Statistical Methods in Medical Research 2004; **13**: 491–506

Planning of randomized early detection trials

Ping Hu Biometry Research Group, National Cancer Institute Bethesda, MD 20892, USA
and **Marvin Zelen** Department of Biostatistics, Harvard School of Public Health,
Dana Farber Cancer Institute, Boston, MA 02115, USA

Consider a randomized clinical trial to evaluate the benefit of screening an asymptomatic population. Suppose that the subjects are randomized into a usual care and a study group. The study group receives one or more periodic early detection examinations aimed at diagnosing disease early, when there are no signs or symptoms. Early detection clinical trials differ from therapeutic trials in that power is affected by: i) the number of exams, ii) the time between exams and iii) the ages at which exams will be given. These design options do not exist in therapeutic trials. Furthermore; long-term follow-up may result in a reduction of power. In general, power increases with number of examinations, and the optimal follow-up time is dependent on the spacing between examinations. Clinical trials in which the usual care group receives benefit are also discussed. Two designs are discussed, for example the 'up-front design' in which all subjects receive an initial exam and then are randomized to the usual care and study groups and the 'close-out design' in which the usual care group receives an exam which is timed to be given at the same time as the last exam in the study group. Both families of designs significantly reduce the power. Power calculations are made for two clinical trials, which actually used these two designs.

1 Introduction

Currently, the most realistic strategy for increasing cure rates and/or lengthening survival for most cancers is to diagnose the disease while it is in an early stage. This is true not only for many cancers, but for many chronic diseases as well (e.g., heart disease, tuberculosis, glaucoma, diabetes, etc). The development of new technologies for disease detection will increasingly make screening for disease a routine part of secondary prevention.

The possibility that early detection of disease may not necessarily result in benefit has motivated the need for well conducted clinical trials which can generate data for evaluating benefit. The most notable example of such trials is breast cancer in which eight randomized trials have been carried out to evaluate the benefits of a mammogram possibly combined with a clinical exam.[1–12] Despite these clinical trials, there is a controversy about the benefits for women under the age of 50. An NIH consensus conference concluded that there was not significant evidence to settle the issue. There is a general agreement that women over 50 years do benefit which is reflected by lower mortality. However, this conclusion is not completely accepted.[13]

Among the problems with these clinical trials is that they have been planned using the ideas of therapeutic trials. The planning has not taken into account the special features

Address for correspondence: Ping Hu, Biometry Research Group, National Cancer Institute, Executive Plaza North, Suite 3131, 6130 Executive Blvd, MSC 7354, Bethesda, MD 20892-7354, USA. E-mail: ph107y@nih.gov

10.1191/0962280204sm379ra

of early detection trials which do not arise in therapeutic trials. To our knowledge, as reflected in the published literature, there was no investigation of the role of the number of examinations and the spacing between examinations and its effect on the statistical power of these studies. As a result, these trials have been planned sub-optimally leading to much controversy about their conclusions.

A typical early detection randomized trial will have two groups, which will be designated as a control and a study group. Individuals in the control group have their usual medical care, whereas individuals in the study group will have one or more special examinations which have the potential to diagnose disease while it is asymptomatic and the individual has no signs or symptoms of the disease. The special features of these early detection trials, which must be considered in the planning stage, are the number of exams in the study group, the spacing of exams and the optimal follow-up time for analysis. One contrast between therapeutic and early detection trials is that long-term follow-up in early detection trials may lead to reduced power. This feature arises because cases diagnosed in the study group, after the last scheduled examination, may either be those which were evaluated as false negatives or cases which did not have disease at the last examination. It is not possible to distinguish between them at the time of diagnosis. As a result, all deaths arising from these cases are included in the mortality comparisons in order to have an unbiased comparison. A relatively short follow-up time may not have enough deaths to have adequate power. A long follow-up time leads to an increase in the number of diagnosed cases not present at the last scheduled examination. This dilution will result in reduced power. This dilution and the resulting reduction in power are not true for therapeutic trials having a long follow-up time. Another consideration in planning early detection trials is the choice of the target population. In therapeutic trials all eligible subjects have disease, whereas in early detection trials, subjects do not have disease. However, subjects may be incident with disease at a later time. Thus, the choice of the target population for the early detection trial is important.[14] A low risk population may require a very large number of subjects to achieve adequate power.

