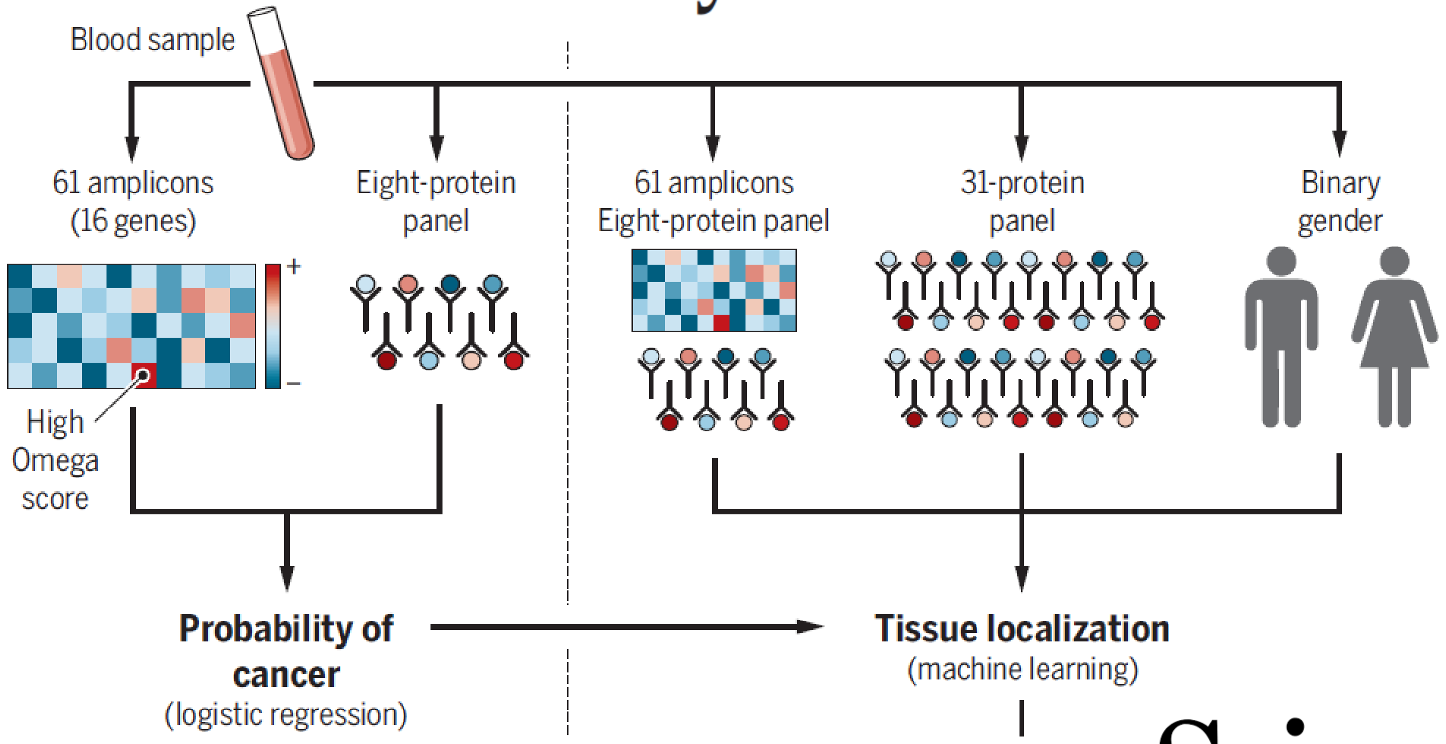
- Measuring diagnostic performance
  - What is the right metric?
  - The importance of disease frequency
- Quantifying benefit
  - What is the right metric? Absolute or relative?
  - The critical importance of time horizon
  - The role of study design
- Assessing harm
  - Why most estimates of overdiagnosis are wrong

Screening development
Focus on performance

CancerSEEK tests were positive in a median of 70% of the eight cancer types. The sensitivities ranged from 69 to 98% for the detection of five cancer types (ovary, liver, stomach, pancreas, and esophagus) for which there are no screening tests available for average-risk individuals

CANCER

Detection and localization of surgically resectable cancers with a multi-analyte blood test

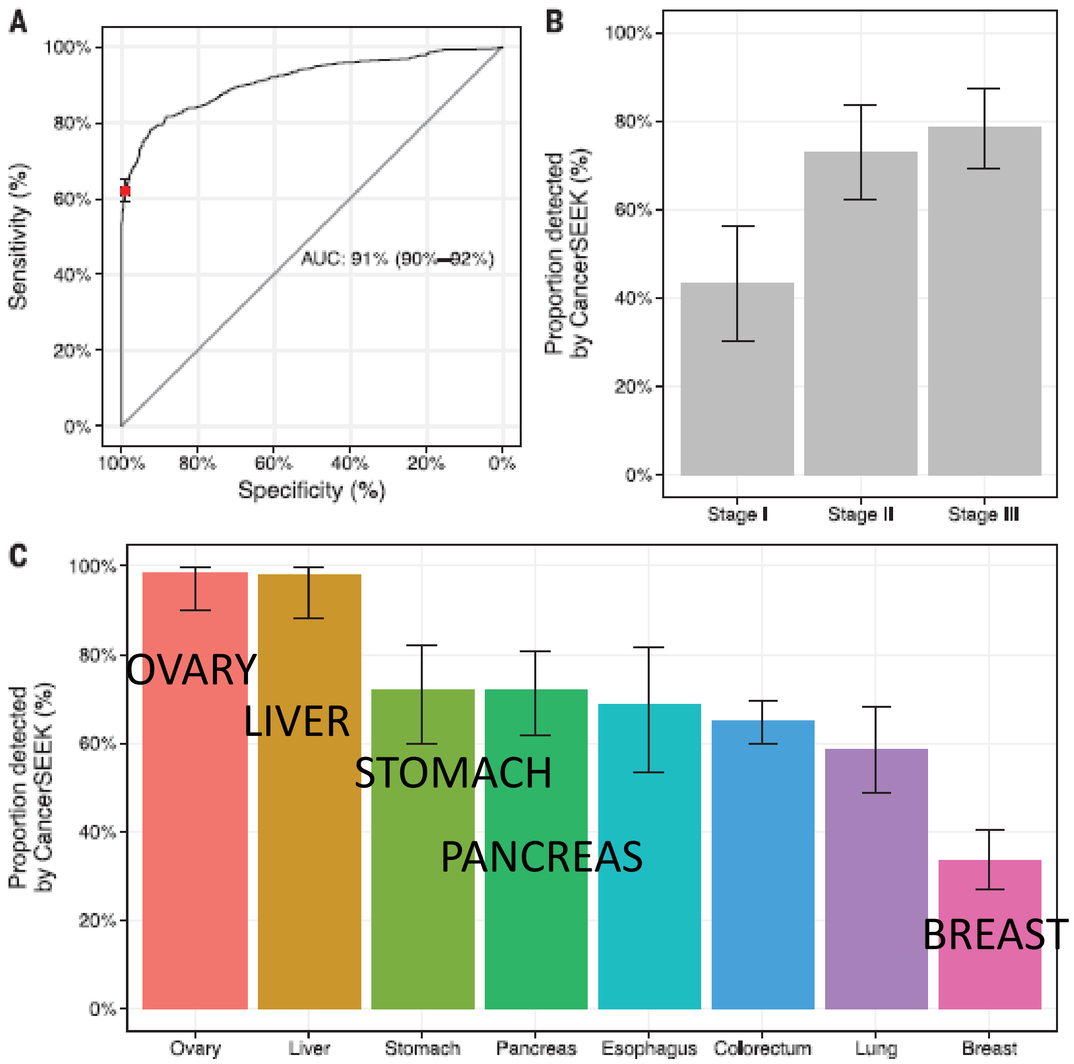


Science

February 23 2018

CancerSEEK performance caveats

- Retrospective study not properly designed to address value for early detection
- Cases had already been diagnosed with cancer – not an early detection setting
- Cases stage I-III; highest specificity in later stage
- Relative frequencies of cases and controls fixed by design
- In prospective screening setting controls far outnumber cases



