



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, the 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.

License

This is an open-access report distributed under the terms of the Creative Commons Public Domain License. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.