At the present time this general topic is important for all chronic diseases in which there appears to be benefit from earlier diagnosis of disease when combined with therapy. Two major topics are discussed in this article: i) the basic model for planning clinical trials and ii) experimental design issues for the early detection of disease: novel designs.

2 The basic model for planning clinical trials

2.1 Basic model

This section describes the basic model for planning clinical trials for the evaluation of the early detection of disease. The theoretical development is given in Hu and Zelen (1997) and will not be duplicated in this article. The model can be used to investigate: i) the optimal time of analysis and length of follow-up; ii) the optimal spacing between examinations and iii) the number of examinations versus sample size for fixed costs. These features are illustrated by numerical calculations on the basis of the input parameters for planning breast cancer trials.

The starting point is to model the natural history of the disease.[15,16] For an individual, this can be idealized as consisting of three possible states, S_0, S_p, and S_c. In S_0, the individual is free of disease or has disease which cannot be detected by any specific diagnostic examination; S_p refers to the pre-clinical disease state where an individual unknowingly has disease, which can be detected by a special examination; S_c refers to the state where the disease is clinically diagnosed by usual care. It will be assumed that the disease is progressive; that is the transitions are $S_0 \rightarrow S_p \rightarrow S_c$.

Assume that at time t_0 individuals are registrated and randomized into two groups–one is a control group and the other is a study group. The study group is offered a number of special examinations, for the early detection of a specific disease at times $t_1 < t_2 < \cdots < t_n$. The ith screening sub-interval is denoted by $[t_{i-1}, t_i)$ for $i = 1, \ldots, n$. The individuals in the pre-clinical disease stage S_p will be identified by generation. An individual entering S_p during the ith interval is identified as an ith generation individual. The argument T considered in this article refers to the chronological time of follow-up; that is the total follow-up time is $(T - t_0)$.

The probability model calculates the cumulative probability of the death of an individual (both for the control and study groups) at chronological time T, which is equivalent to a follow-up time of $(T - t_0)$. The features of the probability model are: i) sensitivity of the early detection examination, ii) the sojourn time distribution in the pre-clinical state, iii) the transition probability of entering S_p from S_0 as a function of time and iv) the survival distributions of the control and the early detection groups. The latter survival distribution is made up of a mixture composed of cases diagnosed by a screening exam and interval cases. We allow for the probability that survival may be different for those in the pre-clinical state at study entry.

In formulating the model, it is necessary to take into account the earlier time of diagnosis for examination diagnosed cases. This so called lead time is the difference between the time the disease would have been diagnosed in the clinical state and when it was actually diagnosed by an early detection examination. The lead time is a random variable, and by definition, is a guarantee time for minimal survival; that is those diagnosed by an early detection exam will, with certainty, have lived to the time of clinical diagnosis. Failure to account for the guarantee time results in a lead time bias in favor of the study group. Note that there is a possibility that a subject may die before being diagnosed under usual care. In the present formulation, we are ignoring this over diagnosis issue.

2.2 Numerical illustrations

We illustrate the use of the model for planning breast cancer trials with four examples. The basic input parameters for all calculations are chosen from literatures (Walter and Day, 1983; Shapiro *et al.*, 1985; Stomper and Gelman, 1989)[9,17–19] and summarized in Table 1.