Establishing performance in prospective settings

Sensitivity or true positive rate

- Requires gold standard test to establish presence of disease
    - Subject to verification bias if not all confirmed
    - Subject to confirmation bias if gold standard not perfect
    - Both biases can lead to overly favorable sensitivity estimates
- Empirical estimates of sensitivity which compare screen and interval cancers measure something quite different!

Specificity or true negative rate

- In prospective setting needs to be almost perfect to avoid many unnecessary confirmation tests
- Even 95% specificity is not adequate in most cancers!

Quantifying benefit

Focus on **disease-specific mortality** not all-cause mortality

- Any single cancer will only cause a small minority of deaths
- Deaths from other causes will swamp those prevented by the screening test
- Consider any increment in other-cause deaths as screening harms

Benefit = reduction in disease-specific mortality

- Relative versus absolute benefit
- The role of the time horizon
- The importance of study design and implementation

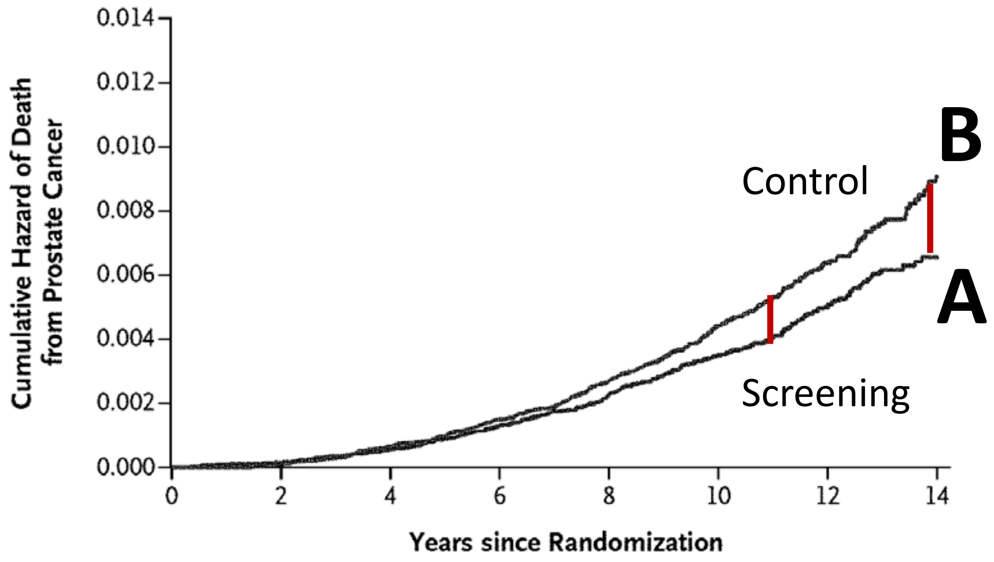
Absolute versus relative benefit

Relative benefit : Deaths in screened group divided by deaths in the control group

$$A/B$$

Absolute benefit: Deaths in the control group minus deaths in the screened group

$$B - A$$



Screening trials: focus on relative benefit

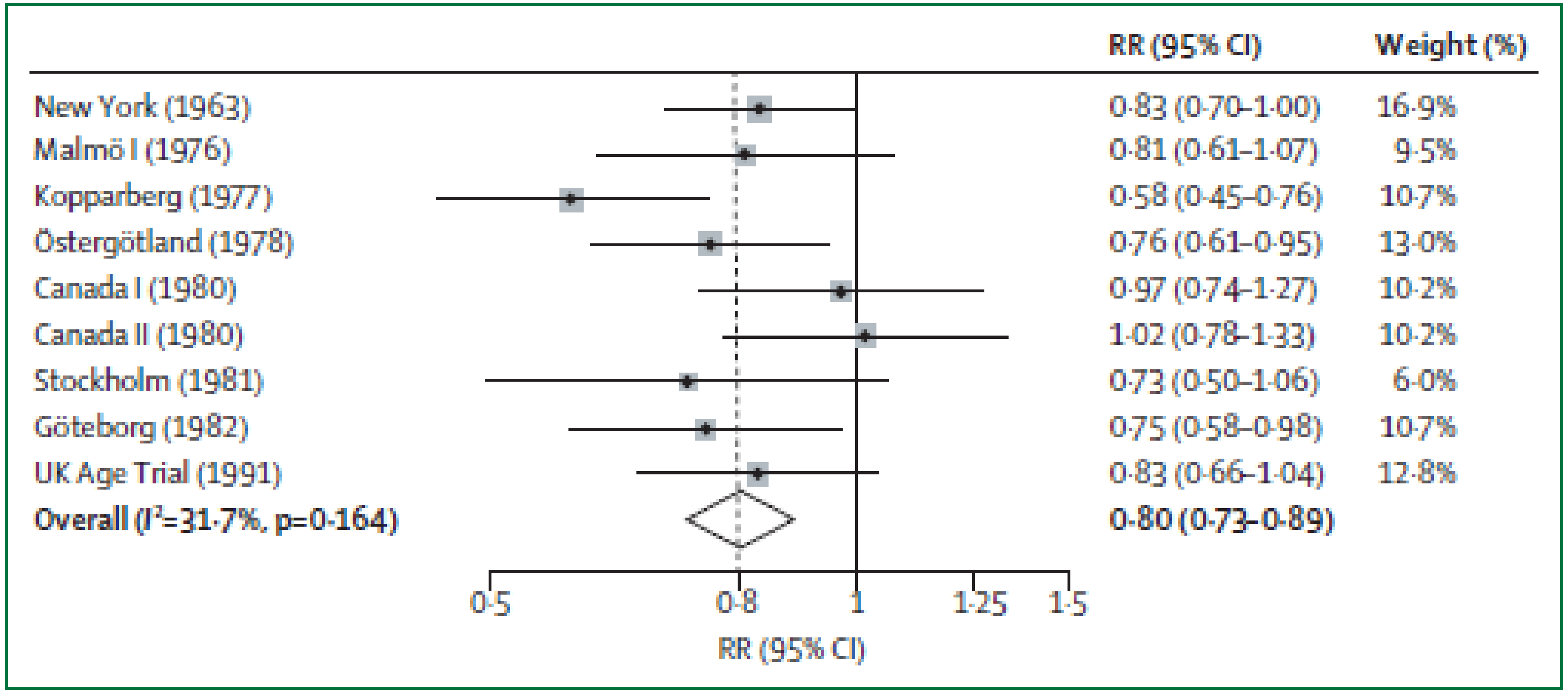


Figure 1: Meta-analysis of breast cancer mortality after 13 years of follow-up in breast cancer screening trials
Adapted from the Cochrane Review.[5] RR=relative risk. Malmö II is excluded because follow-up of about 13 years was not available; the Swedish Two County (Kopparberg and Östergötland) and Canada I and II trials are split into their component parts; the Edinburgh trial is excluded because of severe imbalances between randomised groups. Weights are from random-effects analysis.

