Pamela M. Marcus

Assessment of cancer screening: a primer

Last Updated: November 2019

National Cancer Institute (US)
Bethesda (MD)

National Cancer Institute (US), Bethesda (MD)

NLM Citation: Marcus PM. Assessment of cancer screening: a primer [Internet]. Bethesda (MD): National Cancer Institute (US); 2019 Nov.

Cancer screening is a prominent strategy in cancer control in the United States, yet the ability to correctly interpret cancer screening data seems to elude many researchers, clinicians, and policy makers. Assessment of Cancer Screening: A Primer aims to rectify that situation by teaching readers, in simple language and with straightforward examples, why and how the population-level cancer burden changes when screening is implemented, and how we assess whether that change is of benefit. The book provides an in-depth look at the many aspects of cancer screening and its assessment, including screening phenomena, performance measures, population-level outcomes, research designs, and other important and timely topics. Assessment of Cancer Screening: A Primer is best suited to those with education or experience in clinical research or public health in the United States. No previous knowledge of cancer screening assessment is necessary. Assessment of Cancer Screening: A Primer is the first book dedicated to cancer screening theory and methodology to be published in 20 years. It fills a serious gap in the medical literature: it is a short, accessible, and focused reference on cancer screening as it occurs in the United States.

Author

Pamela M. Marcus, PhD, MS
National Cancer Institute
Corresponding authormarcusp@mail.nih.gov

Table of Contents

Acknowledgements

I have had the opportunity to work with and learn from hundreds of people during my many years as a cancer epidemiologist. To thank everyone by name is impossible. Please know that I learn something from everyone with whom I speak, and that I am grateful to all who have approached me with questions as well as arguments.

Polly Newcomb was the first epidemiologist in my life. Without her support and encouragement at the beginning of my career, it is unlikely that I would have been afforded many opportunities, including the chance to learn about cancer screening in such detail. Many thanks go to my leaders at the National Cancer Institute; they gave me the time, support, and space to bring this primer to fruition. And thanks to my students and mentees: whether they believe it or not, I learn more from them than they learn from me.

Thanks also are due to the friends and colleagues who kept me on schedule, reviewed drafts of the primer, and provided invaluable input. Danielle Durham, John Gohagan, Beth Hoffman, Tony Miller, Phil Prorok, Amy Sayle, and Mark Schiffman were reviewers of a nearly complete draft. Jennifer Croswell, V. Paul Doria-Rose, Richard Fagerstrom, Beth Hoffman, and Robert Hoffman reviewed the complete draft. Rachel Eisenger-Baskin and Anne Julian assisted with reference formatting and copy editing. Stacy Lathrop and Diana Jordon of the National Library of Medicine supported me in this venture.

This primer is dedicated to the memories of Robert Craft Millikan and Seymour Marcus.

Preface

Cancer screening is a prominent strategy in cancer control, yet the ability to correctly interpret cancer screening data seems to elude many researchers, clinicians, and policy makers. My initial attempt to address that problem was to develop a short course on the assessment of cancer screening that focused on methodology and data interpretation. I first taught the course during the 2015 spring semester at the Foundation for Advanced Education in the Sciences, informally known as the National Institutes of Health Graduate School. As the semester went on, it became clear to me that my students and the larger community needed a text. I chose to call the text a primer as it covers the basics and is written as simply as possible.

The primer reflects how cancer screening is perceived and practiced in the United States. It is best suited to those familiar with biomedical research and public health practice in the US. I expect it to be of use regardless of the reader's cancer screening knowledge. Readers less familiar with the topic will want to start at the beginning and read straight through. Those with some experience may be able to read only the sections in which they are interested, or consult the primer for a formula or definition. Please note that the primer does not provide an assessment of the evidence available for or against screening for specific cancers, except as relevant for the purpose of example.

I encourage feedback and can be reached at marcusp@mail.nih.gov.

Pamela Marcus

Bethesda, MD

October 2019

Chapter 1. Foundations

The ability to understand cancer screening data does not require an extensive background in biostatistics, biology, or oncology. Rather, it requires clear thinking, an open mind, and knowledge of a small set of foundational concepts. Those concepts are presented in this chapter.

Cancer

The United States (US) National Cancer Institute's (NCI) webpage, "What is Cancer?," provides an overview of many biomedical aspects of cancer, including its definition, how it arises, and how it progresses (1). The webpage is a great resource for those who are starting out in cancer research. In the next two paragraphs, I summarize relevant topics from the webpage.

Cancer is a complex disease (2), but for the purpose of this primer, it is sufficient to conceptualize it using its most notable features: abnormal cells whose division is usually unchecked. Tumors are collections of those cells. Tumors are classified by their ability to metastasize, that is, the ability of their cells to spread to other regions of the body. Tumors that do not and never will have metastatic potential are called benign, though they can kill by growing large enough to interfere with the proper functioning of organs. Tumors that can or have metastasized are called malignant. Malignant tumors are said to be invasive because they have broken through the basement membrane, the barrier structure on which those cells normally sit. The disruption of that membrane allows cells to utilize the circulatory or lymph systems as routes to spread. Precancer refers to cells that have not broken through the basement membrane but are abnormal in some way that suggests they could break through in the future given the right (though generally unknown) circumstances. The terms precursor, pre-invasive, and pre-malignant sometimes are used instead, but precancer will used in this primer. Strictly speaking, the word cancer (minus the prefix) refers only to malignant tumors and will be used as such in this primer. Be aware, however, that the word cancer often is used in conjunction with precancer. For example, cervical cancer screening rarely leads to the detection of malignant disease; instead, it usually leads to the detection of early cellular changes that are consistent with our understanding of the natural history of cervical cancer.

Cancer is not one disease; it is many diseases. Cancer behavior differs, for example, by and within organ site, by the type of cell that gave rise to the tumor, and by DNA mutations found in the tumor cells. Treatment and prognosis often vary by these characteristics. In the past it was assumed that all cancer would be fatal if left untreated, but we know now that some tumor types regress, stall, or grow so slowly that they are of no clinical relevance.

It is expected that about 1.8 million people in the US will be diagnosed with cancer in 2019, and about 607,000 will die of the disease (3). A little under half the deaths will be due to cancer of the lung and bronchus (143,000), colorectum (51,000), female breast (42,000), prostate (32,000), and cervix uteri (typically referred to as cervix; 4,300). Cancer screening activity in the US is focused on those five organ sites, although screening for other organ sites does occur, often in high-risk populations.

Cancer statistics

The first step in characterizing the extent of any public health problem is to collect data. In the US, our go-to source for cancer data is the Surveillance, Epidemiology and End Results Program, known world-wide simply as SEER (4). SEER was established by the 1971 National Cancer Act (5) and has provided authoritative data on US cancer incidence, survival, and mortality for the years 1975 and later. SEER collects data on every cancer in 19 geographic areas, covering about 34% of the US population (6). SEER data are available in both summary and raw form (7-9).

Cancer data also are collected through the National Program of Cancer Registries, which was established by the Centers for Disease Control and Prevention (CDC) in 1992. Through this program, high-quality cancer registry

data has been collected for 97% of the US population and Puerto Rico, the US Pacific Island Jurisdictions, and the US Virgin Islands (10).

Cancer screening

Cancer screening refers to routine, periodic testing for signs of cancer among individuals who have no symptoms. It is a form of secondary prevention. In the context of cancer screening, the goal of secondary prevention is to improve outcomes by shifting stage at diagnosis to one that is less advanced and deleterious, relative to what occurs in the absence of cancer screening.

Cancer screening is a sorting process. Screenees are sorted into two groups: those with a negative test and those with a positive test. A negative test finds nothing suspicious for cancer and does not require additional medical attention. A positive test reveals something that is suspicious for cancer or with unknown significance regarding cancer; it requires additional medical attention, referred to as diagnostic evaluation. That process is intended to definitively determine whether cancer is or is not present, but in practice can range from active surveillance to the removal of an abnormality. Active surveillance (sometimes called watchful waiting) refers to a schedule of minimally- or non-invasive testing to monitor for clinically important changes. Resection of an abnormality is considered diagnostic evaluation rather than treatment if a definitive diagnosis has not yet been made or cannot be made otherwise.

Cancer screening is not intended in and of itself to provide a definitive diagnosis. Its intent is to identify abnormal medical conditions, such as growths, occult blood, or a biomarker that may suggest cancer. Cancer screening aims to lead to the detection of cancers whose prognosis will improve with earlier detection, and it needs to lead to the detection of enough of those cancers to make screening a worthwhile public health activity. Cancer screening is neither intended to nor is able to lead to detection of every cancer, as the natural history of cancer is erratic, technology has limitations, and frequent screening is impractical.

In the United States, lung and prostate cancer screening tend to detect invasive cancer and not precancer. Screening for colorectal and breast cancer leads to the detection of invasive cancer and precancer. Cervical cancer screening leads to the detection of precancer, certain human papilloma virus (HPV) infections (the causal agent), and on occasion invasive cancers. Cervical cancer screening also can detect cellular changes that occur very early in the cancer process. Those abnormalities are classified as precancer in this primer.

The reader may come across the phrases early detection and early diagnosis in discussions of cancer screening and wonder how the two differ. Early diagnosis refers to a strategy of symptom awareness to lead to a change in the time of diagnosis. The phrases symptom-aware detection and symptom-vigilant detection are more descriptive than early diagnosis but are rarely used. Early detection comprises early diagnosis and screening. Other phrases that can be confusing are cancer prevention screening and early detection screening. Cancer prevention screening refers to cancer screening that leads to the detection of precancer, and early detection screening refers to cancer screening that leads to the detection of invasive cancer.

Principles of early diagnosis will not be discussed in this primer. The remainder of this primer, with the exception of Chapter 8, is written for the assessment of early detection screening, though the material is equally applicable to cancer prevention screening in nearly all instances. Any material that is not is noted as such.

Population-based cancer screening

Population-based cancer screening refers to a cancer control practice in which all individuals who meet certain minimal criteria can choose to receive cancer screening. The term population-based is intended to connote that nearly everyone – that is, almost the entire eligible population – is targeted for cancer screening. Sometimes the phrase mass cancer screening is used instead.

In population-based cancer screening, individuals who are eligible for screening are offered a relatively standard screening regimen, standard in terms of the test and frequency. Population-based screening regimens are not intended for individuals who are at extremely elevated cancer risk due to an unusual exposure or a personal or family history of cancer. When we speak of population-based cancer screening, we exclude the aforementioned individuals, because these individuals usually employ a more intense screening regimen than that employed in population-based cancer screening. These individuals are a very small fraction of the entire population.

The focus of this primer is population-based cancer screening, but principles regarding methodology and assessment still apply when screening individuals at unusually elevated cancer risk. Individuals at that level of risk may weigh benefits and harms of cancer screening differently than those at average risk. Oftentimes more intense cancer screening regimens are offered to individuals at extremely elevated cancer risk. For those individuals, the term surveillance, rather than screening, typically is used.

The phrase population-based often will be excluded as a modifier of the phrase cancer screening in this primer when it is clear that population-based cancer screening is under discussion. The phrase is excluded for reasons of conciseness. Therefore, the reader should assume that population-based cancer screening is being discussed unless otherwise noted.

Readers who are interested in the features of ideal population-based disease screening programs can consult Principles and Practice of Screening for Disease, published in 1968 by Wilson and Jungner (11).

Choosing the cancers for which we screen

Population-based cancer screening occurs routinely in the US for five cancers that, in the absence of screening, typically present as invasive cancer: female breast, cervical, colorectal, lung, and prostate. We screen for these cancers because their invasive forms can lead to morbidity and premature mortality. We also screen because there is evidence, or in some instances suspicion, that cancer screening is beneficial. The fact that cancer screening is recommended by professional organizations or has become established in community settings does not necessarily mean that conclusive evidence of a benefit exists. Adoption of unproven cancer screening tests has occurred in the US and elsewhere.

This primer will not delve into the evidence that supports (or does not support) population-based cancer screening for the five aforementioned cancers. Many well-respected and up-to-date resources for that information already exist (12,13). This purpose of this primer is to teach the reader how to assess and interpret cancer screening through the use of data, not to provide a review of literature on the benefits and harms of screening for specific cancers.

Choosing who to screen

Consideration of who to screen begins with identification of the factors that are known to meaningfully increase cancer risk. Next, prevalence of the risk factors is considered. Sufficient risk and sufficiently prevalent risk factors are necessary to affect an absolute reduction in cancer morbidity and mortality that is large enough to justify population-based cancer screening (assuming, of course, that cancer screening is of benefit). Population-based cancer screening is a resource-intense cancer control method and generally is not used for rare cancers.

Age is the strongest risk factor for adult cancer and as such cancer screening recommendations are based on that factor. For lung cancer screening, smoking history also is a criterion because of its prevalence and strong association with the disease. We do not screen males for breast cancer or never smokers for lung cancer because the chance of individuals in those groups developing the respective cancers is very low. The day may come when age is augmented by genomic or other biologic information to drive cancer screening recommendations, both

for and against screening. We are not yet in that era of personalized cancer screening for individuals at average risk, however.

We choose to screen those for whom we believe the benefit outweighs the harm, though we can only assess that for a population, not for an individual. The term individual refers to the person who is offered screening, while the term population refers to the entire group of individuals who have been offered screening. At the population level, we can examine changes in beneficial outcomes and harmful experiences with the advent of screening. At the individual level, we can never know who will or did benefit from screening, as we do not know what will happen or what would have happened in the absence of screening.

The cancer screening process

Cancer screening cannot result in benefit without the successful completion of other components of the screening process, which encompasses all activities that lead up to and come after application of the screening test. The screening process begins when potential screenees are notified of the option to be screened and ends, at the earliest, when results of the screening test are relayed to the screenee. For those who receive a positive result, the process will extend to diagnostic evaluation and may include cancer diagnosis and treatment.

The resources that are needed to carry out a successful cancer screening effort include more than just those required to administer the cancer screening test. Consideration of resources employed in population-based cancer screening must include, at a minimum, those associated with screening invitation, assessment of eligibility, informed decision making, test interpretation, reporting of results, and diagnostic evaluation and cancer treatment as needed. Other considerations include time and wages lost by individuals who are attending screening, and other manners, perhaps more critical or economically efficient, in which screening resources could be used. Readers who would like to learn more about the screening process can consult Zapka et al (14) and Beaber et al (15).

Cancer screening tests

Cancer screening tests also are known as cancer screening modalities. The screening tests we use in the US are either image-based or biospecimen-based. Imaging tests are used for breast cancer (mammography, digital tomosynthesis), colorectal cancer (sigmoidoscopy, colonoscopy, virtual colonography), and lung cancer (low dose computed tomography (LDCT)). Biospecimen-based tests are used for cervical cancer (pap smear, HPV testing, colorectal cancer (fecal occult blood testing (FOBT)), and prostate cancer (prostate-specific antigen (PSA)).

Some cancer screening tests also are used as diagnostic tests. Colonoscopy is used as a colorectal cancer screening test as well as for evaluation of symptoms or follow-up of a positive FOBT. A positive PSA screening test may lead to serial PSA tests to monitor for changes in PSA. The term indication refers to the reason for performing a test.

Organized screening programs versus opportunistic screening

Cancer screening practices vary from country to country. Reasons include cultural differences, differing interpretations of evidence, and varying public health needs. Central to these choices, however, is the manner in which health care is administered and delivered. Organized screening programs are found in countries with nationalized health care, a setting in which a government body decides on the best medical practices, including cancer screening, and offers and administers, free of charge, only those services deemed appropriate. Infrastructure usually exists to facilitate screening and to manage the experiences of those individuals who receive a positive screening result. Opportunistic screening occurs in the US and in other countries without nationalized health care. Opportunistic screening provides more choice, but individuals typically are left on their

own to navigate the process. Opportunistic screening also occurs in countries with organized screening programs if the primary care physician arranges it or the screenee requests it, but in some jurisdictions the costs of the test must be borne by the individual.

The methods described in this primer can be used to interpret data from organized or opportunistic screening settings. Readers who wish to learn more about organized screening can consult Raffle and Gray's Screening: Evidence and Practice (16).

Benefit versus harm

Assessment of cancer screening tests can be contentious because disagreements exist regarding what constitutes benefit, what constitutes harm, and how to balance the two. We can measure and have measured the impact at a population level by looking for reductions in cause-specific mortality rates. Cause-specific refers to the cause of death that we aim to prevent by cancer screening. Reduction in cause-specific incidence rates is employed for tests that detect precancer and will be discussed in Chapter 8. As the reader will learn, a reduction in cause-specific incidence rate will lead to a reduction in cause-specific mortality rates in most instances.

Cause-specific mortality rates are unable to reflect any harms other than those that affect length of life or cause of death. Yet there are many potential harms of screening, including psychological impact of screening results, diversion of resources away from other health care needs, and late effects (also known as downstream effects) of diagnostic evaluation or cancer treatment. These harms often are difficult to measure, difficult to attribute to the screening process, and vary by screenee. Nevertheless, they are real, and metrics need to be developed that can incorporate them so the net impact of population-based cancer screening programs can be measured.

Benefits and harms can occur at an individual level or a population level. Individual-level harms are more perceptible than population-level harms, but it is at the individual level that the trade-off between benefit and harm is most murky. Acceptable benefit-to-harm ratios differ by individual, because fear of cancer, risk tolerance, risk illiteracy, and other factors vary from person to person.

Reduction in cause-specific mortality rates remains the standard by which most organizations and researchers judge the benefit of population-based cancer screening programs, as it reflects advances in reducing the rates of cancer death, as well as extension of life among those who die of the disease. Lack of a reduction in cause-specific mortality is typically interpreted to mean that cancer screening does not result in benefit.

Breast and colorectal cancer screening have been shown to reduce cause-specific mortality in randomized controlled trials (RCTs), though the tests examined in those trials are now outdated. Newer tests have become the cancer screening standard of care, based on those tests' improvement in performance measures (Chapter 3) relative to the previous and RCT-tested cancer screening standard of care, and without evidence that the newer tests reduce cause-specific mortality rates. The methodological issues involving the adoption of a newer test based on a comparison with the current standard of care test are discussed in Chapter 9.

Efficacy and effectiveness of cancer screening

Efficacy refers to the ability of cancer screening to reduce cause-specific mortality rates in an experimental setting. Effectiveness refers to the ability to affect the same reduction in a community setting, one in which individuals choose whether to be screened as part of their usual health care. Ideally, efficacy is studied first, and the cancer screening test does not disseminate into community settings until it is known to be efficacious.

Efficacy does not guarantee effectiveness. Given their rigor and intense oversight of patient experiences in experimental settings, efficacy studies are considered to provide the best-case scenario regarding cancer screening's ability to reduce cause-specific mortality rates. In community settings, failures in the screening process, such as delayed communication of screening results, inadequate diagnostic evaluation, and lack of

access to appropriate cancer treatment can hinder the realization of a cause-specific incidence or mortality reduction. However, cancer screening can be effective even in the presence of challenges and imperfections.

Cancer screening: turning healthy people into cancer patients

Individuals who present for cancer screening are healthy for all intents and purposes; neither they nor their doctors have any reason to believe they have cancer. A fraction of those screened will be diagnosed and become cancer patients. The diagnosis may lead to prevention of death from cancer. However, it may reflect screen-detection of a cancer that never would have been life-threatening. To say the latter is unfortunate is an understatement. Cancer is a disease that significantly affects every aspect of life.

There is evidence that screening for breast, lung, cervical, colorectal, and perhaps prostate cancer reduces cause-specific mortality relative to the absence of screening, even if there is disagreement regarding the extent of benefit or for whom the benefit exists. In addition to possible benefits, potential screenees need to be informed of the possible harms when the option of cancer screening is raised. Some individuals may opt out of cancer screening; for them, the possible harms outweigh the possible benefits. The choice is reasonable, as it reflects what matters to them.

References

1. U.S. National Cancer Institute. What is cancer? [Internet]. Bethesda (MD): National Cancer Institute; c1990-2000. [updated 9 Feb 2015; cited 1 Oct 2019]; [about 5 screens]. Available from: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.
2. DeVita VT, Lawrence TS, Rosenberg SA. DeVita, Hellman, and Rosenberg's cancer: principles and practice of oncology [Internet]. 9th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2011 [cited 20 October 2019]. 2686 p. Available from: https://www.nihlibrary.nih.gov/agency/nih.
3. Cancer Facts & Figures 2019. [Internet]. Atlanta: American Cancer Society; 2019 [cited 20 October 2019]. Available from: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2019.html.
4. U.S. National Cancer Institute. Surveillance, Epidemiology, and End Results Program [Internet]. Bethesda (MD): National Cancer Institute; [cited 8 Nov 2019]; [about 2 screens]. Available from: seer.cancer.gov.
5. United States House of Representatives. Office of the Law Revision Council, United States Code. National Cancer Act of 1971 (Pub. L. 92-218, Dec. 23, 1971, 85 Stat. 778). Cited 2019 October 29. Available from: http://uscode.house.gov/statutes/pl/92/218.pdf.
6. U.S. National Cancer Institute. Surveillance, Epidemiology, and End Results Program Overview [Internet]. Bethesda (MD): National Cancer Institute; 2018; [cited 8 Nov 2019]; [about 8 screens]. Available from https://seer.cancer.gov/about/factsheets/SEER_Overview.pdf.
7. U.S. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Stat Facts [Internet]. Bethesda (MD): National Cancer Institute; [cited 8 Nov 2019]; [about 3 screens]. Available from https://seer.cancer.gov/statfacts.
8. U.S. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Cancer Statistics [Internet]. Bethesda (MD): National Cancer Institute; [cited 8 Nov 2019]; [about 2 screens]. Available from https://seer.cancer.gov/statistics.
9. U.S. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. SEER Data and Software [Internet]. Bethesda (MD): National Cancer Institute; [cited 8 Nov 2019]; [about 2 screens]. Available from: https://seer.cancer.gov/data-software/.
10. U.S. Centers for Disease Control and Prevention. National Program of Cancer Registries. Atlanta: Centers for Disease Control and Prevention: [cited 8 Nov 2019]; [about 2 screens] Available from: https://www.cdc.gov/cancer/npcr/index.htm

11. Wilson JMG, Jungner G. Principles and practice of screening for disease. [Internet]. Geneva: World Health Organization;1968 [cited 20 October 2019]. Available from: https://apps.who.int/iris/handle/10665/37650.
12. PDQ® Cancer Information Summaries: Screening/Detection. [Internet]. Bethesda: National Cancer Institute. [cited 20 October 2019]. Available from: https://www.cancer.gov/publications/pdq/information-summaries/screening.
13. Cochrane database of systematic reviews. [Internet]. London: Cochrane Library. [cited 20 October 2019]. Available from: https://www.cochranelibrary.com/.
14. Zapka JG, Taplin SH, Solberg LI, Manos MM. A framework for improving the quality of cancer care: the case of breast and cervical cancer screening. Cancer Epidemiol Biomarkers Prev. 2003 Jan;12(1):4–13. PubMed PMID: 12540497.
15. Beaber EF, Kim JJ, Schapira MM, Tosteson ANA, Zauber AG, Geiger AM. Kamineni, Weaver DL, Tiro JA on behalf of the Population-based Research Optimizing Screening through Personalized Regimens consortium. Unifying Screening Processes Within the PROSPR Consortium: A Conceptual Model for Breast, Cervical, and Colorectal Cancer Screening. Unifying Screening Processes Within the PROSPR Consortium: A Conceptual Model for Breast, Cervical, and Colorectal Cancer Screening. J Natl Cancer Inst. 2015 May 7;107(6):djv120. PubMed PMID: 25957378.
16. Raffle A, Gray M. Screening: evidence and practice. 1st ed. New York: Oxford University Press; 2007. 317p.

Chapter 2. Behind the scenes

Cancer screening aims to interfere with disease progression by detecting cancer at a point in its natural history when it is either curable or, if not curable, when treatment will extend life beyond what it would have been in the absence of cancer screening. The phrase screen detected and similar terms are used in this chapter but are a bit of a misnomer, as cancer screening tests are not the final arbiter of the presence of cancer. Screen detected is intended to mean that cancer screening initiated a process that led to the diagnosis of a cancer.

A simple model of the natural history of cancer

The natural history of cancer is complex and for the most part not well understood. Furthermore, there is great variability, even among tumors of the same organ. Figure 1 depicts how cancer progresses. It is an gross over-simplification of the process but it is a useful aid in explaining how cancer screening aims to interfere in the disease process.