Our first illustration investigates the power of a trial comparing: i) two different schedules (two and four exams), ii) varying the time between exams [1(1)5 years], iii) varying the time of the follow-up period (5,10,15, optimal in years) for sample sizes of 25 000 in each group. Figure la and b depict the power calculations for a one-sided level 0.05 test. The numbers shown on the top curves are the required follow-up years in order to reach the maximum power. The fractions in the figures indicate the

Table 1 Summary of parameter for breast cancer used in the examples

Parameter	Value
Median survival (control group), not in S_p at time t_0	10 years
Median survival (control group), in S_p at time t_0	11 years
Median survival (study group) not in S_p at time t_0	17 years
Median survival (study group) in S_p at time t_0	20 years
Sensitivity	0.90
Prevalence	7 per 1000/ year
Transition from S_0 to S_p	2 per 1000/ year

proportion of participants receiving all scheduled exams. The calculations show that the greater the interval between examinations the greater the power, provided the follow-up time is long enough to achieve optimal power. However, larger intervals between

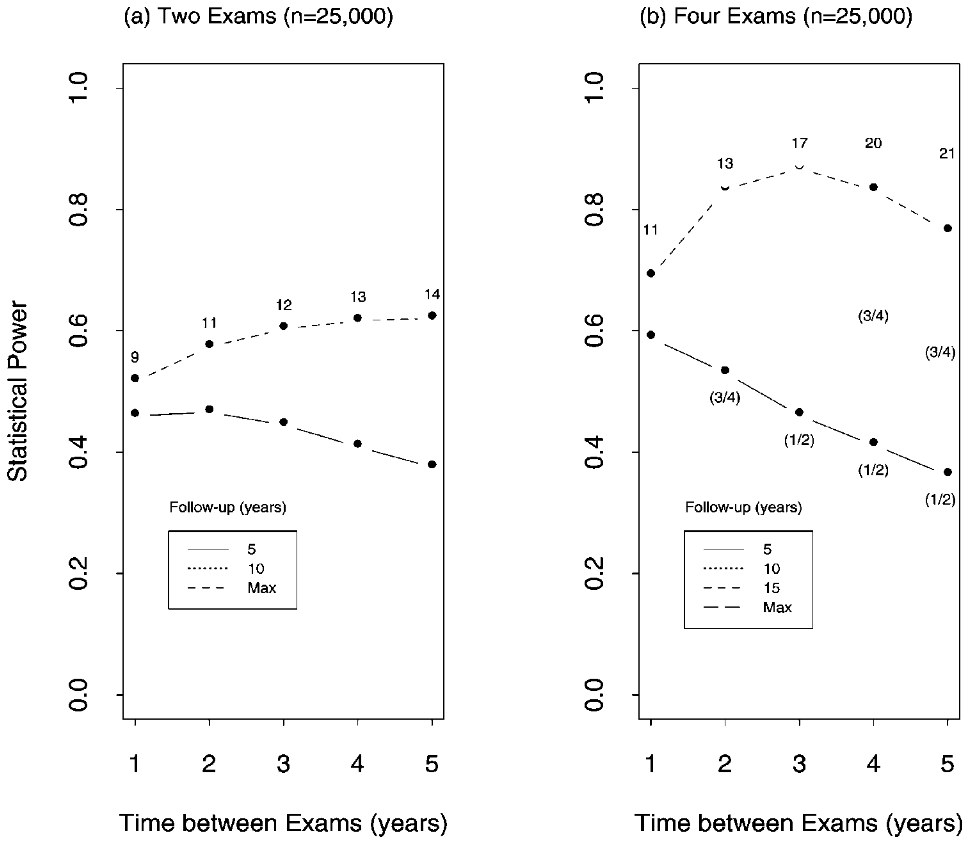

Figure 1 Power calculations for a one-sided level of 0.05 test with $n = 25\,000$ in each group showing relationship with time between exams and follow-up times. (a) and (b) depict results for two and four exams.

examinations require a longer follow-up to reach maximal power. Inadequate follow-up time will result in reduced power with longer intervals between examinations.