Guidelines: focus on absolute benefit

CLINICAL GUIDELINE | Annals of Internal Medicine

Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement

Virginia A. Moyer, MD, PhD, on behalf of the U.S. Preventive Services Task Force*

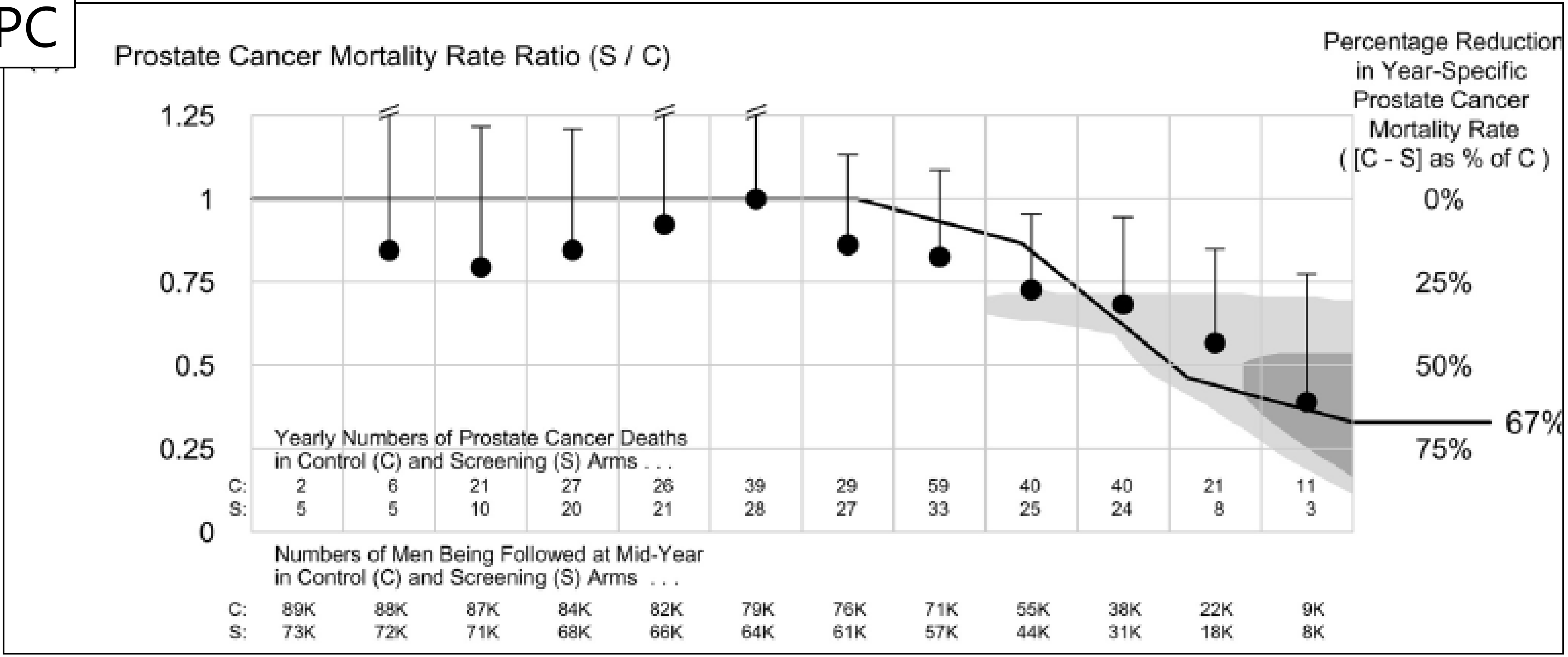
There is adequate evidence that the benefit of PSA screening and early treatment ranges from 0 to 1 **prostate cancer deaths avoided** per 1000 men screened

At the same time, **overdiagnosis and overtreatment** of prostatic tumors that will not progress to cause illness or death are frequent consequences of PSA-based screening. Although about 90% of men are currently treated for PSA-detected prostate cancer in the United States, the vast majority of men who are treated do not have **prostate cancer death prevented** or **lives extended** from that treatment, but are subjected to **significant harms**.

The USPSTF concludes that there is moderate certainty that the **harms** of PSA-based screening for prostate cancer outweigh the **benefits**.

Relative benefit
Depends on time horizon

ERSPC



Hanley, JA: J Med Screen. 2010;17(3):147-51.