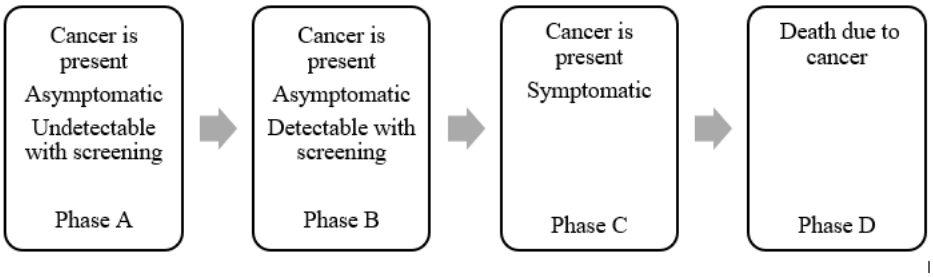


Figure 1: Four phases of cancer progression

Figure 1 displays four phases of cancer progression that are relevant to cancer screening. Cancer is present, asymptomatic, and not yet detectable by screening in Phase A. In Phase B (previously known as the detectable pre-clinical phase or DPCP), cancer is still asymptomatic but has characteristics that should, in the best of all possible worlds, make it detectable through cancer screening. Examples of characteristics are size and shedding of tumor cells that could be detected in a biospecimen. A cancer in Phase B may not be screen detected, however; the individual may not be screened or the test may give an inaccurate result due to its limitations. In Phase C, cancers come to clinical attention due to symptoms. Phase C includes cancers that are curable as well as those that are not. In Phase D, cancer causes death.

It is important to note that Phases A and B are a function of the cancer screening test. A cancer may be classified as being in Phase B if a technologically advanced test is used to screen, but in Phase A otherwise. For example, some lung cancers that can be detected with low dose computed tomography (LDCT) screening would not have been detectable with traditional two-dimensional chest x-ray screening given chest x-ray's poorer resolution and capture of substantially less radiologic information. At a specific point in time, a cancer could be in Phase B if cancer screening with LDCT and Phase A if cancer screening with chest x-ray. Use of chest x-ray screening for lung cancer was common in the later decades of the 20th century but is no longer standard of care.

The purpose of the four phase model is not to demonstrate all possible paths that cancer or an individual with cancer can experience. It assumes that all cancers progress through each phase and do so only in a forward fashion, even though we know that some cancers regress or stall. It assumes that all cancers would be fatal if not treated, though experience tells us otherwise. It assumes that death due to causes other than the cancer of interest cannot occur. Even with those exclusions, the model is useful in conceptualizing our goals in cancer screening and provides a vocabulary that helps us discuss cancer screening.

Cancer screening attempts to shift the diagnosis of Phase C cancers to Phase B. Cancer screening is predicated on the belief that treatment at Phase B is more likely to lead to cure or extension of life than treatment at Phase

C. If treatment at Phase B offers no prognostic advantage over treatment at Phase C, cancer screening will not lead to a reduction in cause-specific mortality. Treatment options may be more palatable, however, for Phase B cancers, and morbidity associated with the cancer and its treatment may be reduced. Then again, a diagnosis that occurs at an earlier time leads to more time spent as a cancer patient, which has psychological and clinical implications as discussed in the Benefits and Harms section of Chapter 1.

Three important phenomena in screen detection of cancer

Lead time, length-biased sampling, and overdiagnosis are three terms that are used frequently in the assessment of cancer screening. They refer to the shift to an earlier date of diagnosis with cancer screening (lead time) and selection of a prognostically favorable (and thus non-random) sample of cancers (length-biased sampling and overdiagnosis). The remainder of this chapter explains these phenomena, while Chapter 5 describes how they complicate interpretation of cancer screening data.

The phrase length-biased sampling becomes awkward when we speak of bias due to length-biased sample. The phrase length time, which is sometimes used instead of length-biased sample, isn't a better choice as it is not particularly descriptive. The phrase length-weighted sampling will be used instead in the remainder of this primer.

In the remainder of this primer, the phrase three cancer screening phenomena refer to lead time, length-weighted sampling, and overdiagnosis.

Lead time

Screen-detected cancers are diagnosed at an earlier point in time than they would have been in the absence of cancer screening. Lead time is the amount of time by which the diagnosis date was advanced by cancer screening. It is that shifting of the diagnosis date to an earlier time that leads many to refer to cancer screening as early or earlier detection.

Lead time for an actual individual is impossible to calculate because we cannot know what the date of symptomatic diagnosis would have been in the absence of cancer screening. But for the purpose of illustration, we will pretend that we know that date. Lead time would be 3 months if, in the absence of cancer screening, a cancer would have been detected on June 1, 2018, but, in the presence of cancer screening, was diagnosed on March 1, 2018.

Length-weighted sampling

The term length-weighted sampling refers to the fact that the chance of screen detection is dependent on the length of time (sometimes referred to as sojourn time) the cancer remains in Phase B. The term sampling is used because cancer screening is merely a selection of some cancers (those that are screen detected) from the pool of all cancers. In elementary probability classes, sampling is often demonstrated using a jar of marbles. If the marbles are all of the same size, each marble has the same chance of selection. If the marbles are of different sizes, the chance that a given marble is selected is determined by its size, with chance of selection positively correlated with size of a marble. Cancer screening is similar to the latter situation; as the reader will see, one particular tumor characteristic, time spent in Phase B, drives the chance of detection.

Recall that not all cancers can be detected through screening. The chance that a cancer will be screen detected is dependent to differing degrees on many factors, including cancer characteristics, screenee characteristics, the cancer screening test, and screening interval. Screening interval refers to the amount of time between screens. An annual screening interval is used (or was used until we learned more about cancer progression) for a number of cancer screening tests.

Most notably, the chance of screen detection is inversely associated with the speed of tumor growth: faster growing cancers spend less time in Phase B and consequently have less time to be screen detected. A cancer that spends only three months in Phase B, for example, will have no opportunity to be detected on an annual screen, unless the annual screen happens to occur during that three-month window. A cancer that spends two years in Phase B will have two annual screens on which it can be detected. Cancers with a longer Phase B are assumed to be slower growing than those with a shorter Phase B, and cancers that are slower growing are assumed to have better prognosis. Therefore, length-weighted sampling leads to detection of cancers through screening that are expected to have better prognosis than those that are not detected though screening.

The term weighted is used to mean that the sample of cancers detected through screening will be skewed in favor of cancers with more favorable prognosis. In other words, the cancers that screening detects do not represent a random sample of all cancers.

Figure 2 depicts the fictional experience of three screenees, Mike, Molly, and Mary. The example demonstrates the interplay of screening interval and length of Phase B.

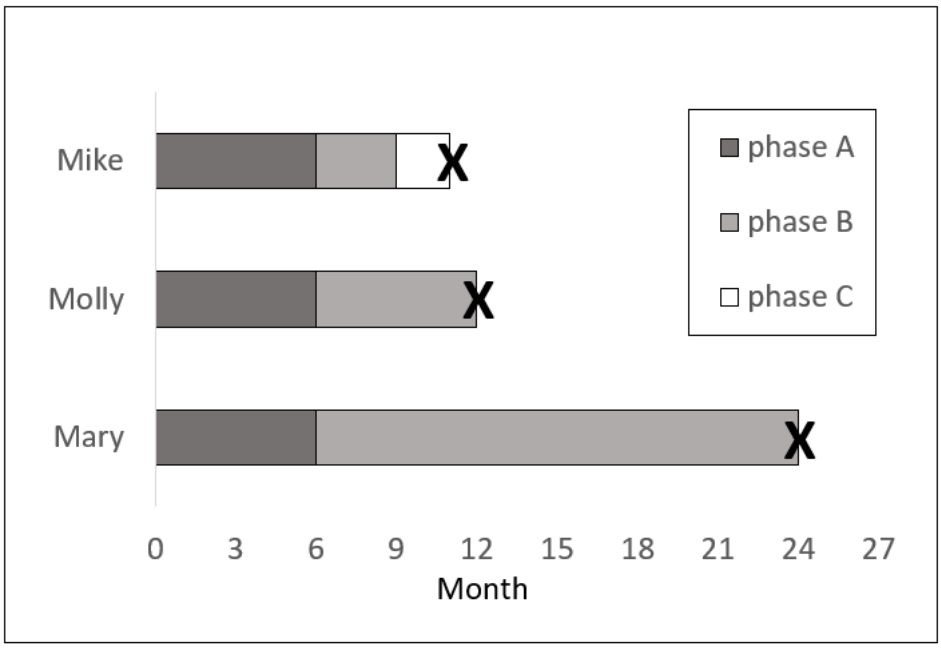


Figure 2: The interplay of Phase B length, a one-year screening interval, and cancer diagnosis
X indicates cancer diagnosis. Experience is fictional.

Mike, Molly, and Mary have Phase A cancers at the time of the first screen (month 0), so none of the three has cancer detected. Each cancer enters Phase B at month 6. Mike's cancer enters Phase C at month 9, leading to a symptom-driven diagnosis prior to the second screen. Molly's cancer is in Phase B at month 12, the time of her second screen, and the cancer is screen detected. Mary's cancer also is in Phase B at month 12, but her cancer is missed at the second screen. Because Mary's cancer is still in Phase B at the time of her third screen (24 months), there is another opportunity to detect it through screening, and that happens.

What would have happened had a different screening interval been used? A six-month interval could have led to screen detection for Mike, because the shorter the screening interval, the more likely that cancers with short Phase B lengths will be detected through cancer screening. The impact of a two-year interval on Molly's experience depends on the length of Phase B: if it had been less than 18 months, Molly's cancer could not have been screen detected.

Overdiagnosis

Overdiagnosis is the detection through cancer screening of cancers that never would have been diagnosed in the absence of cancer screening. These are cancers that, in the absence of screening, would not progress beyond Phase B during the lifetime of the screenee. The existence of overdiagnosis in cancer screening was once quite controversial, but nowadays many researchers and clinicians are open-minded to the possibility that some screen-detected cancers are diagnosed unnecessarily. Many also believe that screening for cancer of any site would result in overdiagnosis. The biggest controversy in overdiagnosis today surrounds its magnitude, which is further discussed in Chapter 9.

Overdiagnosis includes, but is not restricted to, the detection of indolent cancers. Indolent cancers are those that are screen detected (by definition, in Phase B), but in the absence of cancer screening and even in the longest of lifetimes, would have remained in Phase B, regressed to Phase A, or completely resolved. The screen detection of indolent cancers is an extreme example of length-weighted sampling. Because overdiagnosis refers to screen-detected cancers, cancer detected as a result of symptoms (Phase C cancers) cannot be overdiagnosed, even though they too, theoretically, can stall, regress, or resolve.

Non-indolent screen-detected cancers can be overdiagnosed if death, due to a cause other than the screen-detected cancer, occurs between the date of screen detection and the hypothetical date of symptom detection. The phrase competing cause of mortality is used to describe this scenario. In Figure 3, John's cancer is not overdiagnosed because he is alive on the day that the cancer would have been detected due to symptoms. James, however, dies soon after his cancer is diagnosed, and he is not alive on the day that it would have been detected due to symptoms had he lived. If not screen detected, James' cancer would have been in Phase B when he died, and neither he nor his health care providers would have known of its existence.

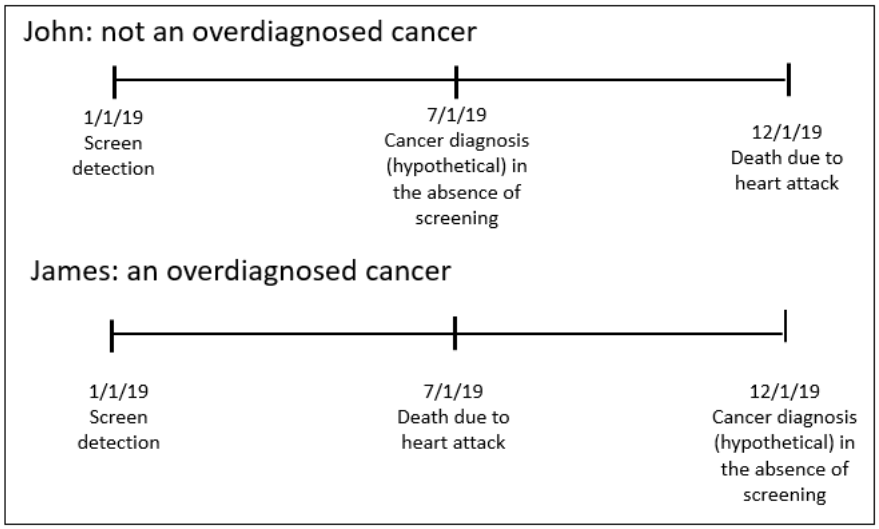


Figure 3: Overdiagnosis due to death from other causes

Experience is fictional

Today's technology generally does not allow for the classification of a tumor as clearly indolent or not. In some instances, tumor characteristics can suggest a likely course, be it an innocuous or highly aggressive one. Death from a different cause soon after screen detection may suggest overdiagnosis associated with a competing cause of mortality, while death long after argues against it. Of course, uncertainty always exists because the life course an individual would have experienced in the absence of cancer screening is unknowable.

Chapter 3. Performance measures

Performance measures reflect the link between cancer screening test results and cancer diagnoses. They provide no information about cause-specific mortality. Performance measures are used in the initial assessment of proposed cancer screening tests and also are used to monitor performance once cancer screening has disseminated. There are six key performance measures, with each interpretable as a probability (ranging from 0 to 1) or percentage (ranging from 0% to 100%).

Performance measures are calculated from the experience of individuals who have been screened. The cancer screening test result and whether cancer was present at the time of the screen need to be available for each individual to calculate performance measures.

The building blocks of performance measures

Cancer screening test result

The cancer screening test result is classified as either positive or negative. A positive result indicates a suspicion of cancer and the need for diagnostic evaluation. A negative result indicates no suspicion of cancer and no need for diagnostic evaluation. The definition of a positive test result is not etched in stone; instead, the medical community makes recommendations as to what constitutes a positive test. In practice, any abnormality deemed suspicious by the test interpreter is called positive, regardless of whether it meets the recommended definition of a positive test. For many cancer screening tests, particularly those than employ imaging, it is impossible for recommendations to include every finding or constellation of findings that creates a suspicion for cancer.

Recommendations are made after many factors are weighed, including the burden of positive tests and the gravity of missing a cancer. Medical communities may arrive at different recommendations. In the US, for example, a prostate-specific antigen (PSA) blood level of 4.0 ng/mL or higher is typically considered a positive test for prostate cancer, but in parts of Europe, a value of 3.0 ng/mL or higher is used.

At the extremes, there tends to be agreement as to whether a cancer screening test result should be classified as positive or negative. For example, a large spiculated lung mass observed on low dose computed tomography (LDCT) would be classified as positive for lung cancer, while a mammogram that shows only the anatomic structures of the breast would be called negative for breast cancer. The challenge comes when it is not obvious what a finding represents: a result that isn't exactly negative and isn't exactly positive. There is a move towards classifying these grey-zone findings as indeterminate and employing a less intense and usually non-invasive form of diagnostic evaluation. Some may disagree with use of the phrase diagnostic evaluation in the instance of indeterminates, as the recommended medical intervention is intended to watch for change in the abnormality rather than determine whether it is cancer. In that instance, the term monitoring can be used. For the purpose of calculating performance measures, I classify indeterminate cancer screening test results as positive. In my opinion, any cancer screening test that is not negative is positive, as it leaves uncertainty in the mind of the clinician and screenee.

Some biospecimen-based cancer screening tests return a numeric value or other quantitative measure. These values correlate with the chance of the presence of cancer. PSA is one such test. A value greater than 4 ng/mL is usually considered a positive result in the United States, but active surveillance rather than biopsy is often recommended if the PSA is between 4 ng/mL and 10 ng/mL. A value of 10 ng/mL or greater, however, typically leads to imaging or biopsy. Other biospecimen-based cancer screening tests, such as cervical cytology, indicate whether abnormal cells are present. One form of cervical cancer screening, human papilloma virus (HPV) testing, indicates whether certain cancer-causing strains of HPV are present rather than indicating whether an abnormality suspicious for cancer is present.

Imaging-based cancer screening tests are used to determine if abnormalities are present. A cancer screening test will be called positive if an abnormality suspicious for cancer is revealed. These tests also can reveal abnormalities that are not suspicious for cancer and abnormalities whose significance with regard to cancer is unknown. Lung cancer screening with LDCT, for example, can lead to detection of non-calcified nodules (positive if above a certain size), calcified nodules (usually negative), or ground glass opacities (oftentimes of uncertain significance). Some imaging-based cancer screening tests also can lead to detection of abnormalities that represent or are suspicious for non-cancer conditions, called incidental findings or incidentalomas. For example, LDCT screening for lung cancer can lead to the detection of coronary artery calcification.

Cancer: present or not?

Cancer is either present or not present at the time of the cancer screening test, though only some cancers that are present can be detected through cancer screening. Recall from Chapter 2 (Figure 1) that Phase A cancers are present but not detectable, while phase B cancers are present and have characteristics that should make them detectable. Knowing whether a cancer is present and detectable at the time of a cancer screening test is often not as simple as the four phase model implies, though. The most challenging aspect is determining whether a negative screen that occurred prior to a symptom-detected cancer represents a true negative or a false negative, terms that are fairly self-explanatory and will be discussed later in this chapter. The following fictional scenarios represent quandaries that researchers face when trying to assess whether a Phase B cancer was present at the time of a negative screen:

- Amanda had a lung cancer screening test and the result was negative. Three months later, she receives a symptom-prompted diagnosis of lung cancer. Was the cancer missed on screening, or was the cancer in Phase A at the time of the screening but moved through Phase B very quickly? Did the cancer exist at the time of the screen?
- Arnie had a prostate cancer screening test and the result was positive. He received standard diagnostic evaluation and his clinician concluded that he did not have prostate cancer. Nine months later, he receives a symptom-prompted diagnosis of prostate cancer. Was the cancer in Phase B at the time of the screen but diagnostic evaluation failed in some way? Is the diagnosed cancer a new and fast growing abnormality. In other words, did the diagnosed cancer arise from an abnormality other than the one that prompted the positive result?
- Astrid schedules her screening mammogram. Two days before the test, she finds a breast lump but does not tell anyone. Her mammogram is positive and diagnostic evaluation indicates that the lump she found is cancer. Astrid's cancer was present at the time of her mammogram, but should the test be considered a screening mammogram or a diagnostic mammogram?

The phrase interval cancer is used to describe cancers that occur between screening rounds and follow either a negative test or a resolved positive test. Resolved means that the conclusion of the diagnostic evaluation was that cancer was not present. Amanda's cancer and Arnie's cancer are interval cancers regardless of whether they were in Phase A or B at the time of the screen. If in Phase B, the previous screening test would be classified as a false negative.

It is clear that Astrid's cancer was present and in Phase C at the time of the screening test. The cancer could be classified as an interval cancer because it was symptomatic before the screen. Then again, it could be classified as screen detected because the screening test result was positive, even though it was beyond Phase B. Cancer screening tests can miss Phase C cancers, and that could have been Astrid's experience.

Most screen-detected cancers are in Phase B at the time of the cancer screening test. For simplicity's sake Phase C cancers that are detected as the result of cancer screening will be excluded for the remainder of this primer.

Calculating cancer screening performance measures

The six cancer screening performance measures are sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR), and false negative rate (FNR). The receiver operating characteristic (ROC) curve is a graph of sensitivity versus FPR, which is equal to 1 minus specificity. The ROC curve demonstrates how those two values vary as the definition of a positive test changes. Its summary measure, area under the curve (AUC), is calculated so that ROC curves can be compared.

The formulas

Table 1 presents the quantities that are needed to calculate performance measures. The four quantities in the center of the table are at the heart of performance measure calculations. They are true positive tests (a), false positive tests (b), false negative tests (c), and true negative tests (d). True positive tests are those positive tests that led to the diagnosis of a cancer, and true negative tests are those negatives tests that correctly indicated no suspicion of cancer. False positive tests are sometimes called false alarms; the test suggests something suspicious, but diagnostic evaluation reveals that cancer is not present. False negative tests are incorrectly negative: cancer is present and in Phase B, but the cancer screening test result is negative. Because Phase A cancers cannot be detected by cancer screening, they are considered not to be present when calculating performance measures.

Table 1: The components of performance measure formulas

		Truth		
		Cancer present (in Phase B)	Cancer not present (includes Phase A)	Total
Screening test result	Positive	a *true positives*	b *false positives*	a+b *all positives*
	Negative	c *false negatives*	d *true negatives*	c+d *all negatives*
	Total	a+c *cancers present*	b+d *cancers not present*	a+b+c+d *all screenees*

Cancers in Phase C can be screen detected, but most screen-detected cancers are in Phase B. For simplicity's sake Phase C cancers are not included as cancers that are screen detected.

The six performance measures are defined as follows. The formulas use the notation in Table 1.

- Sensitivity, sometimes abbreviated as S_e, is the percentage of people with cancer who had a positive test; a/(a+c).
- Specificity, sometimes abbreviated as S_p, is the percentage of people without cancer who had a negative test; d/(b+d).
- PPV is the percentage of people with a positive test who had cancer; a/(a+b).
- NPV is the percentage of people with a negative cancer screening test who did not have cancer; d/(c+d).
- FPR is the percentage of people without cancer who had a positive test; b/(b+d). FPR equals 1 minus specificity.
- FNR is not typically reported but will be defined here for completeness' sake. It is the percentage of people with cancer who had a negative cancer screening test; c/(a+c). FNR equals 1 minus sensitivity.

Positivity and negativity rate usually are not referred to as performance measures, but it is important to present them nonetheless:

- Positivity rate is the percentage of people screened who have a positive test; (a+b)/(a+b+c+d)
- Negativity rate is the percentage of people screened who had a negative test; (c+d)/(a+b+c+d).

Table 2 presents data from the Breast Cancer Surveillance Consortium (BCSC) Data Explorer, a public-access database of mammographic breast cancer screening experience from 1994 through 2009 (1). These data are used in Table 3 to calculate the performance measures.

Table 2: Screening mammogram classification among women ages 50 to 59 at the time of screening. Breast Cancer Surveillance Consortium Data Explorer, 1994-2009

		Truth		
		Cancer present	Cancer not present	Total
Screening test result	Positive	7,044 *(true positives)*	165,115 *(false positives)*	172,159
	Negative	1,534 *(false negatives)*	1,623,399 *(true negatives)*	1,624,933
	Total	8,578	1,788,514	1,797,092 *(all screens)*

Table 3: Performance measures, positivity rate, and negativity rate: formulas and calculations using data from Table 2

Performance measure	Formulas using Table 1 notation	Calculations using Table 2 data
Sensitivity	a/(a+c) true positives/cancers present	7,044/8,578 82%
Specificity	d/(b+d) true negatives/cancer not present	1,623,399/1,788,514 91%
PPV	a/(a+b) true positives/all positives	7044/172,159 4%
NPV	d/(c+d) true negatives/all negatives	1,623,399/1,624,933 >99%
FPR	b/(b+d) false positives/cancer not present also equal to 1-specificity	165,115/1,788,514 9%
FNR	c/(a+c) false negatives/cancers present also equal to 1-sensitivity	1,534/8,578 18%
Positivity rate	(a+b)/(a+b+c+d) all positives/all screened	172,159/1,797,092 10%
Negativity rate	(c+d)/(a+b+c+d) all negatives/all screened	1,624,933/1,797,092 90%

In the BCSC example, sensitivity and specificity are fairly high, as is often the case with cancer screening tests that are used in population-based cancer screening. The manner in which a positive test is defined generally drives sensitivity and specificity, as do the capabilities and limitations of the cancer screening test itself. The definition of a positive screen is chosen so that most cancers are found (high sensitivity) and the absolute number of false positives is kept as low as possible (low FPR, or high specificity). NPV is high as well but PPV is very low.

The relationship between PPV, NPV, and prevalence

PPV and NPV are driven by sensitivity and specificity, and they also are driven by the prevalence of disease. PPV and NPV can be calculated from sensitivity, sensitivity, and the prevalence of disease using the formulas in Box 1.