The next illustration is summarized in Figure 2a and b. In this calculation, the number of examinations is fixed at three examinations, but the sample sizes vary (25 000, 50 000 and 75 000), the intervals between exams vary [1(1)5] and the time of follow-up is five and 15 years. An inadequate follow-up time may not enable all planned exams to be given. For example with five years of follow-up time, only the first two of the three exams will take place when the time between examinations is three years or more. Note also that the power does not increase dramatically with long-term follow-up if the interval between examinations is small (one year). The power for $n = 25\,000$ and one year spacing is 0.54 for five year follow-up and only rises to 0.58 for a 15 year follow-up study. However, the power rises dramatically with a 15 follow-up, when the time between exams become larger.

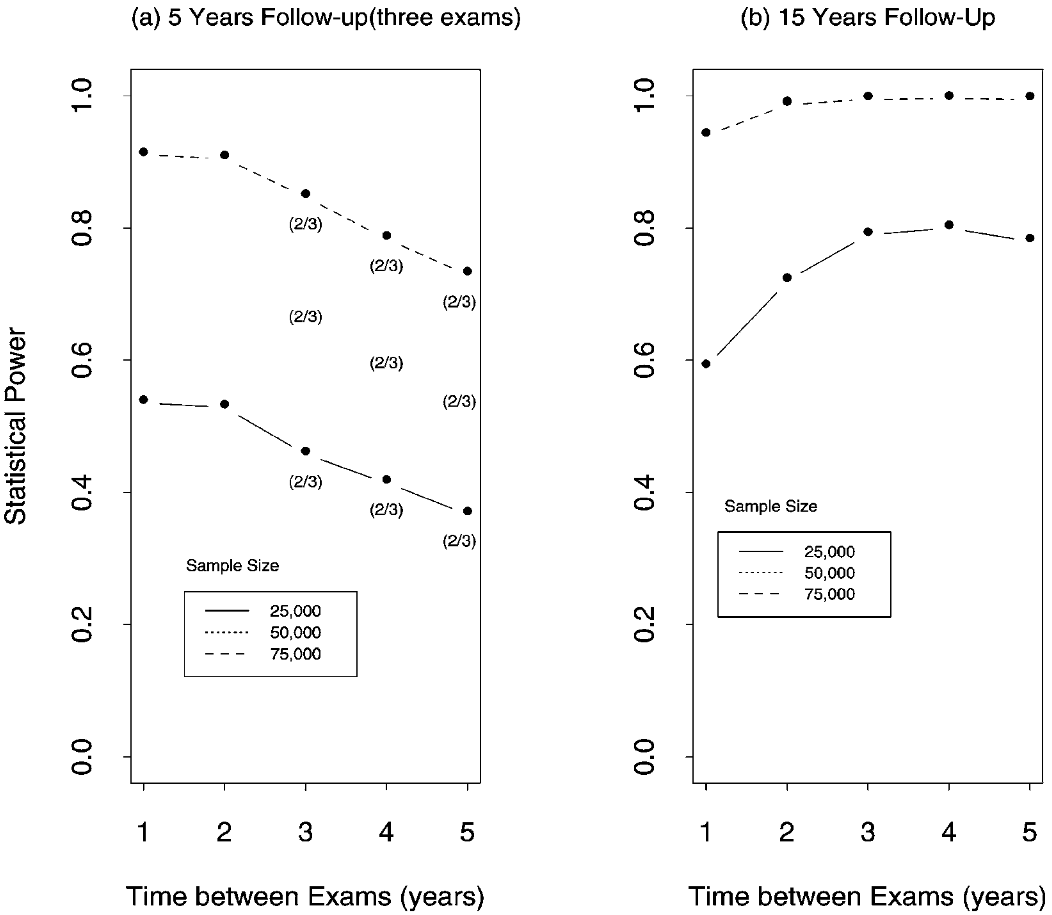

Figure 2 Power calculations for a one-sided level of 0.05 test with three exams showing relationship between sample size and time between exams. (a) and (b) give results for five and 15 years of follow-up.