UKCTOCS

Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial

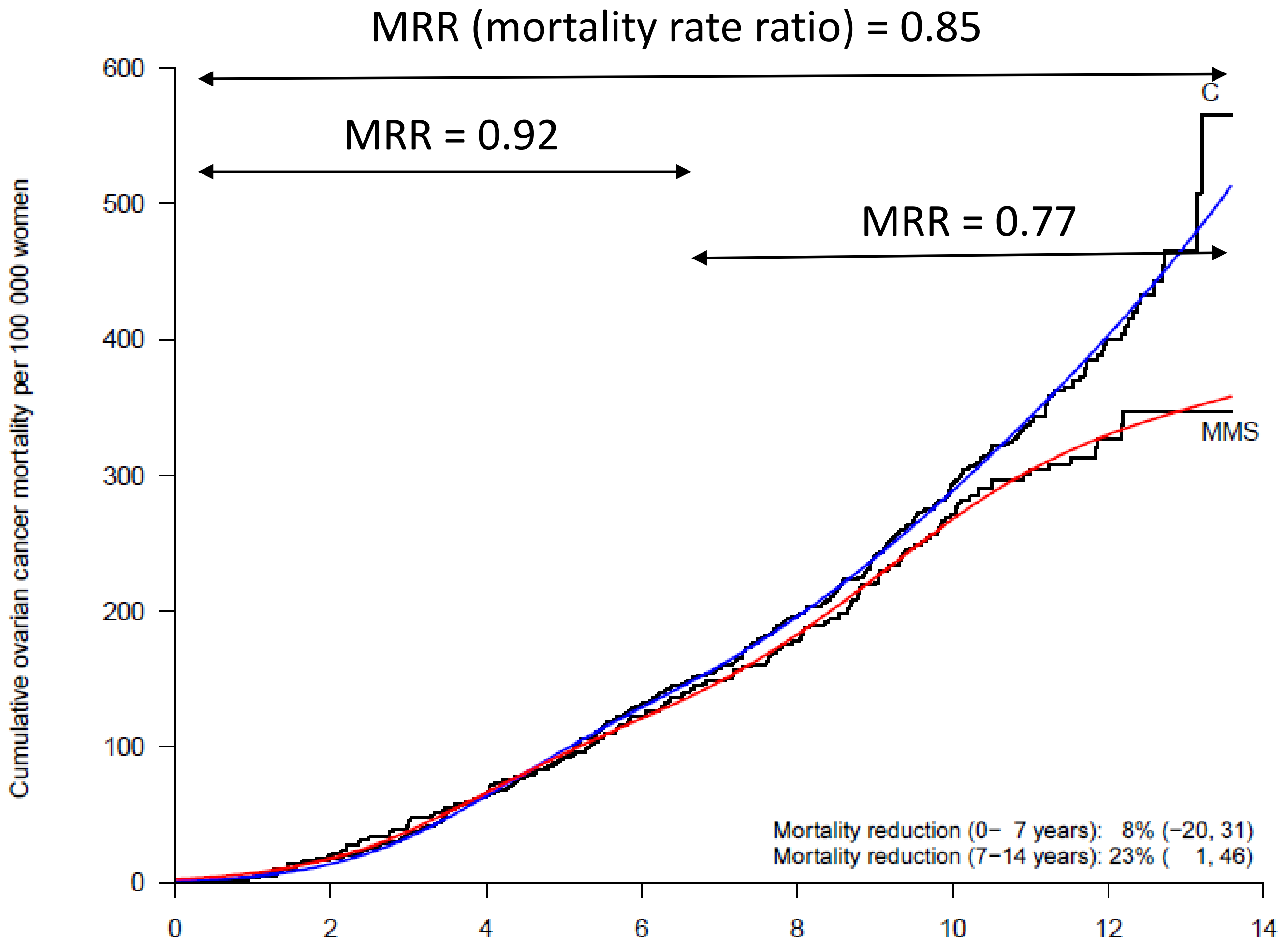
- Began in 2001: N=202,000

- MMS: Multi-modal screening using CA-125
- USS: ultrasound screening

- MMS uses ROCA algorithm – learns by observing serial CA125 values over time

- **MRR=0.85** in MMS arm compared to no screening (p=0.1)

- Royston/Parmar model yields **MRR=0.92** in first 7 years (ns) and **MRR=0.77** in second 7 years (p<0.05)



Jacobs et al, Lancet, 2017

Rethinking benefit estimation

- Relative benefit or mortality rate reduction
  - Usually presented as a single number – average benefit
  - But relative benefit changes over time
- Rethink analysis of screening trials to accommodate MRR shape
  - Early cases are prevalent; enter trial with disease
  - Later cases are incident; transition to disease during the trial
  - Screening is plausibly more effective for later cases
- In UKCTOCS, this phenomenon may be even more exaggerated
  - Precision algorithm for classifying women as screen positive

Methods that accommodate changing MRR

BIOMETRICS 57, 837–843
September 2001

Maximum of the Weighted Kaplan–Meier Tests with Application to Cancer Prevention and Screening Trials

Yu Shen and Jianwen Cai

STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2175–2197

Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects

Patrick Royston[*,†] and Mahesh K. B. Parmar

BIOMETRICS 51, 44–50
March 1995

A Likelihood Ratio Test for Cancer Screening Trials

Steven G. Self[1] and Ruth Etzioni[2]

International Statistical Review (2015), 83, 3, 493–510

A Conditional Approach to Measure Mortality Reductions Due to Cancer Screening

Zhihui (Amy) Liu[1,2], James A. Hanley[3], Olli Saarela[2] and Nandini Dendukuri[3]

Absolute benefit

Dependence on time horizon

The NEW ENGLAND JOURNAL of MEDICINE

Screening and Prostate-Cancer Mortality in a Randomized European Study

0.7 lives saved per 1,000 men screened

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 MARCH 15, 2012 VOL. 366 NO. 11

Prostate-Cancer Mortality at 11 Years of Follow-up

1.1 lives saved per 1,000 men screened

eau European Association of Urology

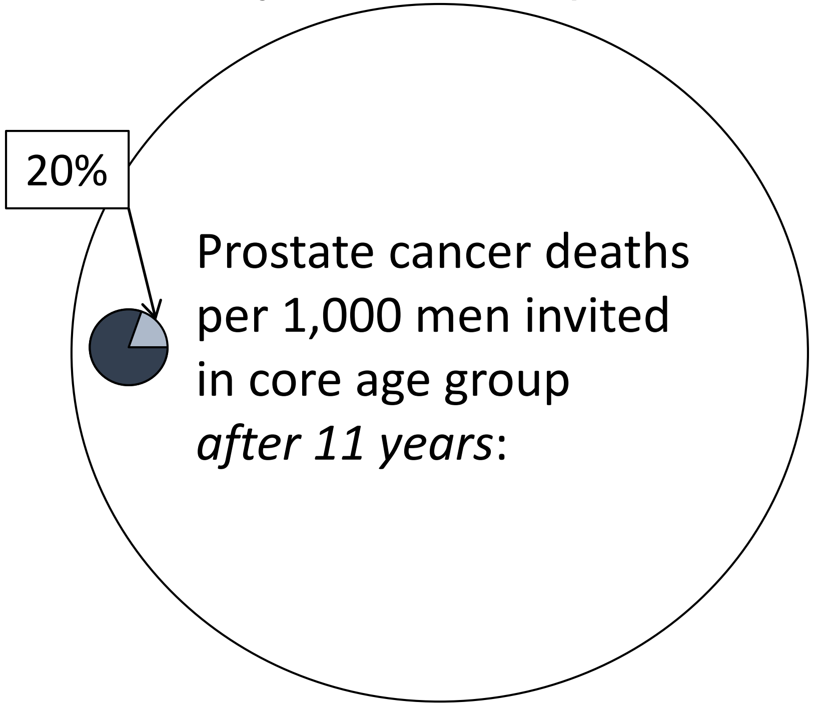
A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer

1.8 lives saved per 1,000 men screened

What is the appropriate horizon?

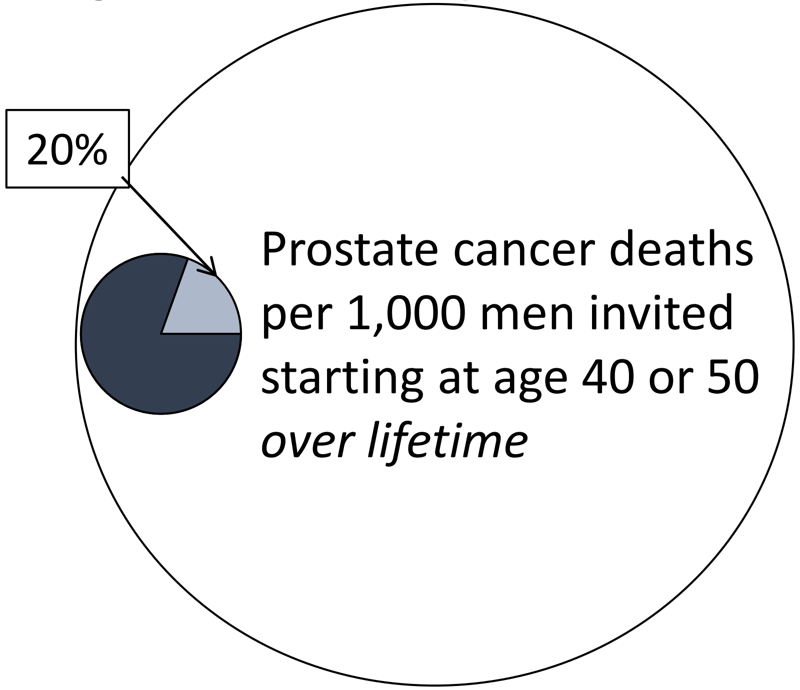
- For policy decisions consider screening lifetime
- Lives saved may be dramatically higher than limited duration trials