Box 1: Calculating PPV and NPV from sensitivity (S_e), specificity (S_p), and prevalence.

$PPV = (S_e \times prevalence)/(S_e \times prevalence + (1 - S_p) \times (1\text{-}prevalence))$

$NPV = (S_p \times (1 - prevalence)/(S_p \times (1 - prevalence) + (1 - S_e) \times prevalence)$

The Box 1 PPV formula indicates that PPV always will be low in the instance of a rare disease (low prevalence) because the numerator will be substantially smaller than the denominator. The Box 1 NPV formula indicates that NPV always will be high in the instance of a rare disease because the numerator and denominator will be nearly the same. Those statements are true because the quantity (S_e x prevalence) will be close to zero when prevalence is low. Table 4 presents, using the BCSC sensitivity (82%) and specificity (91%), values of PPV and NPV for a range of prevalence values. The annual prevalence in the BCSC cohort is approximately 500 per 100,000 women. In Table 4, notice that PPV increases as prevalence increases, but it takes an implausible prevalence, 100 times that of the prevalence observed in the BCSC cohort (50,000 per 100,000 women), for PPV to rise to 90%. A prevalence of 50,000 per 100,000 women means that every other woman has breast cancer, something that is far from true for any cancer.

Table 4: PPV and NPV by prevalence of disease (sensitivity of 82% and specificity of 91%)

Prevalence	PPV	NPV
250 per 100,000	2.2%	>99%
500 per 100,000	4.3%	>99%
1,000 per 100,000	8.2%	>99%
50,000 per 100,00	90.8%	>99%

Data are fictional

Those who are new to assessment of cancer screening often are amazed that PPV is so low for cancer screening tests even when sensitivity and specificity are high. Table 5 demonstrates, for a typical cancer prevalence of 500 per 100,000, how changes in sensitivity and specificity affect PPV. Notice that even at a sensitivity and specificity of 99%, values that are yet to be achieved for cancer screening modalities, PPV is only 33%. The data in Table 5 demonstrate that it is virtually impossible for PPV to rise above 10% given typical prevalence, sensitivity, and specificity associated with today's cancer screening tests.

Table 5: PPV as a function of sensitivity and specificity (disease prevalence of 500 per 100,000)

	Sensitivity		
	90%	95%	99%
Specificity			
90%	4.3%	4.6%	4.7%
95%	8.3%	8.7%	9.0%

Table 5 continued from previous page.

	Sensitivity		
	90%	95%	99%
Specificity			
99%	31.1%	32.3%	33.2%

Data are fictional

The implications of low PPV

A low PPV indicates that most positive cancer screening tests are false alarms. A PPV of 4% means that 96% of positive tests do not lead to a cancer diagnosis. In the BCSC data (Table 2), there are 7,000 true positives but 165,000 false positives. There is disagreement as to whether false positives should be classified as a harm of cancer screening. One point of view is that any test, including the diagnostic evaluation tests that accompany a false positive, is a test worth having if it rules out cancer. The other point of view is that false positives are a harm of cancer screening as they cause patients to worry unnecessarily and to receive unneeded medical tests and procedures, some of which can be risky.

Can PPV be improved?

As was demonstrated in Box 1, PPV depends on three quantities: prevalence, sensitivity, and specificity. Disease prevalence is, for all intents and purposes, not modifiable (and definitely not in the short term), and while we do have some control over sensitivity and specificity, their upper bounds are determined, realistically, by the abilities of the cancer screening tests. So PPV will remain low. And cancer screening will continue to generate many more false than true positive tests.

Recall that the intent of cancer screening is not to diagnose; rather it is to identify individuals who need additional medical attention to determine if they have cancer or to rule that out. A cancer screening test with a value of 100% for sensitivity, specificity, PPV, and NPV would be possible if a cancer screening test has perfect discriminatory ability, which is contrary to the goal of cancer screening. We could guarantee 100% sensitivity by assigning a positive test result to every screening test, but in that instance, PPV will still be low: it will equal the prevalence of the cancer. We could guarantee 100% specificity by assigning a negative test result to every cancer screening test, but in that instance, no cancers would be screen detected.

ROC curves and AUC

An ROC curve demonstrates the trade-off between detecting more cancers and increasing the FPR. The curve is formed by graphing the sensitivity and FPR for different definitions of positivity. Usually an established screening cohort with information on cancer diagnoses and specifics of what was observed on the cancer screening test (rather than only a positive/negative test result classification) is used and scenarios are created. Prostate cancer screening provides a straightforward example. A PSA of 4 ng/mL or greater is the usual definition of a positive prostate cancer screening test in the US, but what would have happened if the cut-off was 3 ng/mL or 5 ng/mL, say? How many additional cancers would be detected with the lower cut-off, and how many additional cancers would be missed with the higher cut-off? The FPR would increase with the lower cut-off and decrease with the higher cut-off, but by how much?

ROC curves provide useful comparisons, though it is necessary to make assumptions when using the scenarios. We must assume that the experience that follows the cancer screening test is the same regardless of the positivity definition employed. For example, we must assume that any cancer diagnosed through cancer screening ultimately would be symptom-detected (no overdiagnosis), and we also must assume that the intensity of

diagnostic evaluation is the same regardless of the positivity definition employed. An ROC curve is built by selecting a finite number of positivity definitions, graphing the sensitivities and FPRs that would have resulted from those positivity definitions, and connecting the dots either in a linear fashion or by way of smoothing.

Examples of cancer screening test ROC curves can be found in the biomedical literature (2-4). For illustrative purposes, the BCSC data presented in Table 2 were used to lay the foundation for a fictional ROC curve.

ROC curves

Figure 4 presents our fictional ROC curve. Sensitivity is plotted along the y-axis and the FPR is plotted along the x-axis. The ROC curve rises steeply as sensitivity moves away from zero, indicating a large gain in sensitivity with only small increases in FPR. All ROC curves have a turning point, a point at which the incremental ability to improve sensitivity becomes increasingly more expensive in terms of FPR.

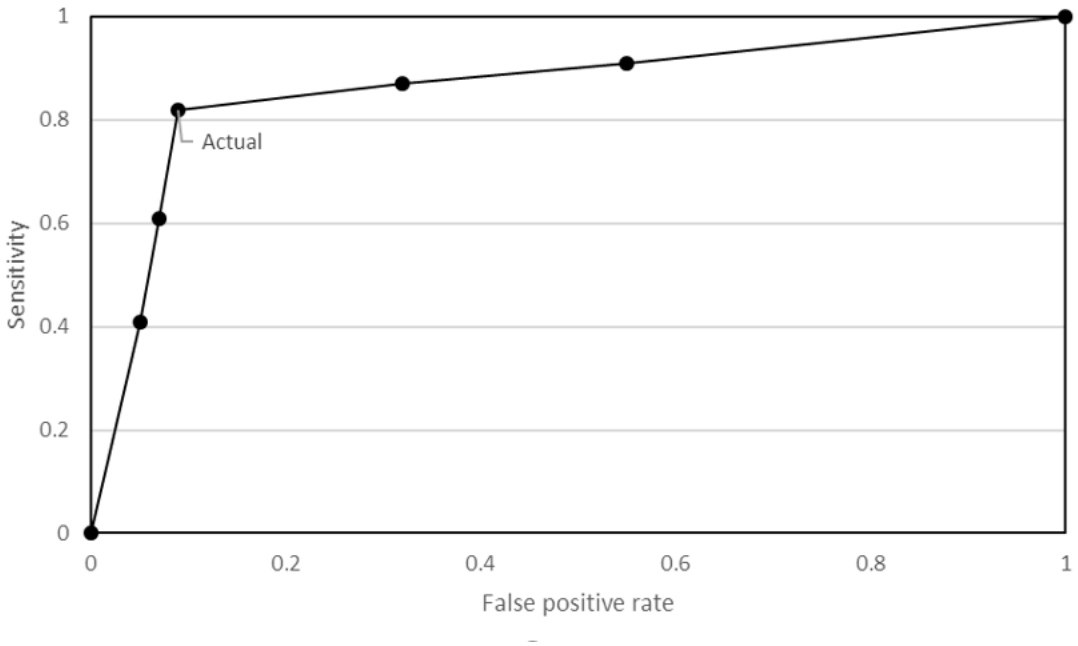


Figure 4: Fictional ROC curve

All ROC curves include the points [0,0] and [1,1]; it is the path the curve takes from [0,0] to [1,1] that varies. [0,0] represents the unrealistic situation in which all test are negative, which results in a sensitivity of zero and an FPR of 0. [1,1] represents the unrealistic situation in which all tests are positive, which results in a sensitivity of 1 and an FPR of 1.

ROC curves can be created for cancer screening tests that return continuous measures, such as PSA, by selecting and varying the value that defines positivity. They also can be used for tests that return categorical classifications, such as the BI-RADS classification for breast abnormalities (5), by collapsing the categories into only two: positive and negative. Let's say that a cancer screening test returns a value of 1, 2, 3, 4, or 5. To create the ROC curve, a positive test result could be defined as a value of 2 or greater, a value of 3 or greater, or a value of 4 or greater. Sensitivity and FPR would then be calculated for each of the three scenarios to create the ROC curve.

The ROC curve (Figure 4) was created using a small number of data points for ease of calculation and presentation. The points were developed by modifying the BCSC values of sensitivity (82%) and FPR (9%) (Table 3): the number of false positives and false negatives were varied by percentages rather than examining test findings and reclassifying according to new positivity definitions. The actual point and the derived points are presented in Table 6.

Table 6: Values of sensitivity and FPR used to calculate the ROC curve in Figure 4

Sensitivity	FPR	Veracity
0.41	0.05	Fictional
0.62	0.07	Fictional
0.82	0.09	Actual
0.87	0.32	Fictional
0.91	0.55	Fictional

Calculating AUC

ROC curves can be summarized and compared by calculating the area underneath them. That area, the AUC, is circumscribed by the curve itself, the x-axis, and a right sided y-axis, and can be calculated using simple formulas for area or, if desired, integral calculus. The AUC for the ROC curve in Figure 4 is 0.87 and was calculated by dividing the area into 5 rectangles and 6 triangles and summing those areas (Figure 5). Many ROC curves presented in the literature are smoothed, however. Smoothing involves advanced mathematics, which is beyond the scope of this primer. Smoothing an ROC curve should change the AUC only slightly.

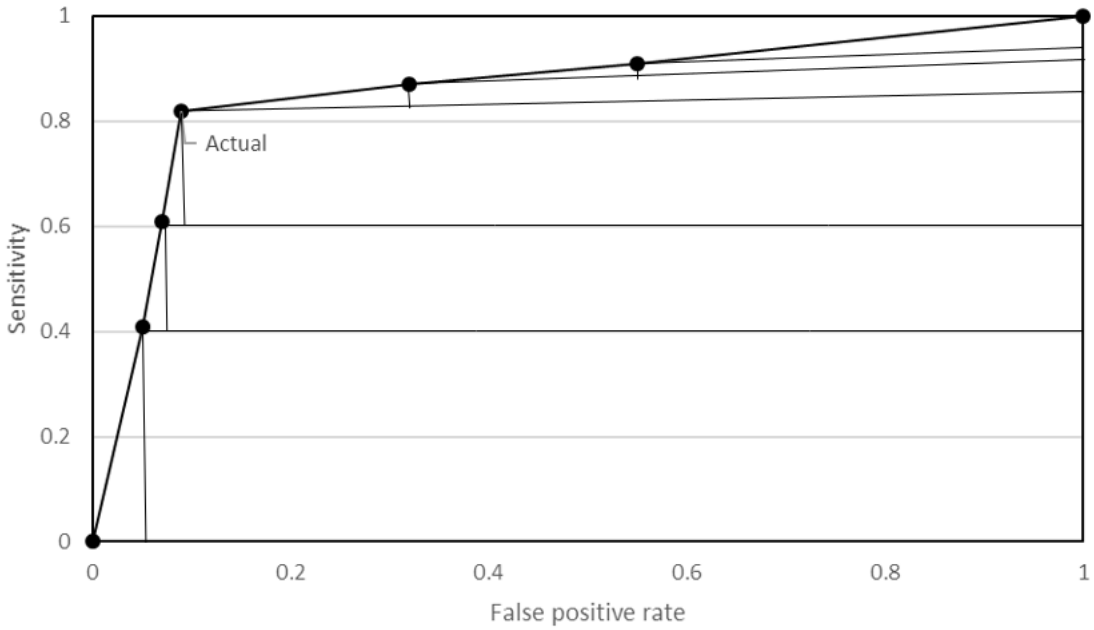


Figure 5: Calculating AUC by partitioning the space under the fictional ROC curve into rectangles and triangles

AUC ranges from 0.5 to 1.0. An AUC of 0.5 represents a cancer screening test with no discriminatory ability, meaning that the result does not depend on whether cancer is present. The cancer screening test is, in effect, no better than flipping a (fair) coin to assign the result. An AUC of 1.0 indicates perfect discriminatory ability: the point [1,0] defines the curve. In that instance, sensitivity is 1 and the FPR is 0. The points [0,0] and [1,1] are not viable scenarios in cancer screening, but they create standard anchors for the curve so that AUCs can be calculated and compared.

Performance measures: Evidence or not?

Performance measures are useful for describing the discriminatory ability of cancer screening tests and for comparing one cancer screening test to another. But they measure the ability of cancer screening to lead to

detection of cancer, not the ability of cancer screening to reduce cause-specific mortality. Chapter 5 explains that improvement in cancer detection does not guarantee a reduction in cause-specific mortality.

Performance measures are rarely considered sufficient evidence to implement cancer screening for the first time. However, a new cancer screening test, one that is similar to an established test known to reduce cause-specific mortality, often disseminates into practice if its performance measures are superior to those of the older test. Examples include the change from film mammography to digital mammography (breast cancer) (3) and the change from guaiac FOBT to immunochemical FOBT, also known as FIT (colorectal cancer) (6). Adoption of new cancer screening tests based on comparison of performance measures with that of past tests is discussed in more detail in Chapter 8.

References

1. BCSC Data Explorer [Internet]. Seattle: Breast Cancer Surveillance Consortium. 2011- [cited 2019 Oct 20]. Available from: http://tools.bcsc-scc.org/dataexplorer/.
2. Thompson IM, Ankerst DP, Chi C, Lucia MS, Goodman PJ, Crowley JJ, Parnes HL, Coltman CA. Operating Characteristics of Prostate-Specific Antigen in Men With an Initial PSA Level of 3.0 ng/mL or Lower. JAMA. 2005 Jul 6;294(1):66–70. PubMed PMID: 15998892.
3. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LF, Bassett L, D'Orsi D, Jong R, Rebner M; Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. New Engl J Med. 2005 Oct 27;353(17):1773–83. PubMed PMID: 16169887.
4. Pinsky PF, Gierada DS, Nath H, Kazerooni EA, Amarosa J. ROC curves for low-dose CT in the National Lung Screening Trial. J Med Screen. 2013 Aug 30;20(3):165–8. PubMed PMID: 24009092.
5. ACR BI-RADS Atlas 5th Edition [Internet]. Reston (VA): American College of Radiology. 2013- [cited 2019 Oct 20]. Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads.
6. Allison JE, Sakoda LC, Levin TR, Tucker JP, Tekawa IS, Cuff T, Pauly MP, Shlager L, Palitz AM, Zhao WK, Schwartz JS, Ransohoff DF, Selby JV. Screening for colorectal neoplasms with new fecal occult blood tests: update on performance characteristics. J Natl Cancer Inst. 2007 Oct 3;99(19):1462–70. PubMed PMID: 17895475.

Chapter 4. Population measures: definitions

Performance measures, the subject of Chapter 3, describe the accuracy of a cancer screening test and its ability to lead to cancer detection in a set of screened individuals. They do not, however, describe characteristics of the detected cancers or experience after diagnosis. Intermediate and definitive outcomes do reflect that information. Intermediate and definitive outcomes are measured at the population level and therefore are called population measures. The term population refers to a group of individuals who are either formally offered cancer screening or for whom cancer screening is available. Though population can mean residents of a geographic region, it does not need to. Population also can refer to the types of research populations discussed in Chapter 6 and Chapter 7. When assessing the impact of cancer screening on the population-level cancer burden, intermediate and definitive outcomes must incorporate the experience of all cancers, regardless of the method of detection, and include all individuals who were eligible to be screened. For example, assessment of intermediate and definitive breast cancer screening outcomes would need to be calculated using both screen-detected cancers and cancers diagnosed due to symptomatic presentation.

Intermediate outcomes can be measured earlier in time than definitive outcomes. Definitive outcomes are not affected by the three screening phenomena (lead time, length-weighted sampling, and overdiagnosis) that were described in Chapter 2. Favorable intermediate outcomes are necessary but not sufficient for favorable definitive outcomes. However, intermediate outcomes that clearly are not favorable are sufficient evidence that cancer screening will not reduce cause-specific mortality.

Intermediate and definitive outcomes are defined in this chapter. Examples and associated calculations are presented. Chapter 5 will address reasons why intermediate outcomes are necessary but not sufficient to guarantee a reduction in cause-specific mortality as well as why definitive outcomes are not affected by the screening phenomena. Phenomena that can lead to inaccuracies in assignment of cause of death also will be discussed in Chapter 5.

The first intermediate outcome to be discussed, cancer incidence, is an intermediate outcome for early detection cancer screening, but a definitive outcome for cancer prevention screening. The reasons for the different classification are discussed in Chapter 8. The discussion of cancer incidence in this chapter is pertinent only to early detection cancer screening.

Intermediate outcomes

Cancer incidence

Incidence reflects the number of cancers that are diagnosed. Absolute numbers of cancers can be reported, although an incidence rate is more commonly used to allow for comparisons across populations of different sizes or with different lengths of observation. A rate incorporates a unit of time or is stated per an amount of time. Incidence rates are calculated as the number of cancers diagnosed (numerator) divided by the number of persons or person-years at risk for the cancer (denominator). Person-years are merely a measure of cumulative time, and are most often used in characterizing data from prospective research in which participants are monitored for different amounts of time. Incidence rates that do not use person-years of experience must state the unit of time over which the experience occurred. Most cancer incidence rates are age-adjusted because incidence of cancer varies by age.

There are many examples of cancer incidence rates in the literature. A widely-used and well-respected source is SEER, which was discussed in Chapter 1. SEER has been collecting data on cancer incidence, cancer mortality and other cancer outcomes in parts of the US since the early 1970s (1).

Here are two examples:

- The SEER age-adjusted incidence rate of breast cancer in 2016 was 129.81 per 100,000 women per year. SEER reports rates in that manner (rather than person-years) as its focus is on annual measures. The equivalent person-years rate would be 129.81 per 100,000 person-years (2).
- The lung cancer incidence rate in the low-dose computed tomography (LDCT) arm of the National Lung Screening Trial (NLST) was 645 per 100,000 person-years. That rate was based on 1060 lung cancers and 164,341 person-years of experience (3).

Calculating a cancer incidence rate: a fictional example

Table 7 presents a cancer incidence rate calculation that uses person-years. The data are fictional and not reflective of the typical magnitude of cancer incidence. To calculate the incidence rate, experience (person-years) is truncated at the date of cancer diagnosis because cancer incidence is the outcome of interest. Put another way, cancer diagnosis is our endpoint for cancer incidence. Information on vital status, date of death, and cause of death are not needed yet.

The experience of 5 individuals is presented in Table 7. Each of the five is followed from the date of his or her 50^{th} birthday. Two are diagnosed with cancer. The numerator is the number of cancers that were diagnosed (2 in this example). The denominator is the sum of the person-years that each individual contributed. Deborah and David were diagnosed with cancer, so person-years equals the time from the date of the 50^{th} birthday to the date of diagnosis. Don and Dudley were not diagnosed with cancer and died prior to their 55^{th} birthday, so person-years equals the time from the 50^{th} birthday to the date of death. Their person-years are truncated at the date of death because they are only at risk of cancer while they are alive. Douglas was not diagnosed with cancer and was alive at his 55^{th} birthday. He contributes five person-years, the maximum time possible in this example. Douglas was at risk of cancer for the entire period of observation.

Table 7: Calculating cancer incidence

	Date turned 50	Status on 55^{th} birthday	Date of diagnosis	Date of death or 55^{th} birthday	Person-years contributed
Douglas	1/1/18	Alive, never diagnosed with cancer	N/A	1/1/23	5 years
Deborah	4/1/18	Dead, diagnosed with cancer	10/1/19	6/1/21	1.5 years
Don	7/1/18	Dead, never diagnosed with cancer	N/A	1/1/20	1.5 years
David	10/1/18	Alive, diagnosed with cancer	4/1/22	10/1/23	3.5 years
Dudley	1/1/19	Dead, never diagnosed with cancer	N/A	1/1/21	2 years
Number of cancers during the 5 year period: 2 Number of person-years: 5+1.5+1.5+3.5+2=13.5 Five-year cancer incidence rate among these individuals: 2/13.5, or 14.8 per 100 person-years					

Experience is fictional

Stage distribution

A stage distribution is fashioned from a series of diagnosed cancers. It presents the number and percent of cancers that have and have not spread. For cancers that have spread, stage distribution captures the extent of spread. The predominant staging system is the TNM system (4). The T value indicates the size of the primary tumor and the spread into nearby tissue; the N value describes spread of cancer to nearby lymph nodes; and the M value describes metastasis (spread of cancer to other parts of the body). A T1N0M0 breast tumor is one that is

invasive, smaller than 20 millimeters in its greatest dimension, and has not spread to lymph nodes or to other organs.

The TNM staging system is quite extensive, and in population-level research, TNM codes usually are combined to create the summary staging categories: local, regional, distant, and if need be unknown. Local refers to cancer that is invasive and confined to the primary site, regional refers to cancer that has spread to regional lymph nodes, and distant refers to cancer that has metastasized. An example of a stage distribution from SEER is found in Table 8 (5). The TNM and summary staging systems include categories for in situ disease (the most advanced form of precancer), though that category is not included in the example.

Table 8: Stage distribution of breast cancers diagnosed 2008 to 2014 and reported to the SEER 18 registry grouping

Stage	Number*	Percentage
Local	213,258	62
Regional	106,629	31
Distant	20,638	6
Unknown	6,879	2
Total	343,965	100

* SEER reports percentages, but not numbers, by stage. The stage-specific numbers in this table were calculated by multiplying the total number of breast cancers by the stage-specific percentages.

The terms early stage, advanced stage, and late stage are frequently used when discussing cancer screening. Early stage generally refers to cancers that are curable, local-stage cancers, or cancers that are in a relatively early phase of their existence. Cancer screening aims to detect those cancers. Advanced and late stage generally refer to distant-stage cancers, cancers that are not curable, or cancers that are expected to be fatal. Cancer screening does not aim to detect cancers at late stages as their prognosis is unlikely to be improved.

Case survival

Case survival is the time from cancer diagnosis to death from any cause, and in assessment of cancer screening is typically measured in months or years. Individual case survival is calculated by subtracting the date of diagnosis from the date of death. Summary measures of case survival are calculated for a series of cancers.

Case survival can be reported using medians or means, but is frequently reported as a percentage of cases alive after a certain amount of time, usually 5 years. Relative case survival is typically used; it is a measure that takes into account the hypothetical mortality the cancers patients would have had had they not been diagnosed with cancer. Relative case survival is calculated by dividing the number of cancer patients alive at the end of a given period of time by the number of individuals in a comparable but cancer-free population alive after the same period of time. It is the latter group that represents the aforementioned hypothetical mortality. A non-relative case survival percentage is calculated by dividing the number of living cancer patients by the total number of cancer patients, and is smaller than a relative case survival percentage because it does not consider that some cancer patients would have died of a cause other than cancer had they not been diagnosed with cancer. Table 9 presents relative and non-relative case survival percentages for a fictional sample of 100 cancer patients.

Table 9: Calculating relative and non-relative case survival

Population	Number	Number alive 5 years later	Number dead 5 years later because of the cancer	Number dead 5 years later because of other causes
Cancer patients	100	85	6	9

Table 9 continued from previous page.

Population	Number	Number alive 5 years later	Number dead 5 years later because of the cancer	Number dead 5 years later because of other causes
Individuals who are otherwise similar to the cancer patients	100	90	0	10
Relative 5-year case survival: 85/90 = 94%. Non-relative 5-year case survival: 85/100=85%.				