11 year follow-up



Trial arm	Deaths
Control	**5**
Screening	**4**
Difference	**1**

Long-term follow-up (SEER)

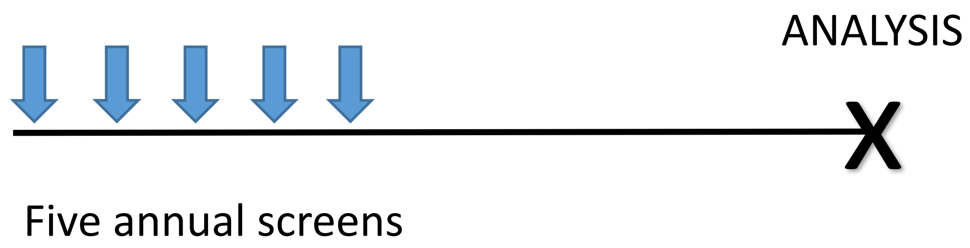


Trial arm	Deaths
Control	**30**
Screening	**24**
Difference	**6**

The importance of study design

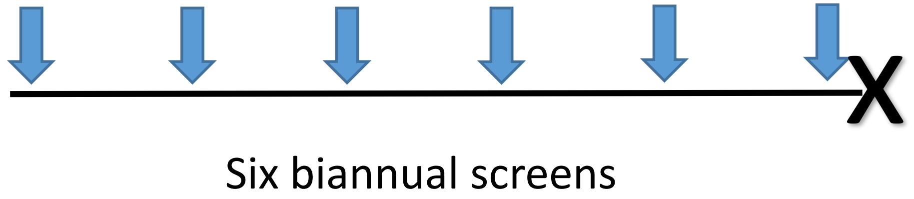
Stop-screen trial

- Screen for several years
- Stop screening and analyze later
- CNBSS breast, PLCO prostate



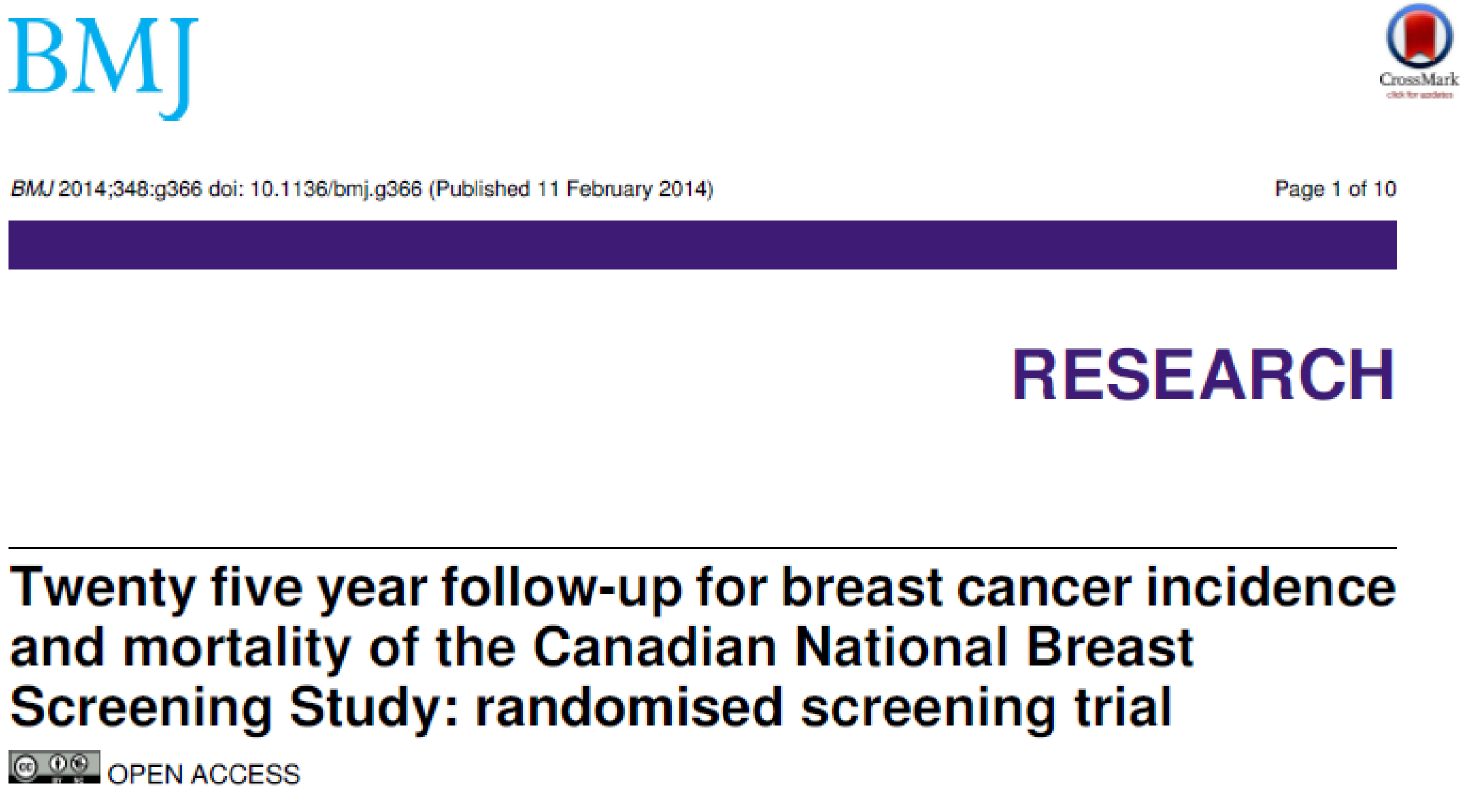
Continued-screen trial

- Screen continuously until analysis
- ERSPC prostate, UKTOCS ovarian



CNBSS

- Stop-screen trial in Canada
- Mamm+CBE vs CBE alone
- CNBSS1: age 40-49
- CNBSS2: age 50-59
- Screen for 5 years
- Follow-up for up to 25 years


BMJ
BMJ 2014;348:g366 doi: 10.1136/bmj.g366 (Published 11 February 2014) Page 1 of 10

RESEARCH

Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial

Miller et al BMJ 2014

Analyzing the CNBSS

Two analyses

1. Compare breast cancer deaths in the two groups over the entire follow-up period
2. Compare breast cancer deaths restricted to cases diagnosed during the screening period

Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial

Analysis options		Screen arm	Control arm
Screening period (5 years)	Cases	666	524
	Deaths (over 25 y)	**180**	**171**
Entire study period (25 years)	Cases	3250	3133
	Deaths (over 25 y)	**500**	**505**

Unfair comparison?
More cases in screened group

Diluted comparison?
More cases in both groups than could have been helped by screening