Experience is fictional

Definitive outcomes

Cause-specific and all-cause mortality

Mortality reflects the number of individuals who die. As is the case with incidence, rates rather than absolute numbers usually are reported. Both cause-specific and all-cause mortality rates use the number of person-years of individuals at risk of any death as the denominator. Cause-specific mortality rates use the number of individuals who died of the cause of interest as the numerator. All-cause mortality rates use the number of individuals who died of any cause.

It is common to confuse the meanings of mortality measures and case survival because they both involve death. The primary difference between the two is that case survival includes only those individuals who have been diagnosed with cancer. Mortality measures include all individuals who are at risk of dying from any cause. Put another way, mortality measures include those with cancer as well as those without cancer.

Figure 6 may help to explain. N, the number at risk of death, includes all individuals. C is the subset of the N individuals who were diagnosed with cancer. D is the subset of the C individuals who died of cancer. C minus D is the number of individuals with cancer who did not die of cancer. A mortality measure would use N (or their person-years of experience) as the denominator, while a case survival measure would use C as the denominator. As mentioned above, a cause-specific mortality measure would use D as the numerator. Figure 7 is a modified version of Figure 6 and depicts the cascade for all-cause mortality measures.

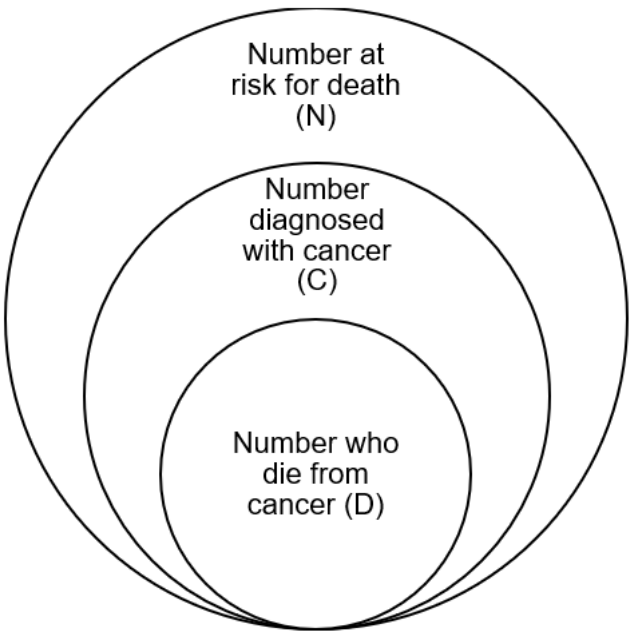


Figure 6: The cascade from at risk for death to cause-specific death

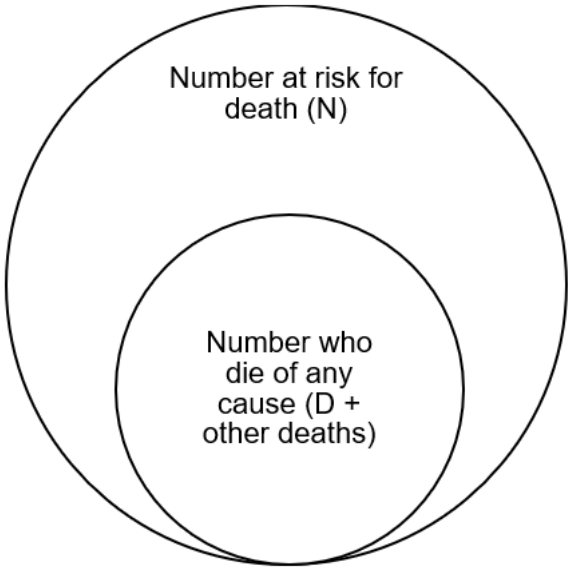


Figure 7: The cascade from at risk for death to death from any cause

Mortality rates reflect two measurable aspects of life: vital status and length of life. The use of person-years as the denominator enables cause-specific rates to reflect extension of life even if the cause of death is the cancer of interest. All-cause mortality rates will reflect the extension as well.

Here are two examples of cancer mortality rates:

- The SEER age-adjusted mortality rate of breast cancer in 2016 was 20.03 per 100,000 women per year or 20.03 per 100,000 person-years (2).
- The lung cancer mortality rate in the LDCT arm of the NLST was 247 per 100,000 person-years. That rate corresponds to 356 lung cancer deaths and 144,103 person-years of experience (3).

Calculating mortality rates: a fictional example

Table 10 displays the fictional experience of 100 individuals. All are alive on 1/1/18, twenty die on 6/30/2018, and the remaining 80 are alive on 12/31/18. Two pieces of information are needed to calculate the all-cause mortality rate: the number of individuals who died (numerator) and the collective amount of time that individuals were alive (denominator). The numerator clearly is 20, but the denominator requires some calculation. We need to separately consider those who were alive on 12/31/18 and those who died before that date. The 80 individuals who were alive on 12/31/18 each contribute a full year of time, for a total of 80 person-years. The 20 who died did so half-way through the year. Each contributes half a year for a total of 10 person-years. The denominator equals the sum of the contributions from the two groups: 80 person-years plus 10 person-years, or 90 person-years. The all-cause mortality rate is 20 per 90 person-years, or 22.2 per 100 person-years.

In some instances, researchers may choose to calculate a simple percentage, reflecting the percentage of individuals who died. In the Table 10 example, the percentage of individuals who were alive on 1/1/18 but who died before 12/31/18 is 20%. It is smaller than the mortality rate and thus suggests more favorable experience. But it does so incorrectly, as it does not take into account that the 20 individuals who died did so halfway through the year. The simple percentage treats the deaths as if they occurred on the last day of the year. While the Table 10 example generated only a small disagreement in the percentage and mortality rate, the two metrics, when calculated using larger numbers of individuals who have a range of life lengths, can be meaningfully different from one another.

Table 10: Calculating an all-cause mortality rate for 2018 using fictional data

	Number	Person-years contribution for 2018
Those alive on 1/1/18	100	Not applicable
Those who die on 6/30/18	20	10
Those alive on 12/31/18	80	80
Mortality rate: 20/(10+80) = 20 per 90 person-years, or 22.2 per 100 person-years		

Table 11 expands on Table 10 by including information on cause of death. Five of the individuals who died on 6/30/18 died of lung cancer, while the remaining 15 died of another cause. When calculating a cause-specific mortality rate, only individuals who died of the cause of interest are included in the numerator, although all who died contribute the amount of time they were alive to the denominator. Table 11 shows calculations for the lung cancer mortality rate. The numerator is now 5, yet the denominator remains at 90. The numerator is smaller because not all deaths were due to lung cancer. The denominator is the same as that of the all-cause mortality rate because the same amount of time was lived. The lung cancer mortality rate is 5 per 90 person-years, or 5.6 per 100 person-years. It is smaller than the all-cause mortality rate because our goal was to measure the rate of dying of lung cancer, which is less common than dying of any cause.

Table 11: Calculating a lung cancer mortality rate for 2018

	Number	Person-years contribution for 2018
Those alive on 1/1/18	100	
Those who die on 6/30/18 of lung cancer	5	2.5
Those who die on 6/30/18 of a cause other than lung cancer	15	7.5
Those alive on 12/31/18	80	80
Lung cancer mortality rate: 5/(2.5+7.5+80) = 5 per 90 person-years, or 5.6 per 100 person-years		

Data are fictional

References

1. Surveillance, Epidemiology, and End Points Program [Internet]. Bethesda (MD): National Cancer Institute. c2000 - 2019 [cited 2019 Oct 22]. Available from: https://seer.cancer.gov.
2. Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2016, Section 4: Breast cancer [Internet]. Bethesda (MD): National Cancer Institute; 2019 Updated 2019 Apr; [cited 2019 Oct 22]. Available from: https://seer.cancer.gov/csr/1975_2016/results_merged/sect_04_breast.pdf.
3. The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011 Aug 4;365(5):395–409. PubMed PMID: 21714641.
4. American Joint Committee on Cancer. Cancer Staging System. Cited 2019 October 29. Available from: http://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx.
5. Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2014, National Cancer Institute. Bethesda, MD, based on November 2016 SEER data submission, posted to the SEER web site, April 2017. Available from: https://seer.cancer.gov/archive/csr/1975_2014/results_merged/sect_04_breast.pdf.

Chapter 5. Population measures: cancer screening's impact

If assessment of cancer screening involved nothing more than calculating the outcomes described in Chapter 4, there would be little need for this primer. The challenging aspect is the interpretation of changes in outcomes, both intermediate and definitive, that accompany cancer screening. The material in Chapter 5 is presented in terms of a change from no population-based cancer screening to the establishment of population-based cancer screening, even though the same principles apply when an established cancer screening test is replaced by one with improved performance measures. Matters specific to the latter scenario are discussed further in Chapter 9.

The three screening phenomena presented in Chapter 2, lead time, length-weighted sampling, and overdiagnosis, feature prominently in Chapter 5. The reader may wish to review that material prior to proceeding.

The material on cancer incidence presented in this chapter is pertinent only to early detection cancer screening. The impact on the incidence of cancer prevention screening is presented in Chapter 8.

Cancer screening's impact on intermediate outcomes

Cancer incidence

Cancer incidence is expected to increase when cancer screening is introduced. Lead time results in diagnosis at an earlier point in time, creating a bunching effect as the shifted screen-detected cancers are diagnosed contemporaneously with symptom-detected cancers. Also adding to the increase are the overdiagnosed cancers.

Cancer screening cannot lead to a reduction in cause-specific mortality if cancer incidence does not increase. If cancer incidence remains stable as cancer screening is introduced and uptake increases, diagnoses are not occurring earlier, and therefore prognosis cannot change. An increase in cancer incidence does not guarantee a cause-specific mortality reduction. The increase may be due to detection of overdiagnosed cancers or detection of cancers that would have the same prognosis regardless of detection in Phase B or Phase C (as defined in Chapter 2).

Example

Figure 8 displays, in a very simplistic manner, how introduction of cancer screening increases the number of cancers that are diagnosed. In the absence of cancer screening, three cancers are diagnosed due to symptoms in 2017 (the X cancers) and three cancers are diagnosed due to symptoms in 2018 (the Y cancers). In the presence of cancer screening, the three X cancers are either screen or symptom detected in 2017, the three Y cancers are screen detected in 2017, and the two Z cancers, the overdiagnosed cancers, also are diagnosed in 2017. The number of cancers diagnosed in 2017 in the presence of cancer screening is 5 more than would have been diagnosed in the absence of cancer screening. If this fictional population included 1000 individuals, the incidence rate for 2017 would be 3/1000 per year in the absence of cancer screening versus 8/1000 per year in the presence of cancer screening.

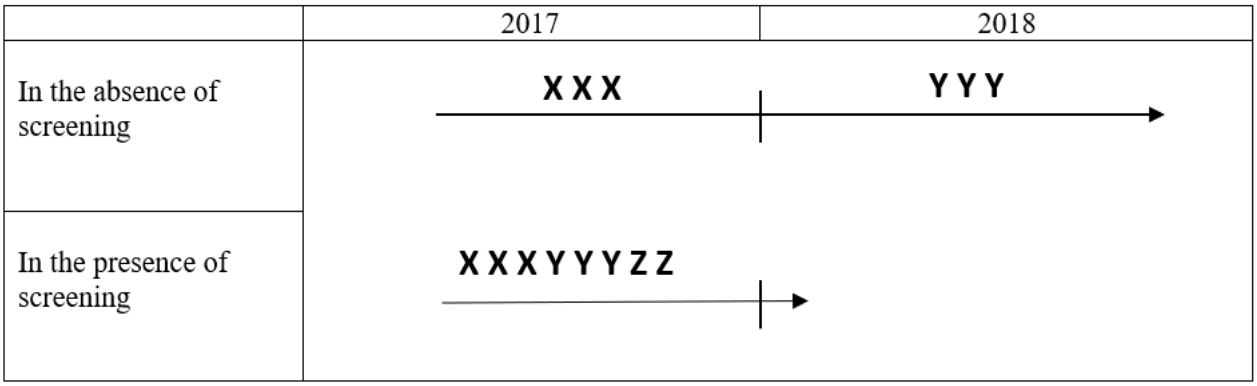


Figure 8: Cancer incidence in the presence and absence of cancer screening.

In the absence of cancer screening, X and Y cancers are symptom detected, with X cancers diagnosed in 2017 and Y cancers diagnosed in 2018. Z cancers are never diagnosed. In the presence of cancer screening, X cancers are still detected in 2017 though they may be screen or symptom detected. Y cancers are now screen detected in 2017. Z cancers (overdiagnosed cancers) are screen detected in 2017. Experience is fictional.

Figure 8 depicts what cancer screening is intended to do: detect cancers at an earlier point in time. Screen detection of the Y cancers in Figure 8 may lead to more favorable experiences for these patients, such as simpler treatment and better prognosis. Their detection could lead to a reduction in cause-specific mortality, although at the point of diagnosis, it is impossible to know. Of course, conjecture is possible and frequently happens. For example, diagnosis at an earlier point in time may be interpreted as advantageous, which can then be prematurely interpreted to mean that cancer screening will lead to a reduction in cause-specific mortality.

Figure 8 does not depict what happens in the presence of cancer screening in 2018, but if the graph were extended for additional years, the same general pattern of shifting should hold. The specifics of the shift depend on how cancer screening interferes with cancer's natural history, other changes in cancer's natural history, as well as changes in cancer screening uptake and performance. Barring any drastic changes in the three, the characteristics of the shift, such as degree and speed, should be fairly similar and stabilize after no more than a few screening rounds.

Stage at diagnosis

Cancer screening aims to detect cancer when prognosis is more favorable than it would have been if detected due to symptoms. Prognosis usually is related to stage at diagnosis. Most local-stage cancers are curable with resection, though these days, some regional- and distant-stage cancers can be cured with surgery, chemotherapy, immunotherapy, radiation, or a combination. As more non-local-stage cancers become curable, cancers diagnosed at those stages could have similar prognosis as those diagnosed at a local stage. But in today's cancer world, it is fair to assume that cure is most likely for local cancers and that those with treated local cancers live the longest.

The number of local-stage cancers is expected to increase when cancer screening is introduced. Soon after, a decrease in the number of regional- or distant-stage cancers is expected, as some cancers that were destined to be diagnosed at a later stage in the absence of cancer screening will have been detected at an earlier stage in the presence of cancer screening. The phrases stage shift and down staging are used to describe that situation. The phrases should be used to refer to changes in numbers, not changes in percentages. While it is true that a stage shift will lead to a change in the percentage of cancers for a given stage, percentages can be misleading if the number of local-stage cancers increase absent a decrease in regional- and distant-stage cancers, which can happen when cancer screening leads to overdiagnosis.

If a stage shift does not occur, cancer screening will not lead to a reduction in cause-specific mortality. Lack of a stage shift indicates no movement in the stage at diagnosis and thus no improvement in prognosis. But the presence of a stage shift does not guarantee a cause-specific mortality reduction. A stage shift reflecting a change from one stage to another that has similar prognosis would confer no reduction in cause-specific mortality. Length-weighted sampling could produce that situation in the instance of curable disease, while lead time could produce that situation in the instance of incurable disease.

Discussion of stage shifts have typically focused on the need to observe an increase in early-stage cancer rather than a reduction in late-stage cancer. But both are necessary for a cause-specific mortality reduction to be possible, and a reduction in distant-stage cancer is unlikely to be due to lead time, length-weighted sampling, or ovediagnosis. The use of distant-stage cancer as a possible surrogate for cause-specific mortality is discussed in Chapter 9.

Example

Table 12 displays fictional stage experience of the Figure 8 cancers in the absence and presence of cancer screening. Scenario 1 excludes overdiagnosed cancers, while Scenarios 2 and 3 include them. Scenarios 1 and 2 present a favorable change: two cancers that, in the absence of cancer screening, would have been diagnosed at a distant stage are, in the presence of cancer screening, diagnosed at a local stage. In Scenario 3, the two distant-stage cancers remain as such even in the presence of cancer screening.

Table 12: Stage distributions in the absence and presence of cancer screening

Disease stage	In the absence of cancer screening		In the presence of cancer screening	
	Cancers	N (%)	Cancers	N (%)
Scenario 1: X and Y cancers only (non-overdiagnosed); two distant cancers are now detected at a local stage				
Local	X, Y	2 (34%)	X, Y, Y, Y	4 (67%)
Regional	X	1 (17%)	X	1 (17%)
Distant	X, Y, Y	3 (50%)	X	1 (17%)
Scenario 2: X, Y, and Z cancers (overdiagnosed and non-overdiagnosed cancer); two distant cancers are now detected at a local stage.				
Local	X, Y	2 (34%)	X, Y, Y, Y, Z, Z	6 (75%)
Regional	X	1 (17%)	X	1 (13%)
Distant	X, Y, Y	3 (50%)	X	1 (13%)
Scenario 3: X, Y, and Z cancers (overdiagnosed and non-overdiagnosed cancer); two distant cancers are detected at the same stage as in the absence of cancer screening.				
Local	X, Y	2 (34%)	X, Y, Z, Z	4 (50%)
Regional	X	1 (17%)	X	1 (13%)
Distant	X, Y, Y	3 (50%)	X, Y, Y	3 (38%)

X, Y, and Z cancers are defined in Figure 8. Experience is fictional.

The numbers of local-stage cancers increase and the numbers of distant-stage cancers decrease in Scenarios 1 and 2. The inclusion of the overdiagnosed cancers in Scenario 2 presents a more favorable picture than in

Scenario 1, but it is an overly-optimistic picture, as the overdiagnosed cases cannot contribute to a cause-specific mortality reduction, should one exist. In Scenario 3, the distant-stage cancers are detected at the same stage, regardless of cancer screening. Screening cannot reduce cause-specific mortality as no stage shift occurred; rather, it has led to the unnecessary detection of the two overdiagnosed cancers. Note that in Scenario 3 the stage-specific numbers do not suggest down staging, but the percentages, when examined alone, do.

Case survival

Measures of case survival will increase when cancer screening is introduced. Cancer screening leads to increased case survival because, for screen-detected cancers, the date of diagnosis occurs earlier (by the amount of lead time) than it would have in the absence of cancer screening. Yet our ability to interpret changes in case survival in the presence of cancer screening, relative to case survival in the absence of cancer screening, is impaired because we do not know what the date of diagnosis or date of death would have been in the absence of cancer screening for a given individual. The fictional Y and Z cancers in Figure 8, in conjunction with additional fictional experience in Table 13, will be used to demonstrate how case survival could change with cancer screening. Mean and median case survival are presented for ease of explanation, although relative case survival will change as well.

If case survival does not increase after cancer screening's introduction, cancer screening will not lead to a reduction in cause-specific mortality. A lack of increase indicates that diagnoses are not occurring earlier and that lives are not being lengthened. It is virtually impossible, however, for case survival not to increase when cancer screening occurs, because shifting the date of diagnosis to an earlier point in time is at the core of cancer screening. An increase in case survival does not guarantee a cause-specific mortality reduction, however. Lead time is usually responsible in that instance, but length-weighted sampling and overdiagnosis can lead to detection of cancers that will have the longest case survival because they have the most favorable prognosis.

Case survival seems to be the most frequently misinterpreted intermediate outcome. Increases in 5-year case survival are quoted as evidence that cancer screening saves lives, but lead time is rarely mentioned as a contributing factor and possible explanation for the observation.

Example

Table 13 presents date of diagnosis, date of death, and case survival for the Y and Z cancers in the presence and absence of cancer screening. The experience of each Y cancer represents a different way that lead time can change case survival. Y_1 is screen detected but the date of death does not change. Case survival, which increases from 12 months to 20 months, suggests a benefit though. Y_2 is screen detected but dies three months earlier than he or she would have in the absence of cancer screening, perhaps due to toxicity of cancer treatment. Case survival increases, though, from 10 months to 15 months because of lead time. Y_3 benefits from screen detection. Case survival increases from 32 to 66 months, though the extension of life is only 26 months. The remainder of the 34-month increase in case survival, 8 months, is lead time.

Table 13: Case survival in the absence and presence of cancer screening

Cancer	In the absence of cancer screening			In the presence of cancer screening		
	Date of diagnosis	Date of death	Case survival	Date of diagnosis	Date of death	Case survival
Y_1	2/1/18	2/1/19	12 months	6/1/17	2/1/19	20 months
Y_2	2/1/18	12/1/18	10 months	6/1/17	9/1/18	15 months
Y_3	2/1/18	10/1/20	32 months	6/1/17	12/1/22	66 months
Z_1	Never diagnosed	6/1/21	Not relevant	6/1/17	6/1/21	48 months
Z_2	Never diagnosed	9/1/21	Not relevant	6/1/17	10/1/20	36 months

X, Y, and Z cancers are defined in Figure 8. Experience is fictional.

The Z cancers do not have a measure of case survival in the absence of cancer screening because they were overdiagnosed. Detection of an overdiagnosed cancer cannot result in extension of life due to treatment. It can result, however, in premature death. Early death can occur in the instance of an adverse event related to cancer screening, diagnostic evaluation, or treatment. In addition, cancer patients have been shown to be at elevated risk of suicide (1).

Z_1 is diagnosed due to cancer screening but his or her date of death does not change. Z_2, on the other hand, dies sooner than he or she would have in the absence of cancer screening. Such a situation needs to be considered when weighing benefits and harms of cancer screening. It also is possible that the experience of having cancer will lead to lifestyle changes that improve overall health and extend life. Both situations would be reflected in mortality rates. The impact of lifestyle changes that do not affect length of life yet lead to improved quality of life is not usually considered when evaluating cancer screening efficacy or effectiveness.

In the absence of cancer screening, the three non-overdiagnosed cancers would have a median case survival of 12 months and mean case survival of 18 months. In the presence of cancer screening, the 5 detected cancers would have median case survival of 36 months and mean case survival of 37 months. Yet only 1 of 5 screen-detected cases lived longer that he or she would have in the absence of cancer screening.

Cancer screening's impact on definitive outcomes

The two definitive outcomes in cancer screening are cause-specific mortality and all-cause mortality. The mortality outcomes are called definitive because it is impossible for them to be biased by the three screening phenomena, as is discussed in the next section of this chapter. That does not mean, however, that they cannot be affected by other factors, something that may not have been appreciated when the term definitive was bestowed upon them many years ago.

Mortality rates and the three screening phenomena

Cause-specific and all-cause mortality rates are not affected by lead time, length-weighted sampling, or overdiagnosis. They are not affected by lead time because date of diagnosis is not used to calculate mortality rates. They are not affected by length-weighted sampling or overdiagnosis because deaths are not restricted to those individuals whose cancer was screen detected.

Recall from Chapter 4 that the numerator in cause-specific mortality rates includes all deaths due to the cause of interest and the numerator in all-cause mortality rates includes all deaths. The denominator includes all persons at risk of death, not only those who were screened. It is for these reasons that mortality rates reflect the impact of cancer screening on the entire population eligible to be screened. They incorporate cancer screening's successes as well as its failures, should either or both exist. Successes are extension of life among those screened. Failures are missed opportunities for early detection due to many factors, including limitations of the test, shortcomings in test interpretation, and non-adherence to cancer screening or diagnostic evaluation for a positive test.

Cause-specific mortality rates

The calculation of a cause-specific mortality rate to assess cancer screening is straightforward, as was demonstrated in Chapter 4. The numerator includes all individuals who died of the cause of interest. The underlying assumption in the calculation is that the numerator correctly captures all relevant deaths. Unfortunately, errors in cause of death assignment are known to occur (2). The cause of death recorded on the death certificate may not be the true cause of death.

When attempting to assess whether cancer screening can reduce cause-specific mortality, it is advised to classify any death that occurred as an adverse effect of the cancer screening process as a cause-specific death. The reason for that is to measure all screening failures. Any death that occurs due to the cancer for which screening is occurring is clearly a failure of the cancer screening process. However, any death due to an adverse effect of the cancer screening process also should be considered a failure because it would not have happened (or might have happened later) if cancer screening had not occurred. Identifying those deaths is a challenge because the death certificate is unlikely to indicate the sort of information that is necessary to link the death to the cancer screening process.