None of these results were statistically significant

Miller et al BMJ 2014

The New York Times

Vast Study Casts Doubts on Value of Mammograms

By GINA KOLATA FEB. 11, 2014

EMAIL

FACEBOOK

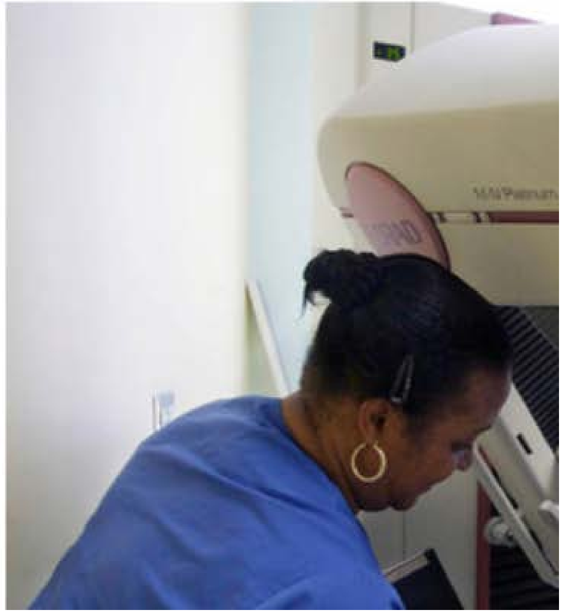
TWITTER

SAVE

MORE

One of the largest and most meticulous studies of mammography ever done, involving 90,000 women and lasting a quarter-century, has added powerful new doubts about the value of the screening test for women of any age.

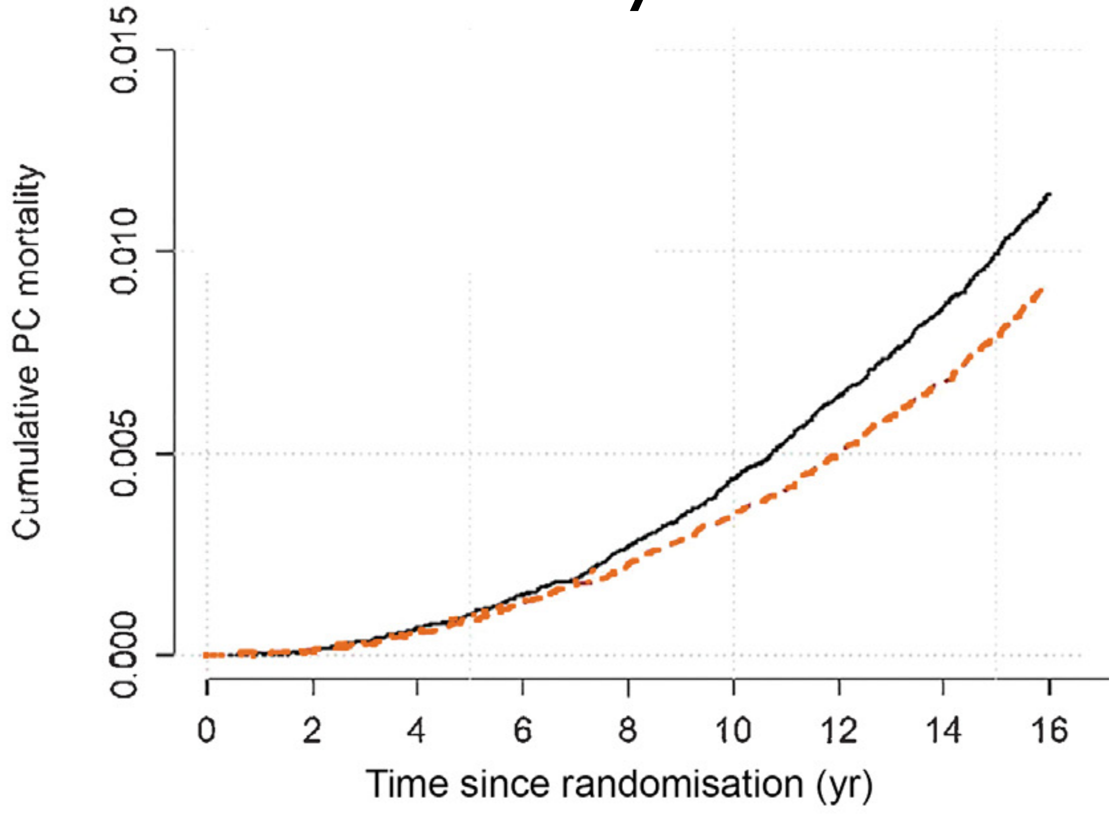
It found that the death rates from breast cancer and from all causes were the same in women who got mammograms and those who did not. And the screening had harms: One in five cancers found with mammography and treated was not a threat to the woman's health and did



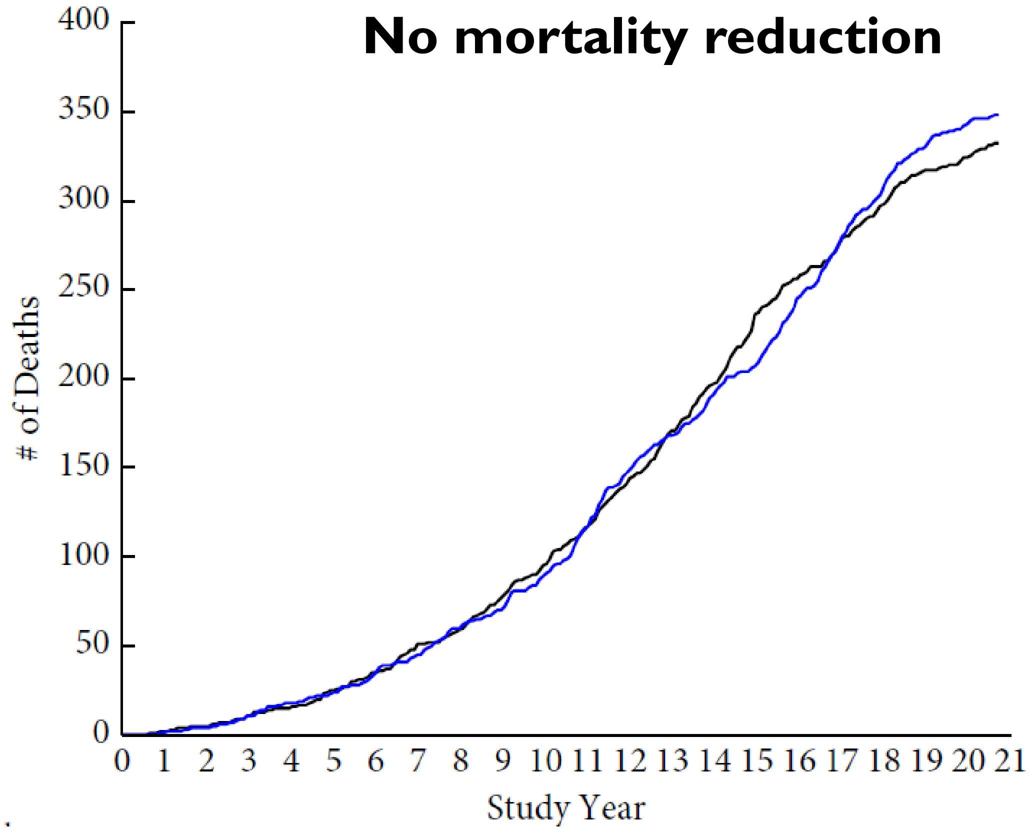
The study as presented is not reflective of the population, time horizon or calendar interval relevant for contemporary breast cancer screening

The importance of study implementation

ERSPC
20% mortality reduction



PLCO
No mortality reduction



ERSPC and PLCO prostate cancer screening trials: cumulative disease mortality

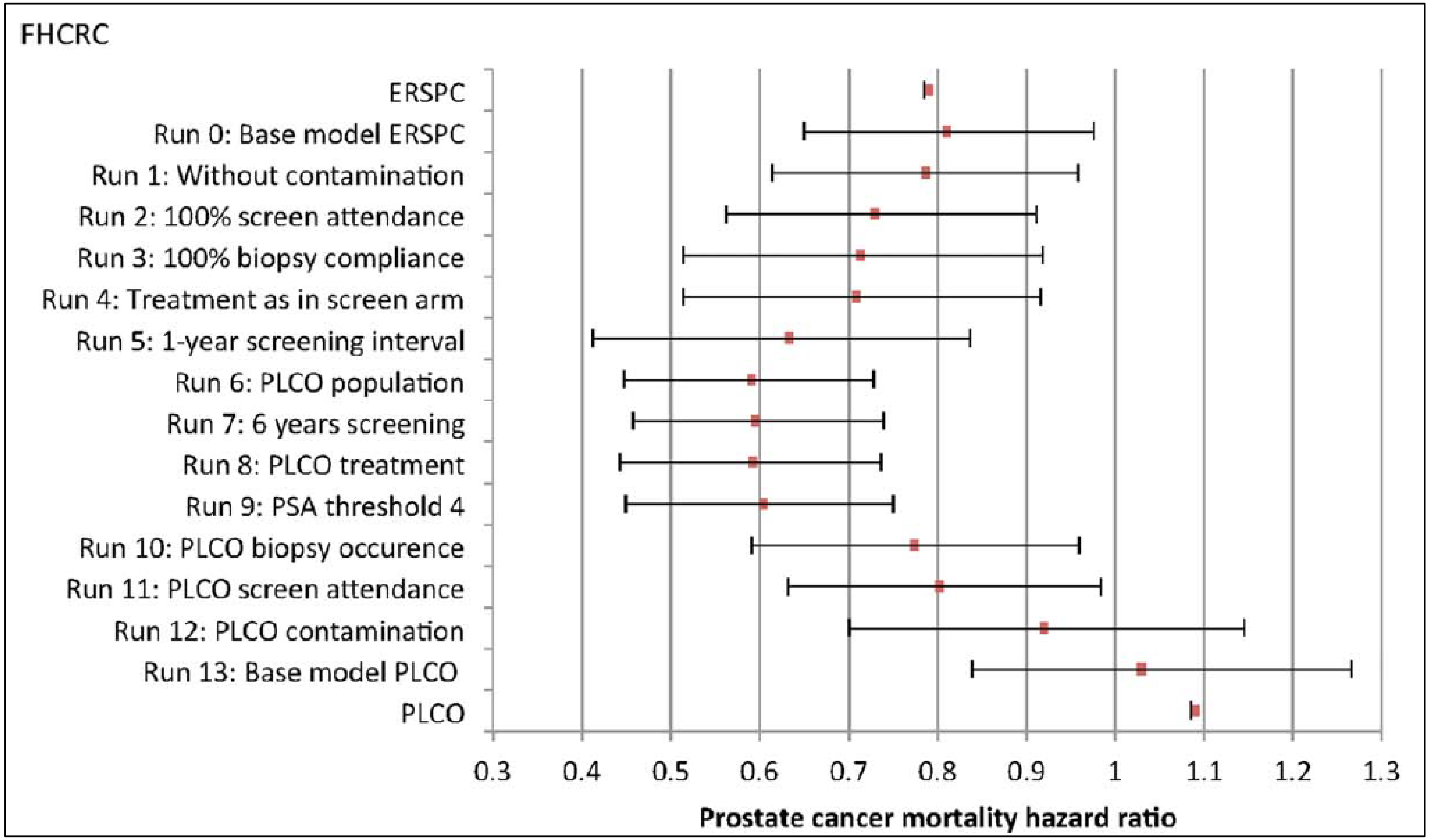
ERSPC and PLCO trials: differences

	ERSPC (n=162,388)	PLCO (n=76,685)
Screening interval	2-4 years	Annual for 6 years
PSA cutoff for biopsy referral	3.0 ng/ml	4.0 ng/ml
Screening on control arm	Infrequent	**74% at least one test** **50% tested each year**
Compliance with biopsy	85%	40%

Schroder et al, Eur Urol, 2010; Otto et al, EJC, 2010

Original Article

The Efficacy of Prostate-Specific Antigen Screening: Impact of Key Components in the ERSPC and PLCO Trials



De Koning et al, Cancer 2017

Annals of Internal Medicine®

LATEST ISSUES CHANNELS CME/MOC IN THE CLINIC JOURNAL CLUB WEB EXCLUSIVES AUTHOR INFO

THIS ISSUE | NEXT ARTICLE ›

ORIGINAL RESEARCH | 3 OCTOBER 2017

Reconciling the Effects of Screening on Prostate Cancer Mortality in the ERSPC and PLCO Trials

Amounts to 25-32% reduction in mortality on screened arms relative to no screening

No evidence that this effect differs across trials

Tsodikov et al, Ann Int Med, 2017

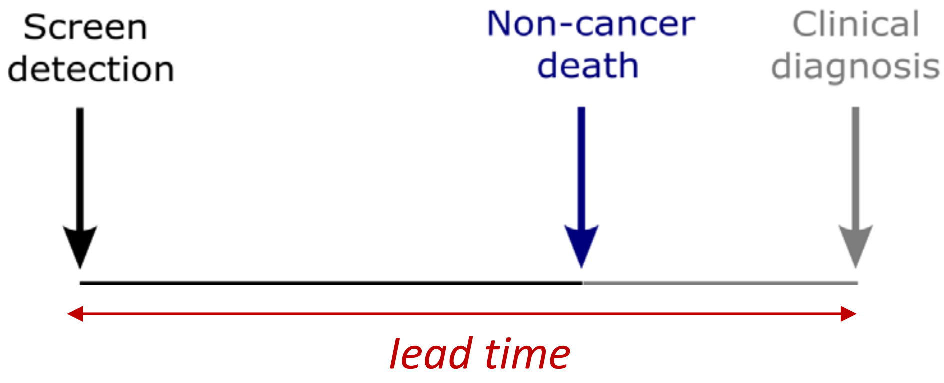
Overdiagnosis

Lead time

- Time by which diagnosis is advanced by screening

Overdiagnosis

- When screening detects a cancer that would not have been detected without screening
- When other-cause death happens during the lead time