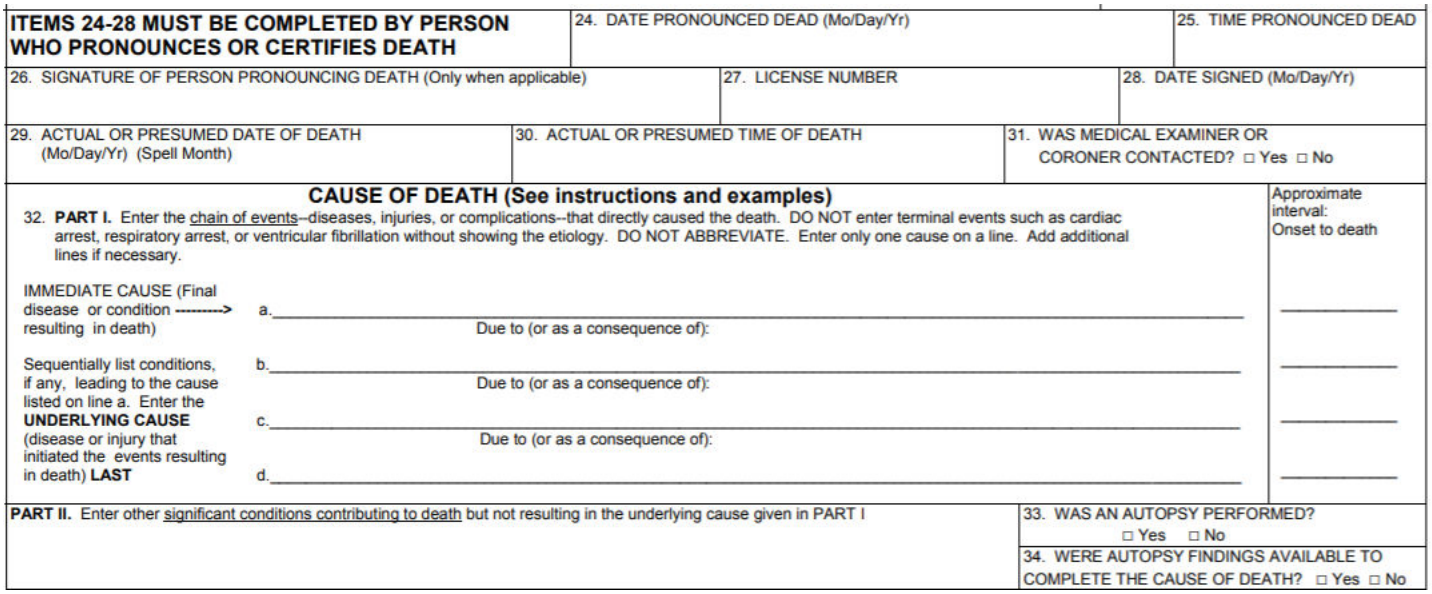
The next section addresses two phenomena that affect the ability of cause-specific mortality rates to measure what we want them to measure.

Sticking diagnosis, slippery linkage, and assessment of cancer screening

Sticking diagnosis occurs when the cancer of interest is erroneously assigned to be the cause of death, which can happen due to cancer's reputation for lethality. Sticking diagnosis can happen in the instance of screen- or symptom-detected cancer, but because incidence rates typically increase with cancer screening, sticking diagnosis generally leads to cause-specific mortality rates that are higher than they should be. In that instance, cancer screening could appear to not reduce cause-specific mortality when it actually does.

Slippery linkage occurs when death certificates do not capture a direct or downstream consequence of cancer screening, or do not capture it in such a way that it can be linked to cancer screening. Slippery linkage leads to cause-specific mortality rates that are lower than they should be and could lead to the conclusion that cancer screening does reduce cause-specific mortality when it actually does not. Slippery linkage would be at work in the instance of death due to a bowel perforation sustained during a screening colonoscopy, or development of fatal breast cancer caused by radiation from extensive imaging for an abnormality observed on lung cancer screening. In the former example, screening played a part in the death, and while the death certificate is likely to note a medical misadventure, it probably will not reflect the reason for the colonoscopy. In the latter example, it would be all but impossible to recognize the death as a downstream effect of cancer screening.

The section of the US standard death certificate that covers cause of death is presented as Figure 9. Note that immediate causes, underlying causes, and significant medical conditions can be listed on the death certificate. Oftentimes a single underlying cause of death is derived using all entries according to rules set forth by the National Center for Health Statistics (NCHS); that cause of death is defined by the World Health Organization as "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury" (3).



ITEMS 24-28 MUST BE COMPLETED BY PERSON WHO PRONOUNCES OR CERTIFIES DEATH	24. DATE PRONOUNCED DEAD (Mo/Day/Yr)	25. TIME PRONOUNCED DEAD
26. SIGNATURE OF PERSON PRONOUNCING DEATH (Only when applicable)	27. LICENSE NUMBER	28. DATE SIGNED (Mo/Day/Yr)
29. ACTUAL OR PRESUMED DATE OF DEATH (Mo/Day/Yr) (Spell Month)	30. ACTUAL OR PRESUMED TIME OF DEATH	31. WAS MEDICAL EXAMINER OR CORONER CONTACTED? □ Yes □ No

CAUSE OF DEATH (See instructions and examples)

32. **PART I.** Enter the chain of events--diseases, injuries, or complications--that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

Approximate interval: Onset to death

IMMEDIATE CAUSE (Final disease or condition ---------> resulting in death) a. _____ Due to (or as a consequence of):

Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**
b. _____ Due to (or as a consequence of):
c. _____ Due to (or as a consequence of):
d. _____

PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I

33. WAS AN AUTOPSY PERFORMED? □ Yes □ No

34. WERE AUTOPSY FINDINGS AVAILABLE TO COMPLETE THE CAUSE OF DEATH? □ Yes □ No

Figure 9: Questions 24 to 34 of the US Standard Certificate of Death (Rev 11/2003) (3)

In the US, researchers can obtain certain death certificate fields as long as the scientific rationale is strong. Requestors often forget, however, that death certificates are not completed with biomedical research in mind. To use death certificate data for research purposes requires an understanding of the rules used to complete them and recognition of their limitations. Additional information about death certificate completion and cause of death coding can be found at the website of the National Center for Health Statistics' (NCHS) National Vital Statistics System website (3).

Cause of death review

To arrive at accurate cause of death information, it may be necessary to review medical records that document the events leading to death. A review of every death could be done, though a thoughtfully-chosen, algorithm-driven, subset of deaths, as was done in the National Lung Screening Trial (NLST) (4), will save time and effort. Death review is usually a large undertaking, given the medical records that must be obtained and the person-power to review them. Nevertheless, death review can help to reverse death certificate cause of death assignment errors caused by sticking diagnosis and slippery linkage.

Cause-specific mortality rates: definitive enough?

Given the possibility of sticking diagnosis and slippery linkage, it is fair to question whether cause-specific mortality outcomes are definitive. Obviously no outcome will be perfect, and by reviewing medical records one may be able to circumvent much of the error that is possible with assigned cause of death. The "definitive-ness" for a given cancer is primarily dependent on the extent of sticking diagnosis and slippery linkage that goes uncorrected. Cause of death in the NLST was expected to be affected by slippery linkage and sticking diagnosis given comorbidities that are often experienced by heavy smokers and the perceived lethality of lung cancer. However, a comparison of death certificate cause of death and death review cause of death indicated that disagreement was minimal (5). The authors concluded death review may not be necessary in lung cancer screening.

It is possible to create a scenario, however far-fetched, in which a cause-specific mortality reduction could be explained by something other than cause of death errors created by slippery linkage. A reduction in mortality could be due, for example, to rapid elimination of a powerful risk factor or rapid introduction of a highly effective treatment. Such dramatic changes would have to be timed just so and be highly correlated with the act of being screened for the cancer of interest to explain away what appears to be a beneficial effect of population-

based cancer screening. Given that the cancer landscape has never been a fast-changing one, that scenario is unlikely. Even cigarette smoking, an exceptionally strong cancer risk factor, took years to make its effect known, and universal smoking cessation, should it ever occur, also would take years for its impact to be realized. Certain molecularly targeted cancer therapies appear to be miracle cures, but they are available for only a few tumor types. The impact of concurrent changes on assessment of cancer screening tests is discussed further in Chapter 9.

Definitive outcomes are considered by most to be superior to intermediate outcomes when assessing the ability of cancer screening to reduce cause-specific mortality.

All-cause mortality

All-cause mortality rates are not affected by sticking diagnosis and slippery linkage because no cause of death is necessary to calculate them. Yet all-cause mortality is not a practical outcome in assessment of most cancer screening tests. Reduction in cause-specific mortality of a typical magnitude (perhaps about 20%) will lead to a small relative reduction in all-cause mortality, because death due to a single cancer usually represents a small percentage of all deaths.

The NLST was an exception: a statistically significant 20% lung cancer mortality reduction was accompanied by a statistically significant 7% all-cause mortality reduction. However, lung cancer deaths accounted for about 25% of all deaths, and when those deaths were excluded, the reduction in all-cause mortality was no longer statistically significant. Results from the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) are more in line with what typically happens. The trial observed a statistically significant 26% reduction in colorectal cancer mortality, though the colorectal cancer deaths represented only about 2% of all deaths. An insignificant reduction in all-cause mortality of about 2% was observed, even with accumulation of over 800,000 person years in each arm.

Randomized controlled trials of cancer screening that utilize an all-cause mortality outcome would require extremely large numbers of individuals to have the necessary statistical power to detect typical mortality reductions. Large simple trials with an all-cause mortality outcome have been proposed (6,7) but have their own shortcomings. Large numbers of screening centers might allow for recruitment of hundreds of thousands of participants, but would require more autonomy on the part of those screening centers, leading to challenges regarding rigor, such as uniform application of the screening protocol. A diffuse trial structure would make tracking factors that impact the outcome, such as contamination, difficult.

References

1. Zaorsky NG, Zhang Y, Tuanquin L, Bluethmann SM, Park HS, Chinchilli VM. Suicide among cancer patients. Nat Commun. 2019 Jan 14;10(1):207. PubMed PMID: 30643135.
2. Smith Sehdev AE, Hutchins GM. Problems with proper completion and accuracy of the cause-of-death statement. Arch Int Med. 2001 Jan 22;161(2):277–84. PubMed PMID: 11176744.
3. Revisions of the US Standard Certificates and Reports [Internet]. Atlanta: Centers for Disease Control and Prevention. c [updated 2017 Aug 17; cited 2019 Oct 22]. Available from: https://www.cdc.gov/nchs/nvss/revisions-of-the-us-standard-certificates-and-reports.htm.
4. Marcus PM, Gareen IF, Miller AB, Rosenbaum J, Keating K, Aberle DR, Berg CD. The National Lung Screening Trial's Endpoint Verification Process: determining the cause of death. Contemp Clin Trials. 2011 Nov;32(6):834–40. PubMed PMID: 21782037.
5. Marcus PM, Doria-Rose VP, Gareen IF, Brewer B, Clingan K, Keating K, Rosenbaum J, Rozjabek HM, Rathmell J, Sicks J, Miller AB. Did death certificates and a death review process agree on lung cancer cause of death in the National Lung Screening Trial? Clin Trials. 2016 Aug;13(4):434–8. PubMed PMID: 27006427.

6. Peto R, Collins C, Gray R. Large-scale randomized evidence: Large, simple trials and overviews of trials. J Clin Epidemiol. 1995 Jan;48(1):23–40. PubMed PMID: 7853045.
7. Prasad V. Powering cancer screening for overall mortality. ecancer 2013 [Internet],7:ed27. Available from: https://ecancer.org/en/journal/editorial/27-powering-cancer-screening-for-overall-mortality.

Chapter 6. Experimental research designs

The first five chapters of this primer present important concepts in cancer screening and evaluation of its data. Examples were provided to reinforce concepts and interpretation, but most were limited, fictional, and not intended to demonstrate how cancer screening efficacy and effectiveness are formally evaluated. Chapter 6 and Chapter 7 present the research study designs that are used to generate the data necessary for cancer screening assessment. Design features, analysis features, and strengths and weaknesses will be presented for each. A synopsis of at least one published report, along with its reference, will be provided for each design. Statistical theory will not be discussed.

There are two classes of study designs: experimental and observational. Randomized controlled trials (RCTs) are experimental study designs and are discussed in this chapter. All other study designs presented in this primer are observational. They are discussed in Chapter 7. In general, efficacy is assessed using RCTs, while effectiveness is assessed using observational designs, though exceptions exist. Recall from Chapter 1 that efficacy refers to the ability of cancer screening to reduce cause-specific mortality in a highly controlled and near ideal setting, and effectiveness refers to the ability of cancer screening to reduce cause-specific mortality in a traditional community health care setting, one that provides numerous and varied services and faces typical US health care challenges. Pragmatic RCTs, which will be discussed, are conducted in community settings. They usually are classified as effectiveness research but are presented in this chapter given their experimental nature.

Readers who would like to learn more about experimental research can consult Fundamentals of Clinical Trials, by Friedman, Furburg, and DeMets (1).

An overview of experimental study designs

RCTs are experimental because the intervention is assigned at random rather than chosen by the study participant or study researcher. Randomization can occur individually for each participant (individual-level randomization) or for entities (cluster-level randomization). Most RCTs are composed of two groups, referred to as trial arms. When the number of participants is large enough, randomization will create, with high probability, trial arms that are equivalent prior to administration of the intervention. Equivalent means that the distribution of all risk and protective factors, both measured and unmeasured, is the same in each trial arm. Large enough means that the trial has adequate statistical power, which can be determined by published formulas (2). The arm that does not receive the intervention is treated as the counterfactual experience of the intervention arm, which is the hypothetical experience that the intervention arm would have had if the intervention had not been administered. It is the counterfactual principle that allows the outcome to be fully and solely attributable to the intervention, as randomization greatly minimizes the possibility of confounding. In the context of cancer screening, confounding occurs when a third factor is related to both screening activity and cause-specific mortality, and will be discussed in detail in Chapter 7.

All RCTs are prospective in nature. Individual-level and cluster-level RCTs share many features. Those features will be discussed in the context of individual-level trials. The manners in which cluster-level RCTs differ will be presented afterwards. Pragmatic RCTs, a type of experimental design used in patient-centered research, will be discussed at the end of the chapter. Pragmatic RCTs incorporate randomization but allow for crossover (that is, assignment to the other trial arm) if the randomization assignment is counter to patient preference.

Individual-level randomized controlled trials of screening

Design features

Individual-level cancer screening RCTs involve randomization of each participant to a trial arm. RCTs have at least one intervention arm and one control arm. For simplicity's sake, a trial with one intervention arm and one control arm will be used to present this chapter's material.

Intervention arm participants are offered the screening test or screening regimen that is hypothesized to be of benefit. Control arm participants are offered either no cancer screening test or cancer screening with the standard of care screening test or regimen. Control arm participants who are offered no cancer screening may be offered an unrelated exam, such as a glaucoma exam, to engender good will and to facilitate follow up for trial outcomes.

Ascertainment of all information, but most importantly intermediate and definitive outcomes, must be conducted with the same amount of rigor for each arm. Death review should be considered. Death reviewers should be blinded to trial arm.

An RCT is designed to have a pre-specified number of screening rounds and years of follow-up. Screening rounds in an RCT are typically called T0, T1, and so on. T0 refers to the first screen and also may be called the prevalence screen, with later screens called incidence screens. A stop-screen RCT is one in which follow-up continues after screening stops. All RCTs should have interim analysis and data monitoring plans so that a trial can be stopped early if evidence is overwhelming that the intervention is efficacious or it is not.

Analysis features

The primary outcome in a cancer screening individual-level RCT is a cause-specific mortality rate ratio (and its 95% confidence interval), which is the ratio of the cause-specific mortality rate in the intervention arm to the cause-specific mortality rate in the control arm. Rate ratios that are statistically significant and lower than 1 indicate that the intervention reduced cause-specific mortality relative to whatever was received (if anything) by the control arm. A rate ratio that is not significantly different from 1 indicates that there is no evidence to suggest that the intervention reduces cause-specific mortality, relative to whatever was received (if anything) by the control arm. An all-cause mortality rate ratio usually will be reported as well, although as discussed in Chapter 5, cancer screening RCTs rarely have the statistical power to detect a significant reduction in all-cause mortality because death due to the cancer of interest usually represents a small percentage of all deaths. Intermediate outcomes often are reported as well.

If it is desired to generate an adjusted ratio due to suspected confounding, proportional hazards models can be used. Confounding is unlikely in well-designed and well-executed RCTs, but it is often worthwhile to explore the possibility. If confounding by measured factors is not present, the unadjusted and adjusted ratios will be similar. Proportional hazards models do not produce rate ratios; instead, they produce hazard ratios, which reflect the instantaneous risk of death. Hazard ratios are comparable to mortality rate ratios as the two types of ratios produce the same information: a relative measure of the chance of death in the intervention arm versus the chance of death in the control arm.

From the counterfactual principle comes the expectation that, prior to application of the intervention, the same number of cancers and cancer deaths would emerge in the two trial arms as time passes. Thanks to randomization, the intervention arm participants have counterparts in the control arm who would have the same experience, including cancer diagnosis and death, if screening did not occur. The intervention arm will quickly begin to accrue more cancer cases than the control arm once screening begins, primarily because of lead time. In the absence of overdiagnosis, he number of cancers is expected to equalize at some point after screening

stops, a phenomenon called catch-up. In the presence of overdiagnosis, catch-up does not occur, because screening found cancers whose control arm counterparts do not present in the absence of screening. A stop-screen design allows the question of overdiagnosis to be addressed by comparing the numbers of cancers in the two arms at a point in time after screening ceases. The appropriate point in time is based on beliefs about the natural history of disease. A stabilization of the difference in the number of cancers as time progresses is a good indication that catch-up is complete. That stable difference is the magnitude of overdiagnosis. This method for calculating overdiagnosis is called the excess incidence method. Another method for estimating overdiagnosis is discussed in Chapter 9. Assessing overdiagnosis in an RCT that does not utilize a stop-screen design cannot be done unless the length of the trial is longer than the longest of lead times. With a long enough observation period, the difference in cancer incidence between the trial arms will stabilize; the difference at that point is the magnitude of overdiagnosis.

Cessation of screening can lead to dilution of the mortality rate ratio. Dilution occurs when a mortality rate ratio that suggested a benefit of screening moves closer to a null result (no benefit; a rate ratio of 1) as time passes without screening. The counterfactual principle explains why dilution occurs: after screening ends, the trial arms eventually will return to their pre-intervention states, a time when they were equivalent in terms of their mortality rates. Any beneficial effect of cancer screening will eventually cease. An RCT that does not utilize a stop-screen design will not experience dilution.

Most RCTs randomize in a 1-to-1 fashion, leading to equal sample sizes in the two arms. Discussion of overdiagnosis and catch-up assumed equal numbers were randomized to each arm. If other randomization schemes are used, expectations regarding catch-up must be adjusted. For example, a trial that employs a stop-screen design and randomizes in a 2 (intervention) to 1 (control) fashion is expected to have twice as many cases in the intervention arm, if overdiagnosis does not exist.

Strengths and weaknesses

The greatest strength of a cancer screening RCT is that results can be attributed to the intervention and not to a confounding factor, but only if randomization achieved its goal of creating two equivalent groups. The chance of that happening is positively correlated with the size of the trial arms. Screening trials that have the necessary statistical power to properly assess a cause-specific mortality rate ratio are almost guaranteed to have equivalent groups as long as nothing in the randomization process is systematically awry.

Other potential differences in the experience of the arms must be considered when interpreting the findings of a cancer screening RCT. Outcome ascertainment methods need to be equivalent for the two arms, as does treatment for a given stage of cancer. Most RCTs collect extensive amounts of data; therefore, the aforementioned two conditions often can be assessed. However, it is important to remember that participants in the intervention arm will have more contact with trial staff during the screening period of the trial, which could lead to the two arms having different experiences at many points in the screening process.

Standardized application of the screening regimen is a strength. An RCT is thought to provide the most favorable setting in which to evaluate a screening regimen; all steps in the screening process, from invitation to treatment, tend to occur with an extra level of forethought and rigor.

Cancer screening RCTs are expensive and take a long time to complete. They require large numbers of participants for reasons of statistical power. If intervention arm participants do not receive the intervention of interest (referred to as non-compliance) or control arm participants do (referred to as contamination), statistical power may be compromised if the degree of observed non-compliance and contamination is greater than what was assumed when the trial was designed. In the instance of extreme non-compliance and contamination, the trial arms become indistinguishable and any comparison in mortality rates is meaningless. If the intervention is

available outside the trial and either is inexpensive or covered by health insurance, high rates of contamination are likely and may make an RCT impractical.

Example of an individual-level cancer screening RCT

There have been a number of cancer screening RCTs conducted, and they vary with regard to rigor and availability of information on their conduct. A well-conducted and a well-documented cancer screening RCT is the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO), which has been mentioned previously. Informative publications include the primary outcome papers (3-6) and methods and operations papers (2,7). The methods and operations papers will be useful to those who are planning to launch a trial or wish to learn more about the nuts and bolts of how cancer screening RCTs are carried out.

Cluster-level randomized controlled trials of cancer screening

Design features

A cancer screening cluster-level RCT is quite similar to an individual-level RCT. The only design difference is that cluster-level trials randomize groups rather than individuals. The number of groups must be at least two but can be more. If a group is randomized to receive the intervention, all eligible individuals in that group are invited to receive it. Groups often are geopolitical entities, such as counties or provinces. The groups to be randomized must be similar enough for the counterfactual principle to hold.

Analysis features

The same principles that hold for analysis of individual-level cancer screening RCTs hold for cluster-level cancer screening RCTs, except in one instance. A cluster-level RCT usually is analyzed at the cluster level, meaning that the cluster, rather than individual, is the unit of analysis (8). When analyzed at the cluster level, statistical analyses are straightforward, but results are applicable to only clusters. For example, a cause-specific mortality rate ratio of 0.80 indicates that clusters that were offered the intervention have a 20% reduction in cause-specific mortality rates relative to those clusters that were not, not that individuals who were screened had a 20% reduction in cause-specific mortality. The conclusions are not guaranteed to be directly applicable to the individuals who reside in those clusters, although many times they are interpreted as if they are.

It is inappropriate to analyze a cluster-level RCT as one would analyze an individual-level RCT; that is, it is inappropriate to use individuals as the unit of analysis rather than the cluster. Individuals within a cluster are rarely independent of one another. Lack of independence invalidates statistical assumptions on which methods rest and can lead to incorrect conclusions. There are, however, advanced statistical methods that can account for the lack of independence that accompanies individuals within clusters and allow for inferences to individuals (9).

Strengths and weaknesses

A cluster-level RCT of cancer screening can have very low rates of contamination if the new screening regimen is available in only certain clusters and it is difficult for individuals to cross into or receive medical services in other clusters. In addition, cancer mortality rates are often available for clusters that are geopolitical entities, eliminating the need for collection of mortality information as part of the RCT. However, compliance within a cluster can be low because individuals are usually not consulted before randomization. The number of clusters is often small, which can impact the ability of randomization to produce a true counterfactual group.

Cluster-level RCTs of cancer screening are difficult to carry out in places with opportunistic screening. In the US, randomization by state could be attempted, but ease of mobility and out-of-network health insurance policy benefits, not to mention entrepreneurial ventures, could foster contamination. Cluster-level RCTs of cancer screening may be more easily done in countries with government-administered health care, although a Swedish

cluster-level RCT of mammography screening still experienced non-negligible rates of contamination in the control arm (10).

Example of a cluster-level cancer screening RCT

In the United Kingdom (UK), the AgeX cluster-level RCT is looking at the impact of offering an additional breast cancer screen to women ages 47-49 and offering breast cancer screening every 3 years to women over 70 (11). Most of the 80 breast cancer screening centers in the UK's National Health Service are participating. Each center is a cluster and is randomized to the intervention arm or the control arm. All women in intervention arm clusters are invited to receive the age-appropriate additional screens. All women in control arm clusters are invited to receive the standard breast cancer screening regimen.

Pragmatic randomized controlled trials of cancer screening

RCTs of cancer screening usually have been carried out in highly controlled and near ideal settings. They have measured efficacy rather than effectiveness. Effectiveness can be addressed by pragmatic RCTs.

A pragmatic RCT is done in the reality of every day health care, which introduces many challenges that can hinder the ability of a cancer screening test to reduce mortality. Pragmatic trials usually have fewer eligibility criteria than in traditional RCTs. Pragmatic RCTs typically do not hire staff dedicated to trial operations; in other words, there usually are no extra resources for recruitment or compliance. Data collection above and beyond what is collected in usual care is not common.

Though randomization still occurs in pragmatic trials, patients may have the opportunity to receive what they want rather than what randomization assigns to them. While that may seem heretical to a strict clinical trialist, the goal of a pragmatic trial is to evaluate the impact of introducing a cancer screening test in a community health care setting. The impact reflects the fact that some patients will accept the test and some will not.

To learn more about pragmatic trials and patient-centered research in general, consult the National Institutes of Health (NIH) Collaboratory's Living Textbook of Pragmatic Clinical Trials (12), a website that presents expert consensus regarding special considerations, standard approaches, and best practices in the design, conduct, and reporting of pragmatic clinical trials.

Examples of pragmatic cancer screening RCTs

There are no completed pragmatic RCTs of cancer screening effectiveness, although there are at least two underway. The HOME trial, conducted in the Kaiser Washington managed care system, is examining the ability of self-sampling to increase cervical cancer screening uptake and effectiveness (13). Self-sampling could overcome certain real-world barriers to being screened, including lack of transportation to a clinic, lack of child care, and needing time off from work. It also could increase cervical cancer screening uptake among women who prefer not to receive a pelvic exam. The WISDOM trial, conducted in clinics in California and South Dakota, is comparing breast cancer screening regimens based on age to screening regimens based on risk (14). WISDOM is using what is known as a preference tolerant design, which encourages randomization but allows women to self-assign if they wish. The reason for choosing such a design was to maximize participation, a factor that may lead to better generalizability of results.

References

1. Friedman LM, Furburg CD, DeMets DL. Fundamentals of Clinical Trials, 4th ed. New York: Springer; 2010. 445 p.
2. Prorok PC, Marcus PM. Cancer Screening Trials: Nuts and Bolts. Semin Oncol. 2010 Jun;37(3):216–23. PubMed PMID: 20709206.

3. Andriole GL, Crawford ED, Grubb RL 3rd, Buys SS, Chia D, Church TR, Fouad MN, Isaacs C, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hsing AW, Izmirlian G, Pinsky PF, Kramer BS, Miller AB, Gohagan JK, Prorok PC; PLCO Project Team. Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. J Natl Cancer Inst. 2012 Jan 18;104(2):125–32. PubMed PMID: 22228146.
4. Oken MM, Hocking WG, Kvale PA, Andriole GL, Buys SS, Church TR, Crawford ED, Fouad MN, Isaacs C, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Rathmell JM, Riley TL, Wright P, Caparaso N, Hu P, Izmirlian G, Pinsky PF, Prorok PC, Kramer BS, Miller AB, Gohagan JK, Berg CD; PLCO Project Team. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. JAMA. 2011 Nov 2;306(17):1865–73. PubMed PMID: 22031728.
5. Schoen RE, Pinsky PF, Weissfeld JL, Yokochi LA, Church T, Laiyemo AO, Bresalier R, Andriole GL, Buys SS, Crawford ED, Fouad MN, Isaacs C, Johnson CC, Reding DJ, O'Brien B, Carrick DM, Wright P, Riley TL, Purdue MP, Izmirlian G, Kramer BS, Miller AB, Gohagan JK, Prorok PC, Berg CD; PLCO Project Team. Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. N Engl J Med. 2012 Jun 21;366(25):2345–57. PubMed PMID: 22612596.
6. Buys SS, Partridge E, Black A, Johnson CC, Lamerato L, Isaacs C, Reding DJ, Greenlee RT, Yokochi LA, Kessel B, Crawford ED, Church TR, Andriole GL, Weissfeld JL, Fouad MN, Chia D, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hartge P, Pinsky PF, Zhu CS, Izmirlian G, Kramer BS, Miller AB, Xu JL, Prorok PC, Gohagan JK, Berg CD. PLCO Project Team. JAMA. 2011 Jun 8;305(22):2295–303. PubMed PMID: 21642681.
7. Marcus PM. Editorial (Thematic Issue The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial: The Operations Behind a Herculean Task). Rev Recent Clin Trials. 2015;10(3):172. PubMed PMID: 26435287.
8. Higgins JPT, Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [Internet]. Updated March 2011, cited 2019 October 23. Available from: http://handbook-5-1.cochrane.org/chapter_16/16_3_1_introduction.htm.
9. Higgins JPT, Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [Internet]. Updated March 2011, cited 2019 October 23. Available from: http://handbook-5-1.cochrane.org/chapter_16/16_3_3_methods_of_analysis_for_cluster_randomized_trials.htm.
10. Tabár L, Fagerberg CJ, Gad A, Baldetorp L, Holmberg LH, Gröntoft O, Ljungquist U, Lundström B, Månson JC, Eklund G, Day NE, Pettersson F. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. Lancet. 1985 Apr 13;1(8433):829–32. PubMed PMID: 2858707.
11. NHS Breast Screening Program. The AgeX trial [Internet]. Oxford: University of Oxford. 2019 [cited 2019 Oct 23]. Available from: http://www.agex.uk/.
12. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials [Internet]. Bethesda (MD): NIH Health Care Systems Research Collaboratory. c2010-2019 [cited 2019 Oct 23]. Available from: https://rethinkingclinicaltrials.org/.
13. Winer RL, Tiro JA, Miglioretti DL, Thayer C, Beatty T, Lin J, Gao H, Kimbel K, Buist DSM. Rationale and design of the HOME trial: A pragmatic randomized controlled trial of home-based human papillomavirus (HPV) self-sampling for increasing cervical cancer screening uptake and effectiveness in a U.S. healthcare system. Contemp Clin Trials. 2018 Jan;64:77–87. PubMed PMID: 29113956.
14. Shieh Y, Eklund M, Madlensky L, Sawyer SD, Thompson CK, Stover Fiscalini A, Ziv E, Van't Veer LJ, Esserman LJ, Tice JA; Athena Breast Health Network Investigators. Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. J Natl Cancer Inst. 2017 Jan 27;109(5):djw290. PubMed PMID: 28130475.

Chapter 7. Observational research designs

An overview of observational study designs

Observational studies do not dictate the cancer screening regimens that their study subjects utilize. Instead, these studies collect data on individuals' cancer screening practices, cancer outcomes, and other factors if needed. Because no regimens are dictated, an observational study can capture information about and evaluate a variety of cancer screening practices, including use of different tests or cancer screening regimens. Observational studies can be retrospective or prospective in nature, with the distinction dependent on how and when individuals are chosen for study inclusion. Prospective observational studies of cancer screening track individuals as they move forward in time until the event of interest happens or the study is complete. Retrospective observational studies of cancer screening look at past experiences of individuals who have had the event of interest and others who have not. Prospective observational studies are said to sample based on exposure (cancer screening experience), while retrospective observational studies are said to sample based on outcome (death).

Observational studies provide weaker evidence than experimental studies because observational studies are subject to confounding. Confounding occurs when a third factor is associated with both the cancer screening practice and cause-specific mortality, meaning that the third factor is not equally present among groups of individuals with different cancer screening practices and is not equally present among groups of individuals with different cancer outcomes. An example comes from observational studies of colorectal cancer screening and colorectal cancer mortality. Individuals who exercise are more likely to have colorectal cancer screening and also are less likely to die of colorectal cancer. If an observational study observes a reduction in colorectal cancer mortality with cancer screening, we cannot be sure what is responsible. Is it cancer screening, exercise, a combination of both, or some unknown protective factor that is more likely among individuals who receive cancer screening and who exercise?

Confounding is a type of bias and leads to an incorrect estimate of the true relationship of cancer screening and cause-specific mortality. It is present in varying degrees in all observational studies and while it can be dampened using statistical methods, these methods cannot eliminate all confounding because it is not possible to measure all confounders or measure them accurately.

Observational studies usually are less expensive and easier to perform than experimental studies. There are many reasons for that: the study usually does not administer or pay for the cancer screening test; existing databases often are used; and retrospective studies do not need to wait for time to pass since the data already have been collected. Some prospective studies can take as long or longer than a randomized controlled trial (RCT), however. Retrospective studies are often used as a first pass to examine a hypothesis about a cancer screening test, especially if use of that test is prematurely disseminating into community practice.

Observational study designs that are frequently used in cancer screening assessment will be discussed: cohort, case-control, and ecologic. Single-arm studies, sometimes known as case series, will be presented as well. Readers who wish to learn more about observational research can consult Modern Epidemiology (3rd edition) by Rothman, Greenland, and Lash (1).

Cohort studies

Design features

A cohort is a group of people with something in common, either by nature or design, who are followed through time for an event of interest. Research cohorts can be created in one of two ways. Prospective cohorts are created in real-time; data is collected as time passes. Retrospective cohorts, also known as historic cohorts, are created

after data have been collected. These cohorts comprise extracted data from pre-existing data sources, such as Medicare or the medical records of health maintenance organization members. Retrospective cohorts usually are analyzed as if their data had been collected prospectively and generally are constructed for the purpose of addressing pre-determined research questions. In prospective cohort studies, individuals usually are recruited to actively participate in the study, but with retrospective cohort studies, individuals usually do not know that their data are being used to answer a specific research question.

Information on cohort experience can come from a variety of data sources. Prospective cohorts usually rely heavily on participant interviews and participant-completed questionnaires, and may use medical records to validate procedures and diagnoses. Retrospective cohorts typically have little-to-no additional information collected on them. In both instances, deaths can be confirmed with collection of death certificates, while death certificate cause of death can be verified with review of medical records that document clinical experiences prior to death. It is recommended that records for at least the three to six months prior to the date of death be considered (2).

No prospective cohorts have been established for the primary purpose of examining cancer screening effectiveness, though some have been established to collect other information about the cancer screening process in community settings. Some pre-existing prospective cohorts have been used to address effectiveness if cancer screening and death information are available. Many retrospective cohorts have been created to address a range of questions regarding cancer screening. Cohorts without information on cancer screening can be repurposed by collecting the needed information if it is available.

Analysis features

Cohort members choose their cancer screening regimens, which means that confounding is all but guaranteed. Therefore, outcome measures need to be calculated using statistical models that allow for adjustment for confounding variables. If timing of death (date of death or person-years of experience) is available, Cox proportional hazards regression or Poisson regression can be used, with the choice determined by assumptions regarding whether the hazard of death changes over time (3). Logistic regression can be used if information on timing of death is unavailable.

Poisson regression produces a cause-specific mortality rate ratio, Cox proportional hazards regression produces a cause-specific hazard rate ratio, and logistic regression produces an odds ratio, which estimates a risk ratio (also known as a relative risk) in the case of a rare outcome like cancer. Each ratio represents a measure of disease burden in the individuals who received the cancer screening regimen of interest divided by those who did not. When assessing cancer screening data, the exact measure used (mortality rate, hazard rate, or odds) is of less importance than the ratio that they will produce. The three methods, when applied to a cancer screening cohort with typical experience, usually will produce ratios that lead to the same conclusion.

Risk difference measures are sometimes used to describe how the absolute rather than relative magnitude of disease burden changes with cancer screening. To calculate a risk difference, the measure of interest (incidence rate, mortality rate, or hazard rate) in the presence of cancer screening is subtracted from the measure of interest in the absence of cancer screening. For example, a cause-specific mortality rate of 4 per 1,000 person-years in the absence of cancer screening and a cause-specific mortality rate of 3 per 1,000 person-years in the presence of cancer screening result in a risk difference of 1 per 1,000 person-years. Difference measures are more useful than relative measures when considering health care resource allocation.

Strengths and weaknesses

Cohort studies allow for evaluation of effectiveness, something of the utmost importance because the manner in which cancer screening is utilized in community settings is often quite different from the idealized regimens in RCTs. For example, an RCT might test an annual regimen, but the regimen that evolves in the community could

have longer or varied cancer screening intervals, especially when the cancer screening test is not fully acceptable to community members. Cohort studies also can be used to examine uptake of a new cancer screening test or test performance measures.

The value of a cohort often depends on the extent of confounding and timing of data collection. The chance and possible impact of confounding must be discussed whenever cohort data are presented. Regarding timing, cohort data are analyzed as if data were collected with the passing of time, meaning that collection of information on exposure (cancer screening and confounding) occurs before the outcome (cause-specific mortality) has occurred or is known. The data for some cohorts are collected that way, but for cohorts that are repurposed, researchers often need to collect information on past events. These data may be affected by recall bias, which happens when the passage of time results in data errors that then lead to incorrect estimates of the true relationship between cancer screening and cause-specific mortality. For example, let's say a cohort is used to examine the ability of colonoscopy to reduce colorectal cancer mortality. When participants are asked about cancer screening use in the past ten years, some may erroneously report a past colonoscopy when in fact their exam was a flexible sigmoidoscopy, which also reduces colorectal cancer mortality but to a lesser degree. Non-trivial error in reporting would lead to an observed association between colonoscopy and colorectal cancer that is weaker than the real association. The recall error led to measurement of the impact of receiving flexible sigmoidoscopy or colonoscopy, rather than only colonoscopy.

Needless to say, it's best to collect information as soon as possible after an exposure occurs. It is probably best to collect information on past cancer screening activities from medical records or health insurance claims rather than participant interviews, although medical records can be lost, and laws may be enacted that make use of both sources more difficult. Also, medical records do not always provide correct information. They are subject to human error and in some instances creative procedure coding to maximize insurance reimbursement.

To have adequate statistical power, cohort studies evaluating cancer screening usually need to be large and have a number of years of follow-up. Establishment of a new cohort and the infrastructure to track the experience of the participants will be expensive, although typically less than that of an RCT, as cancer screening activity and follow-up occurs as part of community care. Repurposing of an existing cohort can save money and time, but the need for additional data will lead to a reduction in resources saved.

Variations

A nested case-control study of cancer screening uses all cause-specific deaths in a cohort as cases, but only a subset of the rest of cohort members as controls. This design is used when additional data collection is needed and is expensive or time-consuming, as in the situation of needing to determine the indication for a medical test. Nested case-control studies of cancer screening are constructed and analyzed in the same manner as case-control studies of cancer screening; the only difference is that cases and controls are drawn from an established cohort rather than another source. Details of case-control studies of cancer screening, nested and otherwise, are presented later on in this chapter.

Examples of cancer screening cohort studies

The BCSC is a prospective cohort study of breast cancer screening. It is a cancer screening test registry: information on screening mammograms and other breast cancer screening imaging tests is collected, as well as information on the women who receive them. The unit of analysis is often a test rather than a woman. The BCSC is not intended to evaluate cancer screening effectiveness; instead, it strives to "assess and improve the delivery and quality of breast cancer screening and related patient outcomes". The cohort has been used to evaluate important issues in breast cancer screening, including screening adherence, test performance, and supplemental screening (4).

The Nurses' Health Study (NHS) and Health Professionals Follow Up Study (HPFS) are on-going prospective cohort studies, each designed to explore causes of major health conditions in the US. The NHS began in 1976 and the HPFS in 1985. Both studies added questions to their self-administered questionnaires in 1990 regarding receipt of lower endoscopy (colonoscopy or sigmoidoscopy). The researchers published findings on the impact of lower endoscopy on colorectal cancer mortality in 2013 (5).

Kaiser Permanente of North California (KPNC) is an integrated health care delivery system with more than 4 million members. Their extensive electronic health databases have been used to address many questions in cancer etiology and prevention, including cancer screening. An example of how a health care organization's databases can be used to conduct a retrospective cohort study of cancer screening can be found in KPNC's report on the long-term risk of colorectal cancer death after a negative colonoscopy (6).

Case-control studies

Design features

A case-control study is retrospective in nature, meaning that all exposures and events have occurred before the study begins. A case-control study includes cases, who are individuals who had the outcome of interest, and controls, who are individuals who did not have that outcome at a point in time that is determined by the case's experience. The design has been used extensively in cancer etiology studies. A case-control study often aims to include the universe of cases: all individuals who experience the outcome of interest during a specific time period. Controls are randomly sampled (usually within age strata) from the population that gave rise to the cases. In case-control studies of cancer etiology a population-based roster, such as a list of drivers' license holders, is used to sample controls.

In principle, case-control studies of cancer screening are the same as case-control studies of etiology. Cases are individuals who have died due to the cancer of interest. Controls are selected for a specific case, with random selection usually stratified on age and sex of the case. In addition, controls must not have been diagnosed with the cancer of interest prior to the case's diagnosis date; the reason is to ensure an equal and cotemporaneous opportunity for cancer screening. Some case-control studies of cancer screening have required that controls be alive on the date of the case's death. Cancer screening experience during a specific period (as discussed below) is compared in cases and controls.

Case-control studies of cancer screening usually select their cases and controls from health system patient rosters because access to medical records is a necessity. Medical records are used to determine whether a test was done for cancer screening as opposed to diagnostic evaluation, and obtain details of cancer diagnoses. As was noted earlier in this chapter, case-control studies of cancer screening also can be constructed by selecting cases and controls from an established cohort.

Analysis features

Case-control studies of cancer screening are designed and analyzed as matched case-control studies because exposure assignment for controls is defined by the experience of a case. Conditional logistic regression models are used to account for the matching and to adjust for other possible confounders. Logistic regression produces an odds ratio; in the instance of a case-control study of cancer screening, it is the ratio of the odds of receiving cancer screening among those who died of the cancer of interest divided by the odds of receiving cancer screening among those who did not die of the cancer of interest.

The primary challenge in analysis of case-control studies of cancer screening is assessing cancer screening exposure. An exposure window, one that reflects the period when cancer screening could have been beneficial to cases (Phase B as defined in Chapter 2), must be defined. The exposure window for cases ends no later than the

date of diagnosis, and usually ends prior to the date of diagnosis to exclude the period when cases were undergoing diagnostic evaluation. Controls are given a reference date, which corresponds to the final date of their matched case's exposure window. Only cancer screening experience prior to that date is considered to be in the exposure window. Cancer screening tests that occurred in the distant past should be excluded if there is reason to believe that they were done prior to the time the case's cancer was in Phase B.

Strengths and weaknesses

Case-control studies of cancer screening are retrospective research and can be done more quickly and inexpensively than cohort studies or RCTs. The number of cases is known at the start of the study, and controls are selected only if they match to a known case. Detailed information, such as that found in medical records, usually is needed to determine whether a test was for cancer screening and whether it occurred within the exposure window.

Confounding is a concern in case-control studies of cancer screening. Recall bias may be of concern if medical record abstractors are aware of the study hypothesis, or if medical records are systematically missing information, or are systematically unavailable. Because the exposure window must be inferred, it never will correctly capture the exact period in which cancer screening could have been of benefit to the cases. The exposure window must be thoughtfully chosen, and sensitivity analyses can be used to explore the impact of varying its definition.

The many methodologic challenges in design and analysis of case-control studies of cancer screening are discussed in Cronin et al (7) and Weiss (8).

Example of case-control studies of cancer screening

Using data on women residing in Saskatchewan, Pocobelli and Weiss conducted a case-control study of breast cancer mortality in relation to receipt of screening mammography (9). Saskatchewan has a universal health care system funded by the government, with nearly all residents eligible for coverage. About 90% of residents also are eligible for province-funded outpatient prescription drug benefits. The cases and controls for the cancer screening study were sampled from a larger study that utilized the roster of women with drug benefits. Cases were women who died due to breast cancer at 50–79 years of age during the years 1990–2008. Controls were selected for each case and were women who had the same birth year as the case and were not diagnosed with breast cancer prior to the case's date of diagnosis. Additional methodologic considerations, including definition of the exposure window, are discussed in the paper.

Ecologic studies

Design features

An ecologic study is the observational equivalent of a cluster-level RCT: the experience of groups, usually geopolitical entities, rather than individuals, is examined. The outcome of interest is cause-specific mortality rates, and the exposure is a measure of cancer screening utilization in the entity. Ecologic studies of cancer screening often compare cause-specific mortality rates for countries with different degrees of cancer screening utilization.

Analysis features

Data from ecologic studies of cancer screening often are presented using simple two-axis plots. Cause-specific mortality rates are plotted on one axis and their associated cancer screening use metric is plotted on the other. Percentage of eligible individuals screened is an example of a metric that has been used in ecologic studies. If

cancer screening reduces cause-specific mortality, a graph of cause-specific mortality rates on the y-axis and the cancer screening metric that measures use on the x-axis should produce a pattern of negative correlation. Figure 10 presents a fictional ecologic study in which utilization of cancer screening and cause-specific mortality are negatively correlated, as suggested by a fitted line that slopes downward. We cannot assume that cancer screening is the reason for the decrease in cause-specific mortality as other factors may be at play. In the instance of an ecologic study that suggests a reduction in cause-specific mortality, changes or regional differences in cancer treatment need to be carefully considered as confounders. Accuracy of the summary measures must be considered as well.

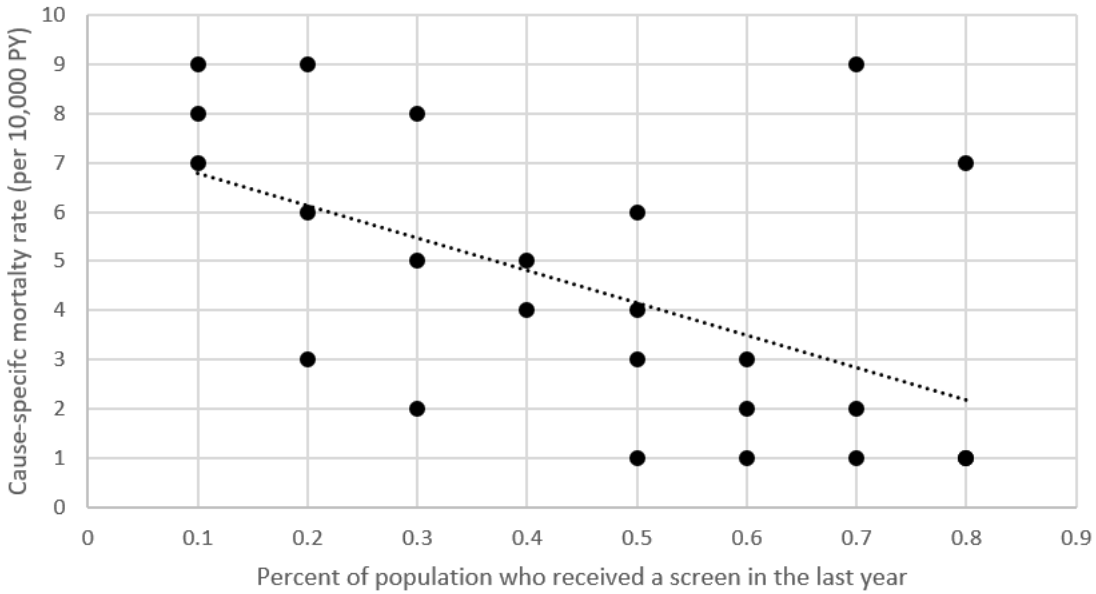


Figure 10: Plot of data from a fictional ecologic study that suggests a benefit of cancer screening. PY stands for person-years.

Ecologic studies can provide compelling evidence that cancer screening implementation has not led to reductions in cause-specific mortality for some cancer sites. In Figure 11, the cause-specific mortality rate hovers between 3.5 and 4.0 per 10,000 person-years regardless of cancer screening uptake, and the fitted line suggests no negative correlation. It is unlikely that such a pattern would mask a benefit of cancer screening due to confounding, as the confounding factor would need to increase cause-specific mortality and increase cancer screening use. Though very high-risk individuals do tend to receive cancer screening more frequently, they are a small percentage of the individuals in a population, and cannot drive entity-level rates unless most deaths from cancer occur in the high risk group.