Overdiagnosed case

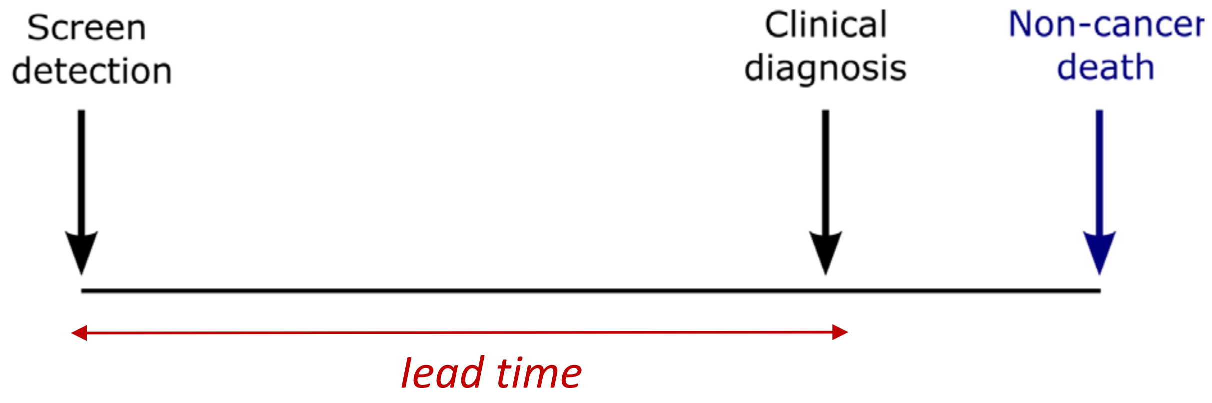
Overdiagnosis

Lead time

- Time by which diagnosis is advanced by screening

Overdiagnosis

- When screening detects a cancer that would not have been detected without screening
- When other-cause death happens during the lead time



Non-overdiagnosed case

Estimating overdiagnosis from screening trials

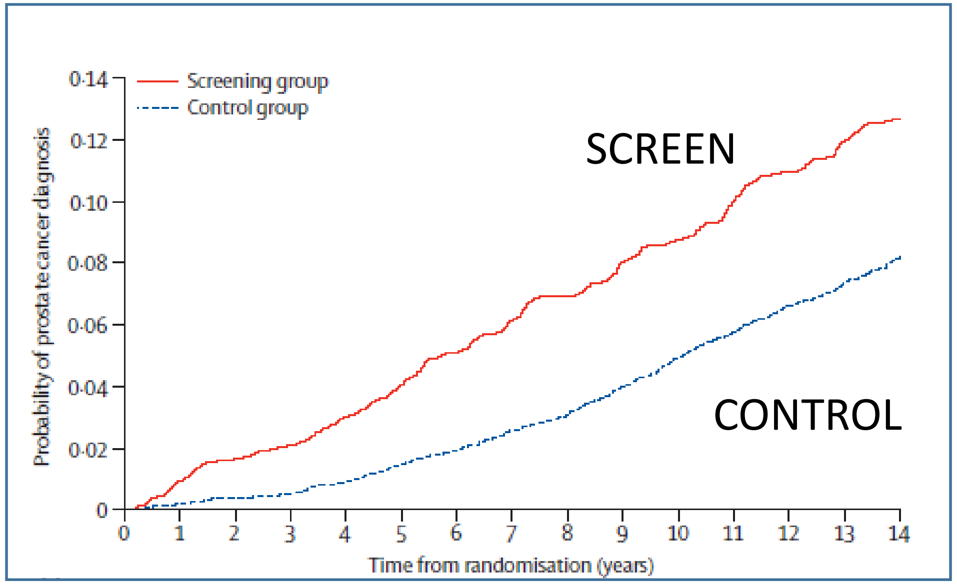
- An overdiagnosed case is an excess case detected by screening
  - Would not have presented clinically in the absence of screening
  - We cannot tell whether any individual screen-detected cases is overdiagnosed
- Control group represents incidence in absence of screening
- Excess-incidence approach to estimating overdiagnosis
  - Incidence on screened arm minus incidence on control arm
  - This idea has also been applied in population settings

ERSPC

Screening and Prostate-Cancer Mortality in a Randomized European Study

Schroder et al NEJM 2009

Prostate cancer incidence in ERSPC



	Cumulative Incidence at 9 years
Screened arm	**8.2%**
Control arm	4.8%
Excess	**8.2% - 4.8% = 3.4%**
Excess/screen-detected	**3.4/5.8 = 58%**

Schroder et al
NEJM 2009

The problem with cumulative excess incidence in ERSPC

- What we know

Cases detected under screening

Represent cases that would have arisen during AND after the trial

Corresponding cases in the absence of screening

- What we do

Take cases detected under screening

Subtract the cases on the control group that arose <u>during</u> the trial

Corresponding cases in the absence of screening

If there is no overdiagnosis this approach will still yield a positive result!

Excess incidence from a stop-screen trial

- What we know

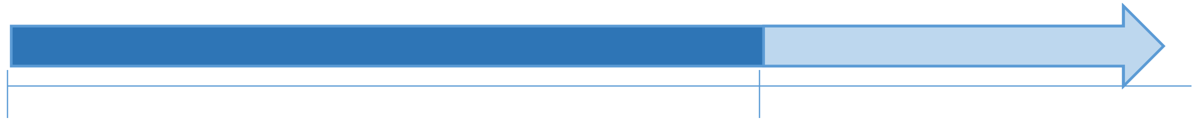
Cases detected under screening

Represent cases that would have arisen during AND after the trial

Corresponding cases in the absence of screening

- In a stop-screen trial

Take cases detected under screening plus an appropriate post-screening interval



Subtract the cases on the control group over the same interval

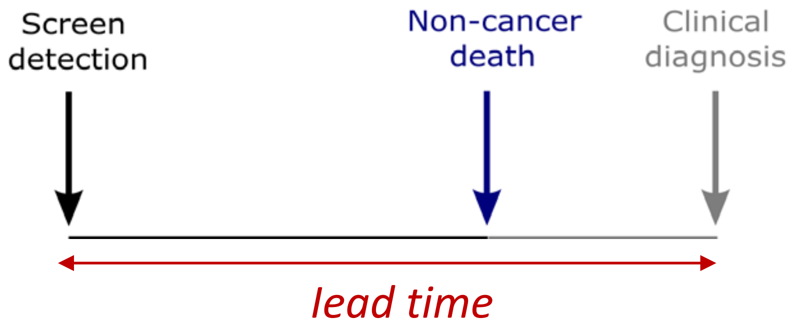
Corresponding cases in the absence of screening

If there is no overdiagnosis excess incidence will be zero with adequate followup

An alternative to learn about overdiagnosis

Learn about lead time via statistical modeling of excess incidence

- Overdiagnosis: chance that lead time is less than other-cause survival



Need

- Rich screen- and interval incidence within screened population
- Assumptions to ensure estimability of lead time distribution

Identification of the Fraction of Indolent Tumors and Associated Overdiagnosis in Breast Cancer Screening Trials

Ryser et al, 2019

Summary and Wrap-up

Our main methodological and statistical challenge

- Accurately quantifying population-relevant benefits and harms

Must recognize limitations of

- Studies and measures of performance
- Simple, empirical measures of benefit and harm

These limitations apply as much if not more to novel screening tests

Careful development and application of cancer models can be very helpful to interpret empirical results and to go beyond the data

Thank you!

Acknowledgements

- Roman Gulati
- Lurdes Inoue
- Jane Lange
- Marc Ryser
- Rocky Feuer
- Angela Mariotto
- CISNET prostate cancer group