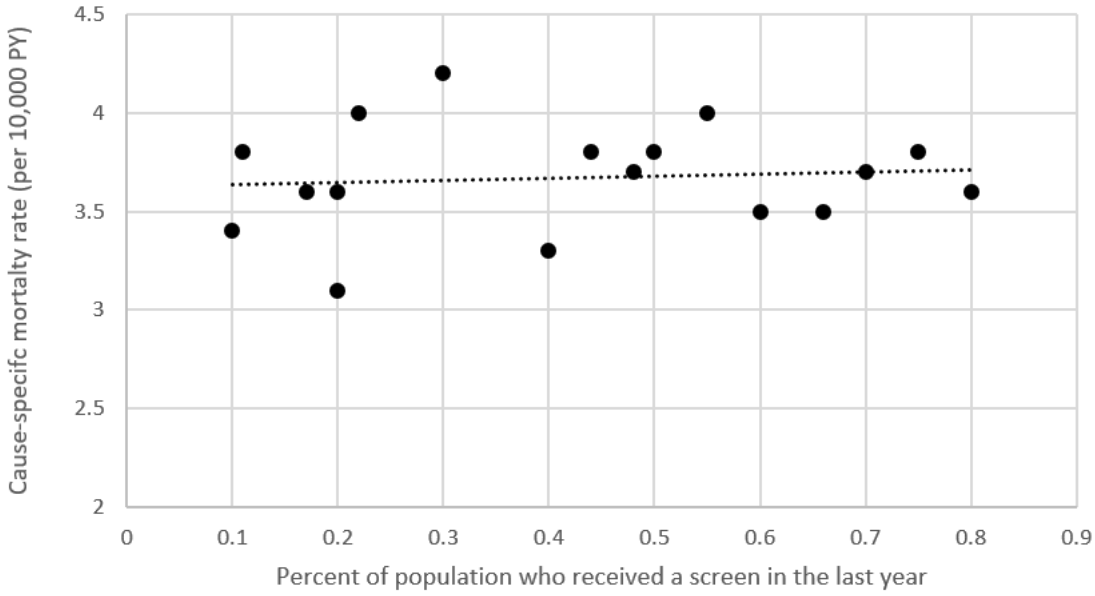


Figure 11: Plot of data from a fictional ecologic study that suggests no benefit of cancer screening. PY stands for person-years.

Strengths and weaknesses

Ecologic studies of cancer screening are usually easier to undertake and less expensive than individual-level observational studies. Cause-specific mortality rates are publicly available for geographic entities in the US and elsewhere. Obtaining data on cancer screening use within a group is more challenging, but data sources already exist in the US for some cancer screening practices. Electronic medical records or administrative claims from a universal provider such as Medicare could be used as well. It can, however, be challenging to identify a set of entities to compare. To determine appropriateness, it is useful to return to the counterfactual principle and choose entities that are as comparable as possible except for cancer screening practices.

Ecologic studies have a number of shortcomings. There can be confounding at the entity level, although linear regression can be used to adjust for some degree of the influence of confounding factors if that information is available. The results may not be applicable to the individuals within the entities, as was discussed in the context of cluster-level RCTs. Measures of cancer screening utilization that are not calculated in conjunction with individual-level medical records often are overestimates, as they can reflect use of cancer screening modalities that can be used for diagnostic purposes. For example, a measure of colonoscopy utilization derived by counting all colonoscopies performed will include both screening colonoscopies and diagnostic colonoscopies.

Variations

A time trend study is a type of ecologic study that examines changes in cancer mortality rates as time passes, with time as a marker for changes in cancer screening practice. Time trend studies are useful for examining changes in cause-specific mortality rates after cancer screening is introduced or after cancer screening regimens change. A two-axis graph can be used, with cause-specific mortality rates on the y-axis and year on the x-axis. A metric of cancer screening utilization, if available, can be included by using a second (right-sided) y-axis. Otherwise, milestones in cancer screening practices, such as the year that cancer screening was recommended for the first time, can be annotated. Figure 12 is an example, again fictional, of such a graph. Rates of cause-specific mortality decrease soon after widespread recommendation of cancer screening (1993). We cannot, however, assume that the decrease in mortality is due to cancer screening; other concurrent changes that might explain the observed pattern need to be considered before drawing any conclusions.

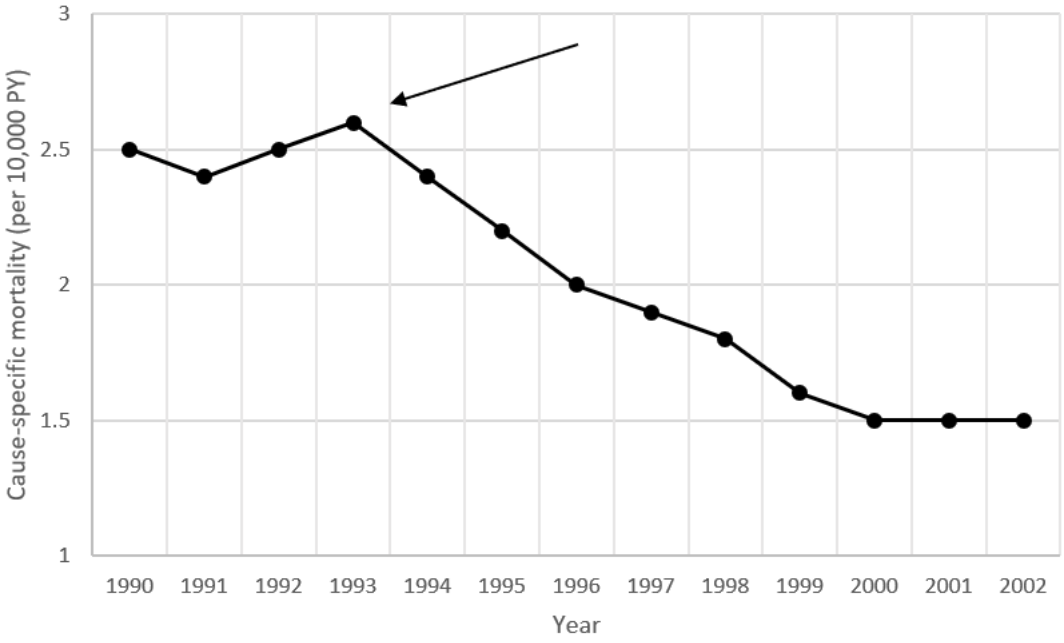


Figure 12: Time trends in cancer mortality before and after recommendation of cancer screening (1993). PY stands for person-years. Data are fictional.

In this fictional example, cause-specific mortality is stable prior to wide-spread recommendation of cancer screening in 1993. Cause-specific mortality begins to drop in 1994. It stabilizes around 2000, perhaps due to a leveling off of cancer screening utilization.

Examples of ecologic studies of cancer screening

A current controversy in breast cancer screening is whether reductions in breast cancer mortality are due primarily to screening or to improvements in treatment. To examine that question, Autier et al examined breast cancer mortality rates in neighboring European countries with different histories of screening use but access to similar treatments (10). Their ecologic analysis suggests that cancer screening has played only a minor role in improvements in breast cancer mortality.

The use of thyroid cancer screening in South Korea began to increase in 1999 when it was offered as a paid add-on test to the set of cancer screening tests offered for free through a national cancer screening program. No changes in thyroid cancer mortality were observed as utilization increased, and in 2013, use began to wane due to the evidence of no benefit and evidence of overdiagnosis (11).

Single-arm studies

Design features

In the context of cancer screening, a single-arm study refers to the experience, within a period of time, of a set of individuals who receive a screen in the context of a medical study. The screen is usually not standard of care. The test is considered to be experimental for cancer screening purposes, but a single-arm study is considered an observational study because it involves no randomization. Single-arm studies are a type of cohort study in which no participants are unscreened.

Analysis features

Cancer screening single-arm studies are used to assess performance of proposed tests, most notably, the ability of a test to lead to cancer detection at an early stage. These studies tend to enroll a small number of participants. Because there is no study comparison group, results either are presented on their own or compared with those from published literature or a population-level database such as SEER.

Strengths and weaknesses

Cancer screening single-arm studies are very limited in the information they can provide. The participants usually are a highly select group, suggesting that their experience is unlikely to be representative of what would occur in the general population. Participants typically are not chosen in a random fashion. They may be paid for their participation, or they may be required to pay to participate. Data collection usually does not include cause-specific mortality experience. Nevertheless, single-arm studies are a useful way to assess whether a proposed cancer screening test should receive further study.

Variations

A case series (or clinical series) is similar to a single-arm study. The difference is that case series include the clinical experiences of individuals not enrolled in a study. They are a culling of patients who have had the same exposure, in this case, cancer screening. Analyses examine their post-screening experience. The terms single-arm study and case/clinical series have been used interchangeably.

Examples of cancer screening single-arm studies

The Mayo Clinic initiated a single-arm study in 1999 to evaluate the performance of lung cancer screening with low dose computed tomography (LDCT)(12). At that time, there was evidence, though not definitive, that LDCT screening might reduce lung cancer mortality. The purpose of this study was to address some of the outstanding questions in LDCT screening, including the magnitude of false positive tests and the prevalence of adverse downstream effects. Participants were offered four annual LDCT screens. They were current or former cigarette smokers, with former smokers having quit less than 10 years ago. They also had at least a 20 pack-year history of smoking.

Two-in-one single-arm studies

Design features

In a two-in-one single-arm study, individuals are offered the chance to receive cancer screening with an experimental test in addition to and at the same time as the standard of care cancer screening test. Each participant receives both tests, and each test is evaluated without knowledge of the results of the other. Action is taken if either test result is positive.

Two-in-one single-arm studies have been used to determine if an experimental test has improved performance measures relative to the standard of care. They also have been used to examine whether an experimental screening test with a favorable feature (for example, lower cost, less invasiveness, or greater patient acceptability) has similar performance measures as the standard of care test. Two-in-one single-arm studies have been used to compare two tests already available in clinical settings. They also have been used to compare an experimental test with a diagnostic test, as diagnostic tests provide a definitive answer as to the presence of cancer.

A two-in-one single-arm study usually cannot be used to evaluate tests beyond diagnosis, although excessively optimistic speculation about the benefits of the experimental test is not uncommon.

Analysis features

The analytic focus of a two-in-one cancer screening single-arm study is a comparison of the performance of the tests. Of most interest is how and when the two tests disagree. Tables 14 and 15 present data from a fictional two-in-one single-arm study with 1000 participants. Table 14 compares positivity rates and Table 15 compares cancer diagnoses.

In Table 14, we see that both tests returned positive results for 80 individuals. Twenty individuals, however, received a positive experimental test result and a negative standard of care test result. The experimental test had a higher positivity rate, which may or may not indicate improvement over the standard of care test. A higher positivity rate could lead to a higher false positive rate or to additional cancer diagnoses. The meaning of additional cancer diagnoses is uncertain as well. They may represent cancers that are curable due to early detection, not curable regardless of early detection, or overdiagnosed.

In Table 15, we see that 35 cancers were diagnosed after both a positive standard of care and experimental cancer screening test. An additional 10 cancers were diagnosed as a result of a positive experimental test, even though the standard of care test was negative. We assume that these cancers were false negatives for the standard of care test, though we will never know what would have happened in the absence of the experimental test.

Table 14: Comparison of results from the standard of care and experimental screening tests

		Standard of care test		
		Positive result	Negative result	Total
Experimental test	Positive result	80	20	100
	Negative result	0	900	900
	Total	80	920	1000

Data are fictional

Table 15: Cancer diagnoses by results of standard of care and experimental screening tests

		Standard of care test		
		Positive result	Negative result	Total
Experimental test	Positive result	35	10	45
	Negative result	2	3	5
	Total	37	13	50

Data are fictional

Strengths and weaknesses

A two-in-one single-arm study may provide useful information if the standard of care test is known to reduce cause-specific mortality and the experimental test appears to have increased sensitivity and positive predictive value (PPV). A demonstrated increase in those two performance measures is usually interpreted to mean that the new test is superior, but to make that leap, one must assume that more asymptomatic diagnoses will lead to a greater reduction in cause-specific mortality. The existence of overdiagnosis and, for some cancers, equally efficacious treatment at a later stage, challenge that assumption.

A cancer screening two-in-one single-arm study cannot provide definitive evidence of efficacy or effectiveness. In the instance of a test that has not yet disseminated, results are best used to make decisions regarding the need for an RCT.

Examples of two-in-one single-arm studies

Blood-based biomarker cancer screening tests are of particular interest in colorectal cancer screening as the available screening tests, lower endoscopy and fecal testing, are not palatable to many individuals. Testing for circulating methylated SEP9 DNA has been under consideration as a way to screen for colorectal cancer. To examine the performance of SEP9 testing, individuals who were scheduled for screening colonoscopy were invited to give blood plasma samples prior to their colonoscopy preparation regimen (13). Performance measures for the SEP9 test were calculated using the results and ultimate outcome of colonoscopy screening, the gold standard in both colorectal cancer screening and diagnosis.

Screening mammography has evolved from the use of film-based to computer-based imaging. Film-based mammography only provides two-dimensional hard copy images. Digital mammography provides three-dimensional images that are read on a computer screen and can be manipulated to allow additional interpretation. The Digital Mammographic Imaging Screening Trial (DMIST) was designed to measure what were expected to be relatively small but potentially clinically important differences in diagnostic accuracy between digital and film mammography (14). Women enrolled in the trial received both tests on the same day, and each test was read by a different radiologist. Diagnostic evaluation was performed if either test was positive. Performance measures were calculated assuming that neither test was definitive.

All study designs: critical data elements

Most studies of cancer screening, regardless of study design, collect a large amount of data. Critical data elements are those that are necessary for proper assessment of screening performance, effectiveness, or efficacy. Other data elements may be collected for ancillary studies, or they may be collected for "what if" situations, but their collection must not jeopardize the collection of the critical data elements. Resources, including participant good will and staff time, always are limited.

Not every research endeavor will be able to collect every data element, even the critical elements. As a result, not every research endeavor will be able to answer every question. Even so, studies that do not collect every critical element may provide useful information, although the limitations of the research in the absence of such data must be clearly stated. The inability to collect most critical data elements should lead to questions about the value of the research.

Critical data elements for individual-level studies include date of birth; receipt, date and results of cancer screening tests; diagnosis of the cancer of interest; date of diagnosis; cancer characteristics (including stage, histology and location); age at death; date of death; cause of death. Indication for all relevant medical tests or procedures that are proximal to either the date of screen or the date of diagnosis should be collected to differentiate cancer screening from diagnostic evaluation. If that information is not available, any information that can be used to derive indication should be collected. Other valuable data elements include cancer treatment procedures, and adverse events of any medical procedure associated with cancer screening, diagnostic evaluation, or cancer treatment. Risk factors for the cancer of interest, as well as other potential confounders, should be collected, especially in observational studies.

Ecologic studies of cancer screening require entity-level cancer mortality rates and a metric of cancer screening use. One option for a test administered annually is to use the percent of residents who received a cancer screening test in the last 12 months. Other useful data elements include measures of cancer screening availability and characteristics of the entity that may predict cancer screening behavior, such as percent of residents with a college degree.

References

1. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. 758 p.
2. Miller, Anthony B. (Dalla Lana School of Public Health, University of Toronto). Email to: Pamela Marcus (National Cancer Institute). 2018 Oct 30.
3. Fagerstrom, Richard M (National Cancer Institute). Conversation with: Pamela Marcus (National Cancer Institute). 2018 Oct 25.
4. Breast Cancer Surveillance Consortium [Internet]. [Seattle]. [cited 2019 Oct 23]. Available from: https://www.bcsc-research.org/
5. Nishihara R, Wu K, Lochhead P, Morikawa T, Liao X, Qian ZR, Inamura K, Kim SA, Kuchiba A, Yamauchi M, Imamura Y, Willett WC, Rosner BA, Fuchs CS, Giovannucci E, Ogino S, Chan AT. Long-Term Colorectal-Cancer Incidence and Mortality after Lower Endoscopy. N Engl J Med. 2013 Sep 19;369:1095–1105. PubMed PMID: 24047059.
6. Lee JK, Jensen CD, Levin TR, Zauber AG, Schottinger JE, Quinn VP, Udaltsova N, Zhao WK, Fireman BH, Quesenberry CP, Doubeni CA, Corley DA. Long-term Risk of Colorectal Cancer and Related Deaths After a Colonoscopy With Normal Findings. JAMA Intern Med. 2018 Dec 17;179(2):153–60. PubMed PMID: 30556824.
7. Cronin KA, Weed DL, Connor RJ, Prorok PC. Case-control studies of cancer screening: theory and practice. J Natl Cancer Inst. 1998 Apr 1;90(7):498–504. PubMed PMID: 9539244.
8. Weiss NS. Clinical epidemiology: the study of the outcome of illness, 2nd edition. New York: Oxford University Press; 1996. 163 p.
9. Pocobelli G, Weiss NS. Breast cancer mortality in relation to receipt of screening mammography: a case-control study in Saskatchewan, Canada. Cancer Causes Control. 2015 Feb;26(2):231–7. PubMed PMID: 25471059.
10. Autier P, Boniol M, Gavin A, Vatten LJ. Breast cancer mortality in neighbouring European countries with different levels of screening but similar access to treatment: trend analysis of WHO mortality database. BMJ. 2011 Jul 28;343:d4411. PubMed PMID: 21798968.
11. Ahn HS, Kim HJ, Kim KH, Lee YS, Han SJ, Kim Y, Ko MJ, Brito JP. Thyroid cancer screening in South Korea increases detection of papillary cancers with no impact on other subtypes or thyroid cancer mortality. Thyroid. 2016 Nov;26(11):1535–40. PubMed PMID: 27627550.
12. Swensen SJ, Jett JR, Hartman TE, et al. Lung cancer screening with CT: Mayo Clinic experience. Radiology. 2003 Mar;226:756–61. PubMed PMID: 12601181.
13. Church TR, Wandell M, Lofton-Day C, Mongin SJ, Burger M, Payne SR, Castaños-Vélez E, Blumenstein BA, Rösch T, Osborn N, Snover D, Day RW, Ransohoff DF; PRESEPT Clinical Study Steering Committee, Investigators and Study Team. Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. Gut. 2014 Feb;63:317–25. PubMed PMID: 23408352.
14. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LF, Bassett L, D'Orsi D, Jong R, Rebner M; Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. New Engl J Med. 2005 Oct 27;353(17):1773–83. PubMed PMID: 16169887.

Chapter 8. Cancer prevention screening

Cancer screening for certain organs leads to detection of precancer. Detection of precancer is a form of cancer prevention if one uses the word cancer to exclusively mean invasive cancer, which is customary but not universal. Bretthauer and Kalager (1) have coined the phrase cancer prevention screening to refer to cancer screening practices that aim to detect precancer, and the phrase early detection cancer screening to refer to cancer screening practices that aim to detect invasive cancer.

Much of the theory and methodology regarding the assessment of cancer screening data arose during a time when the goal of cancer screening was to reduce cancer mortality by detection of invasive cancer at early stages. The reason for that goal was that technology was not advanced enough to detect precancer. It is fair to ask whether that theory and methodology still apply in our current era, one in which both invasive cancer and precancer disease are detected through cancer screening. It does, with one exception: the interpretation of changes in cancer incidence. The remainder of the principles laid out in the first seven chapters also are applicable to cancer prevention screening. This chapter presents material of relevance to cancer prevention screening for each of the first seven chapters of this primer.

Chapter 1: Foundations

The NCI website mentioned in Chapter 1 defines precancerous as "a term used to describe a condition that may (or is likely to) become cancer. Also called premalignant " (2). Most researchers use those terms as well as the term pre-invasive interchangeably. I prefer precancer because I find it to be broader in meaning than pre-malignant or pre-invasive. I use precancer to mean any change that is thought to be on the pathway to invasive cancer, be it DNA mutations in one cell or a tumor consisting of mutated cells that is on the verge of breaking through the basement membrane. In general, the material presented in Chapters 1 through 7 are relevant to whatever abnormality cancer screening aims to find.

Cancer prevention screening will be of value if some precancer detected through cancer screening would have become invasive and ultimately fatal cancer in the absence of cancer screening. Detection of precancer that does not meet that designation represents overdiagnosis. The definition of overdiagnosis can be modified slightly to be inclusive: screen-detected precancer or invasive cancer that never would have been diagnosed, either as precancer or invasive cancer, in the absence of cancer screening.

The overarching goal of both early detection cancer screening and cancer prevention screening is to reduce cause-specific mortality. We should not, however, assume that cancer prevention screening is merely early detection cancer screening at a very early stage, and that the benefits would be more extensive and harms less extensive than detection at a later stage. Precancer, at the time of detection, is not life-threatening as it cannot metastasize. Advances in technology have led to detection of more and more precancerous abnormalities with uncertain clinical relevance, creating quandaries for clinicians and patients. It is almost certain that overdiagnosis is more prevalent in cancer prevention screening as compared with early detection cancer screening. Even so, treatment of precancer has the potential to be less onerous than treatment of invasive cancer.

Chapter 2: Behind the scenes

Chapter 2 presented the four phase model (Figure 1). The model did not incorporate invasiveness of disease as it is immaterial to its purpose: to classify the stages of the natural history of cancer at which an abnormality, invasive or not, could be detected at an asymptomatic stage through cancer screening. While immaterial to the purpose of the model, the invasiveness of an abnormality is not immaterial to the assessment of cancer screening.

Chapter 3: Performance measures

The building blocks of performance measures were presented in Chapter 3 (Table 3); a revised version that includes precancer is presented here as Table 16. Note that Table 16 does not discriminate between positive test results that are suspicious for precancer and invasive cancer. Today's cancer screening tests, with the exception of cervical cytology, do not have that level of discriminatory ability. It is questionable whether they should, as cancer screening is not intended to provide that degree of information about the nature of suspicious abnormalities.

Table 16: The building blocks of performance measures for cancer screening tests that detect precancer and invasive cancer

		Truth			
		Invasive cancer present (Phase B)	Precancer present (Phase B)	Neither present (Phase A or no cancer)	Total
Screening test result	Positive	a_i true invasive positives	a_p true precancer positives	b false positives	a_i+a_p+b
	Negative	c_i false invasive negatives	c_p false precancer negative	d true negatives	c_i+c_p+d
	Total	a_i+c_i	a_p+c_p	b+d	$a_i+a_p+b+c_i+c_p+d$

Performance measures for cancer screening tests that detect both precancer and invasive cancer can be calculated by combining the two if measuring the complete impact and performance of the cancer screening test is desired. Calculations would be the same as in Chapter 3, with a equaling $a_i + a_p$, and c equaling $c_i + c_p$. Cells b and d do not change in this instance. The interpretations do not change, although to be as precise as possible it should be said, for example, that sensitivity is the percent of individuals with precancer or invasive cancer who received a positive test, and that specificity is the percent of individuals with neither precancer nor invasive cancer who received a negative test.

C_p is somewhat of a theoretical quantity, as it is impossible to know whether a symptom-detected invasive cancer that is classified as a false negative was, at the time of the screen, a precancer or an invasive cancer. It is unlikely for a precancer to be detected due to symptoms, but should that occur, it seems fair to count that cancer towards c_p.

There are no hard and fast rules for calculating performance measures for precancer alone or invasive cancer alone when a cancer screening test detects both, though a compelling argument can be made for calculating sensitivity simply as $a_p/(a_p+c_p)$ in the instance of precancer and $a_i/(a_i+c_i)$ in the instance of invasive disease. For the other performance measures, the calculations will depend on how the outcome that is not of interest is classified and whether it is even included. If we wish, for example, to calculate performance measures for invasive disease, we have two options: precancer diagnoses could be excluded entirely from calculations, or screens that are associated with precancer diagnoses can be counted as false positives. Cells b and d are affected, which means that any performance measure that utilizes them will be different for the two methods. Both options return results of similar magnitude if precancer and invasive cancer are rare.

Chapter 4: Population measures: definitions

The manner in which intermediate and definitive outcomes are calculated does not change. Incidence and case survival can be calculated for precancer and invasive cancer alone or combined. A category for precancer can be added to stage distributions. Mortality calculations will not change as they do not utilize diagnoses.

Chapter 5: Population measures: cancer screening's impact

Recall from Chapter 5 that cancer screening that detects only invasive cancer will lead to an increase in invasive cancer incidence. Cancer screening that detects only precancer will lead to an increase in precancer. It also will lead to a decrease in invasive cancer incidence as long as not all precancer detected through cancer screening represents overdiagnosis. If a cancer screening test can detect both precancer and invasive cancer, the impact on invasive cancer incidence is difficult to predict. It will depend on many factors, including the ratio of precancer to invasive cancer detected through cancer screening, as well as the frequency of interval cancers and their stage (precancer or invasive).

The other measures discussed in Chapter 5 will be affected as well, though none will "flip-flop" like cause-specific incidence. Consider, for example, case survival. Detection of invasive cancer inflates case survival, and detection of precancer inflates case survival to even a greater degree, because precancer occurs earlier in the natural history of cancer.

A reduction in invasive cancer incidence is accepted as a definitive outcome in the case of cervical cancer screening and colorectal cancer screening with colonoscopy. Far more cervical precancer is detected than invasive cervical cancer. Years of wide-spread cervical cancer screening combined with unique aspects of cervical cancer natural history have led to extremely low incidence rates of invasive cervical cancer in much of the US. Screening with colonoscopy has led to a meaningful reduction in the number of invasive colorectal cancers, though its impact has yet to match that of cervical cancer screening.

If cancer screening is of benefit, a reduction in invasive cancer incidence should be followed by a reduction in cause-specific mortality. If the former happens but not the latter, it is likely that detection at a precancerous stage offers no prognostic benefit compared with detection at an early invasive stage. Further discussion of benefit in the absence of a cause-specific mortality reduction can be found in Chapter 9.

The use of cancer incidence as a definitive outcome assumes that the benefit-to-harm ratio is similar or better for screen detection of precancer relative to invasive cancer. That may not be the case: precancer, at the time of detection, is not life-threatening as it cannot metastasize. Unfortunately, population-based trends in detection of precancer either are not available or are based on incomplete ascertainment of the precancer that cancer screening can detect. That limits our ability to assess the entire impact of cancer screening, a serious issue given that detection of precancer through cancer screening is becoming a relatively common occurrence.

Chapter 6: Experimental research designs

All study designs described in Chapter 6 can be employed to investigate cancer screening's ability to reduce invasive cancer.

Chapter 7: Observational research designs

Case-control studies, the most complex of the study designs presented in Chapter 7, need some modifications when detection of invasive disease is the outcome of interest (3).

A case-control study to assess the ability of a cancer screening test to reduce invasive cancer utilizes cases, individuals who have been diagnosed with invasive cancer, and matched controls. Controls must be alive at the time of the case's diagnosis and must not have been diagnosed with invasive cancer during the case's exposure window, which is the time during which the case's invasive cancer could have been detected through cancer screening as precancer. The exposure window must not include the time that the case's cancer could have been screen-detected as invasive cancer. Cancer screening activity for both cases and controls is assessed for the exposure window.

Data elements that provide information on death usually are not needed for studies of cancer prevention screening, as death occurs after the definitive outcome of diagnosis.

Example of a case-control study of cancer screening with an outcome of invasive disease

Newcomb et al examined the ability of screening sigmoidoscopy to reduce colorectal cancer incidence (4). Cases and controls resided in one of three counties in Washington State. Cases were identified using the SEER Puget Sound cancer registry, were between ages 20 and 74, and newly diagnosed with invasive colorectal adenocarcinoma. Controls were randomly selected according to the age and sex distribution of the cases (frequency-matching) using Washington State driver's license data (ages 20–64 years) and Medicare files (65 years and older). The exposure window included only those tests performed more than 1 year prior to diagnosis date (cases) or more than 1 year prior to interview date (controls). Information on cancer screening history was collected using structured telephone interviews. The authors present their findings separately for proximal and distal colorectal cancer to reflect the anatomy of the colorectum and the inability of the sigmoidoscope to reach the proximal colon.

References

1. Bretthauer M, Kalager M. Principles, effectiveness and caveats in screening for cancer. Br J Surg. 2013 Jan;100(1):55–65. PubMed PMID: 23212620.
2. U.S. National Cancer Institute. NCI dictionary of cancer terms [Internet]. Bethesda (MD): U.S. National Cancer Institute [cited 2019 Oct 23]; [about 25 screens]. Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms.
3. Weiss NS. Case-control studies of the efficacy of screening tests designed to prevent the incidence of cancer. Am J Epidemiol. 1999 Jan 1;149(1):1–4. PubMed PMID: 9883787.
4. Newcomb PA, Storer BE, Morimoto LM, Templeton A, Potter JD. Long-term efficacy of sigmoidoscopy in the reduction of colorectal cancer incidence. J Natl Cancer Inst. 2003 Apr 16;95(8):622–5. PubMed PMID: 12697855.

Chapter 9. Additional considerations

The topics in this chapter are relevant to assessment of cancer screening data but did not have an obvious home in the earlier chapters of this primer. As you will see, they are quite varied in scope. Each falls in one of three categories: data interpretation, methodology, and policy.

Topics regarding data interpretation

Number needed to screen

Number needed to screen, or NNS, indicates how many individuals need to be screened so that one fewer individual dies of the cancer of interest. NNS is only relevant if cancer screening reduces mortality. NNS estimates for cancer screening tests tend to be in the hundreds to thousands of individuals. For example, the NNS for lung cancer screening with low dose computed tomography (LDCT) calculated from the National Lung Screening Trial (NLST) data was 320 (1).

The first step in calculating NNS is to subtract the cause-specific mortality rate in the presence of cancer screening from the cause-specific mortality rate in the absence of cancer screening. That quantity, which is a rate, is called the absolute risk reduction, and is an indication of extent of death prevented by cancer screening. NNS equals the reciprocal of the absolute risk reduction. A fictional example is presented in Table 17. The absolute risk reduction in that table is 20 per 1,000 person-years. The NNS is 1000/20, or 50.

NNS is calculated assuming that the only factor that contributes to the difference in mortality is cancer screening. It is best to use data from randomized controlled trials (RCTs), as data that come from other sources could reflect confounders of the screening/cause-specific mortality relationship.

Table 17: Calculating number needed to screen (NNS)

	Person-years (PY)	Number who die of the cancer of interest	Cause-specific mortality rate
Screened	10,000	100	10 per 1,000 PY
Unscreened	15,000	450	30 per 1,000 PY
NNS calculations: 30 per 1,000 PY - 10 per 1,000 PY = 20 per 1,000 PY; 1000/20=50 NNS is 50			

Data are fictional

Generalizability of results

Generalizability refers to the applicability of results from a study, experimental or observational, to groups other than the study participants. Issues of generalizability are what drive the need to assess effectiveness. A cancer screening test may be efficacious in an RCT, but its ability to be effective in a community setting is not guaranteed by that finding.

Most cancer screening guidelines are based on findings of RCTs. Because cancer screening RCTs are long, large, and expensive undertakings, few are done. Not surprisingly, the urge to take the results of an RCT conducted in one population and apply them to another population is strong. The populations at hand could be dissimilar regions of one country, two countries in the same part of the world with different health care systems, or two countries far away from one another with dramatically different cultural norms.

It should not be assumed that a beneficial effect of cancer screening seen in one population will be replicated in another population if the two populations have different risk factor profiles. An example is lung cancer screening: the cancer screening process may not confer the same magnitude of benefit in asbestos workers, say, as it does in cigarette smokers. It is not wise to extrapolate results from one population to another if the two populations have different clinical practices, clinical resources, and access to health care. Low and middle

income countries have begun to establish cancer screening programs based on experience in high income countries, yet differences in medical resources, access to transportation, and rurality may not allow easy, frequent, or productive visits to cancer screening or treatment centers. Cultural norms also may impact cancer screening uptake and cancer treatment choices.

The assumption that a null effect of cancer screening is generalizable from one population to another also can be unwise. A region with a preponderance of late-stage, untreatable cancers may benefit from cancer screening, whereas the same cancer screening practice may have little to no impact in a region where most patients have earlier stage disease for which treatment is available.

Studies done in regions assumed to be similar enough to produce comparable findings can and have produced conflicting results. The phenomenon has been observed in breast cancer screening, but the best example comes from prostate cancer screening. There are two notable RCTs of prostate cancer screening: the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) (2) and The European Randomized Study of Screening for Prostate Cancer (ERSPC) (3). PLCO, an RCT done in the US, found no reduction in prostate cancer mortality, while ERSPC, an RCT done in many countries in Europe, did. The two studies employed different cancer screening protocols, which may explain, at least in part, the discordant findings. Nevertheless, discussions regarding the conflicting results have focused on contamination in PLCO's control arm and likely inferior prostate cancer treatment in ERSPC's control arm. Random variation or a systematic difference (that is, the contamination and treatment issues) may very well be responsible, but it also is necessary to consider the possibility that prostate cancer screening may be of benefit in one region but not the other.

Concurrent changes in treatment

Cancer screening does not operate in a vacuum. While cancer screening tests are under investigation, disseminating, or their use reaches a steady state, changes in clinical practice are occurring as well. Advances have led to a better understanding of tumor composition, which in turn have led to new and highly effective therapies for some tumors. Cures are possible today that were not possible 20 years ago. This situation begs this question: if cancer treatment has improved, especially at regional and distant stages, is screen detection at an early stage still necessary?

In the presence of concurrent changes in treatment, an RCT can still evaluate whether cancer screening is of benefit as long as individuals in both arms have access to the same treatments. Concurrent changes do present a problem in time trend studies; it is impossible to know whether reductions in cancer mortality are due to uptake of a new cancer screening regimen or availability of a new treatment.

An RCT to determine whether a cancer screening test affects a benefit cannot be established each time a shift in clinical practice occurs. Creative use of available data can shed some light, however. The ecologic study of Autier et al (4), mentioned in Chapter 7, examined the issue of concurrent changes in breast cancer screening uptake and treatment by examining time trends for three pairs of regions in Europe. Each region in a pair had similar access to breast cancer treatment yet a different date of widespread mammography adoption. While not without limitations, that analysis suggests that recent reductions in breast cancer mortality are not overwhelmingly due to cancer screening.

Topics regarding methodology

Microsimulation modeling

Microsimulation modeling of cancer screening is a technique in which computer-generated (fictional) life histories are manipulated by applying assumptions about factors that affect cancer screening outcomes. Models produce outcomes, such as cause-specific mortality, for a variety of assumptions and cancer screening scenarios,

providing insight into benefits and harms of cancer screening. The National Cancer Institute's (NCI's) Cancer Intervention and Surveillance Modeling Network (CISNET) initiative has taken the lead in microsimulation modeling for cancer screening (5).

Microsimulation modeling is possible given unprecedented improvements in computational power in recent years. The use of microsimulation modeling in lieu of establishing RCTs has been suggested, because RCTs cannot address every proposed cancer screening strategy. Microsimulation modeling is arguably most valuable when done in conjunction with data from population-level databases, completed RCTs, or large, well-conducted prospective cohort studies, as certain assumptions needed to generate life histories can be based on real-life experience.

No microsimulation model will perfectly replicate reality. However, these models have become a popular and useful tool to investigate "what if" situations. Results from CISNET models, in conjunction with RCT and cohort data, are now used by the United States Preventive Services Task Force (6) when developing cancer screening recommendations .

Magnitude of overdiagnosis

The excess incidence method was presented in Chapter 6 as a way to calculate the degree of overdiagnosis in an RCT, but it is not the only method available. Some methods employ assumptions about the distribution of lead time (7), while others compare changes in incidence that have occurred over time, generally in conjunction with other factors (8,9). Statistical modeling, including microsimulation modeling, has been utilized in the effort to determine the magnitude of overdiagnosis or a range of plausible magnitudes.

There has been heated discussion as to which method will produce the correct answer. That assumes, of course, that there is one correct answer. But overdiagnosis only exists in the context of cancer screening, and therefore, the magnitude of overdiagnosis is a function of aspects of the cancer screening regimen, including test, screening interval, compliance, and those who are screened. Magnitude also is a function of the intensity of diagnostic evaluation that follows a positive test. There is no one correct answer; there are many correct answers, with each dependent on many factors.

The desire to quantify the magnitude of overdiagnosis is related to the desire to weigh the benefits and harms of cancer screening, something that is most easily done when a single number can be attached to each. In lieu of a single number, a range of plausible measures of overdiagnosis can be used in sensitivity analyses.

Incidence and prevalence screens

When discussing burden of disease, the terms prevalence and incidence refer to disease that is existing and new, respectively. The terms prevalence and incidence are sometimes used in cancer screening to describe the initial and later screens, respectively, performed as part of a cancer screening program or an RCT. The initial screen is expected to lead primarily to detection of cancers that have stalled in Phase B, while incidence screens are expected to lead primarily to detection of cancers that have moved into Phase B since the last cancer screening test. All other things being equal, the yield on prevalence screens is expected to be higher than the yield on incidence screens. Also, the prognosis for cancers detected on the prevalence screen is expected to be more favorable than for those detected on incidence screens.

Interval cancers

Interval cancers often are considered failings of cancer screening, even though cancer screening is not designed or expected to lead to detection of every Phase B cancer. Some conditions that lead to interval cancers, for example, errors in test interpretation and missed screens, may be addressable, but it is unrealistic to believe that interval cancers can be eliminated. Interval cancers are a reminder of the limits of cancer screening.

Cancer can be detected serendipitously, meaning that an unrelated diagnostic medical test or procedure inadvertently finds an abnormality that is suspicious for cancer. An MRI performed to investigate back pain could identify a colonic mass, for example. Whether serendipitously detected cancers are interval cancers is open to debate. They do not arise from symptoms but they may have been missed on the previous organ-specific cancer screening test.

Topics regarding policy

Selecting a cancer screening interval

The phrase cancer screening interval refers to the time between screens. Though the choice of the screening interval should be based exclusively on the average length of Phase B and how variable it can be, historically, is has not. It is only recently that screening intervals have started to reflect the natural history of cancer. In the past, screening intervals were typically one year, probably because cancer screening was associated with the practice of having an annual physical.

The choice of screening interval will impact effectiveness and the magnitude of harms. It also will drive costs and availability of health care resources. Ideally, these factors are weighed in conjunction with knowledge of the natural history of cancer to arrive at a screening interval that affords benefit but does not strain a health care system.

De-implementation

De-implementation refers to the reduction or cessation of a service provided by health care practitioners. Calls for de-implementation may be made when practices do not benefit patients, including when they are harmful or wasteful. The need for de-implementation may arise in the instance of adoption of a practice whose benefit is uncertain, or if a practice observed to be efficacious is not effective. A well-known instance of de-implementation is the reduction in prescribing of postmenopausal hormone therapy after users experienced an increase in breast cancer risk (10).

De-implementation has been discussed in the context of cancer screening for a number of reasons. Some cancer screening tests have become widely adopted in clinical practice without strong or direct evidence that their use reduces cause-specific mortality; some also have been adopted without complete understanding of the harms they cause. A notable example of the former is thyroid cancer screening. Low-cost ultrasound thyroid cancer screening became available in South Korea in the 1990's even though the practice had never been evaluated in an RCT. Thyroid cancer incidence increased 15-fold from 1993 to 2011, although no change in thyroid cancer mortality occurred concurrently. In 2015, the Korean Committee for National Cancer Screening Guidelines issued a recommendation against thyroid cancer screening with ultrasonography for healthy individuals (11,12).

De-implementation will result in reversal of the effects on intermediate outcomes described in Chapter 5. Incidence of invasive cancer (in the case of cancer screening that detects only invasive disease) and case survival will decrease, and assuming all else remains the same, should approach their pre-screening levels. The number of early stage cancers should decrease due to elimination of overdiagnosis. The number of late stage cancers will not change if cancer screening did not result in down staging, and will increase if it did.

As it is for implementation, it is critical to track the changes in both intermediate and definitive outcomes during a period of cancer screening de-implementation. Both implementation and de-implementation are by necessity based on certain assumptions; therefore, the impact cannot be predicted. It is particularly important to watch for unexpected consequences, be they favorable or deleterious.

Reduction in advanced-stage cancer

A reduction in advanced-stage cancer, usually distant cancer, has been suggested as a surrogate for cause-specific mortality. The push to use advanced-stage cancer has to do, at least in part, with the desire to obtain answers regarding the impact of cancer screening without having to wait for a cause-specific mortality outcome. A reduction in the number of distant-stage cancers may be the best of the intermediate cancer screening outcomes in terms of correlation with reductions in cause-specific mortality, but it still does not reflect experience after diagnosis and does not measure how cancer screening alters length of life.

Legitimate use of a reduction in distant-stage cancers as what is, in effect, a definitive endpoint requires that those cancers are fatal, and often they are. It also assumes that non-distant-stage cancers have a better prognosis, which in most situations they do. Yet consider a cancer that, in the absence of cancer screening, would be diagnosed at a distant stage, but in the presence of cancer screening, is diagnosed at a regional stage. If the prognosis for regional stage cancer is the same as that of distant-stage cancer, no reduction in cause-specific mortality would occur even though the number of distant-stage cancers has decreased.

If the day comes when cancer is no longer fatal even at a distant stage, the goals of cancer screening will need to be reassessed. In the meantime, the choice of distant-stage disease as a definitive endpoint must be made carefully and on a situation-by-situation basis.

Benefit in the absence of a mortality reduction

Once upon a time there was no cancer screening in the US. When discussions regarding establishment of population-based cancer screening began in earnest, the proposed metric of benefit was a reduction in cause-specific mortality, as cancer was considered to be a life-threatening disease. Diagnoses often occurred at late stages and few, if any, effective treatments were available once cancer spread beyond the organ of origin.

The first breast and colorectal cancer screening tests to become established in the US were shown to reduce cause-specific mortality in at least one RCT. Those tests, film-screen mammography and guaiac-based fecal occult blood testing, have since been replaced with tests that are more technologically advanced: digital mammography and breast tomosynthesis, and fecal immunochemical testing, flexible sigmoidoscopy, and colonoscopy. Yet none of the replacement tests was vetted in a study that assessed cause-specific mortality prior to adoption.

When replacement tests are adopted, it is done so under the assumption that the new test will confer the same or a greater reduction in cause-specific mortality as the test it is replacing. The replacement tests also have a characteristic that make them more desirable than the test they are replacing. They may have better performance measures, such as lower false positive rates, or they may be more acceptable to patients. They could be less expensive when all components of the screening process are considered.

In my opinion, future cancer screening tests that target an organ for which no efficacious screening test exists only should be implemented in clinical practice when high-level evidence is available to support a reduction in cause-specific mortality. Others may feel differently. Some have argued that a shift to a stage at diagnosis that is simpler to treat is benefit enough, though the consequences that come with a cancer diagnosis earlier in time must not be ignored. Those include intense surveillance regimens, chemoprevention strategies, and psychological challenges for periods of time that are longer than those that would have occurred if cancer had been diagnosed later.

Whether it is appropriate to adopt replacement tests in clinical practice without formal vetting using a cause-specific mortality endpoint or another measure of the benefit to harm is a matter of the cancer at hand and differences in the replacement and original test. There are some instances in which a strong argument can and have been made for adoption without full knowledge about the impact on benefits and harms. Data are available

to retrospectively support some of the decisions made regarding replacement, including the choice to adopt colonoscopy screening for colorectal cancer.

References

1. The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011 Aug 4;365(5):395–409. PubMed PMID: 21714641.
2. Andriole GL, Crawford ED, Grubb RL 3rd, Buys SS, Chia D, Church TR, Fouad MN, Isaacs C, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hsing AW, Izmirlian G, Pinsky PF, Kramer BS, Miller AB, Gohagan JK, Prorok PC; PLCO Project Team. Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. J Natl Cancer Inst. 2012 Jan 18;104(2):125–32. PubMed PMID: 22228146.
3. Hugosson J, Roobol MJ, Månsson M, Tammela TLJ, Zappa M, Nelen V, Kwiatkowski M, Lujan M, Carlsson SV, Talala KM, Lilja H, Denis LJ, Recker F, Paez A, Puliti D, Villers A, Rebillard X, Kilpeläinen TP, Stenman UH, Godtman RA, Stinesen Kollberg K, Moss SM, Kujala P, Taari K, Huber A, van der Kwast T, Heijnsdijk EA, Bangma C, De Koning HJ, Schröder FH, Auvinen A; ERSPC investigators. A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer. Eur Urol. 2019;76(1):43–51. PubMed PMID: 30824296.
4. Autier P, Boniol M, Gavin A, Vatten LJ. Breast cancer mortality in neighbouring European countries with different levels of screening but similar access to treatment: trend analysis of WHO mortality database. BMJ. 2011 Jul 28;343:d4411. PubMed PMID: 21798968.
5. Cancer Intervention and Surveillance Modeling Network. Overview of CISNET Modeling [Internet]. Bethesda, MD: National Cancer Institute [cited 2019 Oct 24]; [about 4 screens] Available from: https://resources.cisnet.cancer.gov/registry/learn/.
6. Modeling Report: Lung Cancer: Screening. [updated May 2019; cited 2019 Oct 24]. Available from: https://www.uspreventiveservicestaskforce.org/Page/Document/modeling-report/lung-cancer-screening.
7. Ripping TM, Ten Haaf K, Verbeek ALM, van Ravesteyn NT, Broeders MJM. Quantifying Overdiagnosis in Cancer Screening: A Systematic Review to Evaluate the Methodology. J Natl Cancer Inst. 2017 Oct 1;109(10):djx060. PubMed PMID: 29117353.
8. Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. BMJ. 2015 Jan 7;350:g7773. PubMed PMID: 25569206.
9. Welch HG, Prorok PC, O'Malley AJ, Kramer BS. Breast-Cancer Tumor Size, Overdiagnosis, and Mammography Screening Effectiveness. N Engl J Med. 2016 Oct 13;375(15):1438–47. PubMed PMID: 27732805.
10. Steinkellner AR, Denison SE, Eldridge SL, Lenzi LL, Chen W, Bowlin SJ. A decade of postmenopausal hormone therapy prescribing in the United States: long-term effects of the Women's Health Initiative. Menopause. 2012 Jun;19(6):616–21. PubMed PMID: 22648302.
11. Ahn HS, Kim HJ, Kim KH, Lee YS, Han SJ, Kim Y, Ko MJ, Brito JP. Thyroid cancer screening in South Korea increases detection of papillary cancers with no impact on other subtypes or thyroid cancer mortality. Thyroid. 2016 Nov;26(11):1535–40. PubMed PMID: 27627550.
12. Ahn HS, Welch HG. South Korea's Thyroid-Cancer "Epidemic"--Turning the Tide. N Engl J Med. 2015 Dec 10;373(24):2389–90. PubMed PMID: 26650173.

Chapter 10. Closing thoughts

SEER data indicate that cancer incidence has risen slightly (10%), 5-year cancer survival has risen substantially (40%), and cancer mortality has droped modestly (25%) since 1975 (1). An increase in incidence and 5-year case survival are difficult to interpret for reasons discussed in this primer. But a reduction in cancer mortality of any magnitude is a success.

President Nixon spoke of the conquest of cancer when he proposed the National Cancer Act in 1971(2), and others have used similar war-like language over the years. Data from SEER, however, suggest that we have not conquered cancer. We know much more about cancer today than we did in 1971, but we still do not seem to know enough to make a huge impact. Cancer "fads" have come and gone; some have and some haven't made a lasting difference. Chemoprevention has reduced the risk of breast cancer recurrence, and prevention of smoking initiation and smoking cessation have led to meaningful decreases in lung cancer incidence and mortality. On the other hand, autologous bone marrow transplant for breast cancer was used for a number of years to no avail. Our understanding of the relationship of diet and cancer is still poor. It was estimated that in 2018, 600,000 Americans would die from cancer. Though a small percentage of the population, the absolute number is large.

Where does cancer screening fit into the picture? It depends on who you ask. Most researchers believe that earlier detection due to cancer screening has led to some reduction in cancer mortality, though there is widespread disagreement as to the degree of its impact. There is general agreement, however, that cancer screening programs impact health care spending and availability of resources yet benefit only a few of those who are screened. There is less agreement, however, regarding what constitutes benefit and harm, and even less regarding the acceptable ratio of harm to benefit. Discussion of the complexities of cancer screening began to appear in some lay press publications about 20 years ago, yet the predominant feeling among individuals in the general public is that cancer screening is important and worthwhile, and that cancer detection at the earliest stage can only lead to good.

The forces that drive the availability of cancer screening and the choice to be screened are complex. So too are the issues that were covered in this primer. I believe, however, that most people, patients and clinicians alike, are educable, and the complexities of cancer screening need not be out of reach. I hope that this primer has helped you in your quest to understand.

References

1. Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD, /, based on November 2018 SEER data submission, posted to the SEER web site, April 2019. Available from: https://seer.cancer.gov/csr/1975_2016.
2. United States House of Representatives. Office of the Law Revision Council, United States Code. National Cancer Act of 1971 (Pub. L. 92-218, Dec. 23, 1971, 85 Stat. 778). Cited 2019 October 29. Available from: http://uscode.house.gov/statutes/pl/92/218.pdf.