Table 2 Combination of factors resulting in power of 80% and 95% for one-sided 0.05 level of significance (equal sample sizes in the study and control groups)

Sample size	No. of exams	Power of 80%		Power of 95%	
		Time between exams (years)	Optimal follow-up time (years)[a]	Time between exams (years)	Optimal follow-up time (years)[a]
25 000	4	2	10	–	–
50 000	2	2	7–8	–	–
	3	1	5–6	3–4	12–13
	4	0.5[b]	6	2	8–9
75 000	2	0.5	6	3–6	10–13
	3	0.5[b]	4	1	7
	4	0.5[b]	4	0.5	8

[a] Follow-up time is dated from the end of accrual.
[b] Power > 80%.

Our third illustration is concerned with planning trials to achieve a fixed power of 0.80 and 0.95. The results are summarized in Table 2. There is a trade-off between number of examinations, time between examinations and optimal follow-up times. For example 80% power can be achieved with a sample size of $n = 25\,000$ with four exams spaced two years apart, provided the follow-up time is 10 years. Increasing the sample size to $n = 50\,000$ will achieve the same power with two exams spaced two years apart or three exams spaced one year apart with follow-up times of 7–8 or 5–6 years, respectively. Larger number of patients results in a reduction of required follow-up time.

Our final illustration in this section illustrates how to increase power by unbalancing the sample sizes for the control and study groups. Suppose the major cost of the trial for the study group is the number of examinations; for example, one exam for 50 000 subjects costs the same as two exams for each of 25 000 subjects. We will consider an experimental design which has 50 000 in the control group. The study group may have 50 000 subjects receiving one exam or 25 000 subjects receiving two exams, etc. Table 3 summarizes the results of this investigation when there are up to four exams in the study

Table 3 Maximum power for constant screening intervals when cost of study group is fixed and independent of the number of exams

Screening interval (years)	No. of exams			
	One ($n = 50\,000$)	Two ($n = 25\,000$)	Three ($n = 16\,666$)	Four ($n = 12\,500$)
0.0	0.58 (8)			
1.0		0.63 (10)	0.66 (10)	0.67 (11)
2.0		0.70 (10)	0.75 (13)	0.79 (13)
3.0		0.72 (13)	0.79 (12)	0.82 (15)
4.0		0.74 (12)	0.79 (16)	0.82 (19)
5.0		0.73 (12)	0.79 (18)	0.80 (21)
6.0		0.72 (14)	0.77 (20)	0.77 (25)

Note: Maximum power is achieved at the follow-up time (Follow-up time is dated from the end of accrual) shown in parenthesis. $n = 50\,000$ in control group.

group. The power was calculated for varying screening intervals. The larger the number of exams, the greater the power. However, the larger the number of exams, the longer the follow-up time to reach maximum power. For example, one exam with 50 000 subjects requires eight years of follow-up to reach a power of 0.58, but four exams, spaced three years apart, having 12 500 subjects in the study group require 15 years of follow-up to reach a power of 0.82.

3 Experimental design issues for the early detection of disease: novel designs

The 'classical' clinical trial design, in which there is a control and a study group, is an ideal way of planning early detection trials. However, because entry into a clinical trial requires subject consent, there may be resistance in consenting to enter a study in which there is the possibility of being assigned to a control group which has no potential benefit to the individual. The purpose of this section is to discuss two experimental designs for early detection trials for which there is potential benefit even if subjects are assigned to the control group. These two experimental designs are called the 'up-front design (UFD)' and the 'close-out design (COD)'. Both designs have been used in breast cancer early detection trials.[1–4,18] Etzioni *et al.*[21] have commented on some features of the analysis of the COD. The UFD essentially results in every patient receiving an initial early detection exam. The COD is a trial where subjects in the control group receive a special early detection exam which coincides with the timing of the last exam in the study group. Both of these designs raise special issues which affect the resulting power of the design and in some instances may lead to bias. The UFD presents an issue of whether the cases diagnosed by the initial exam, given to all participants, should be included in the final analysis. We note that the choice will answer two different scientific questions. The COD presents similar issues regarding which diagnosed cases should be included in the analysis; the options are to include: all cases, only cases diagnosed up to and including the last exam, or when the number of cases in the control and study groups are the same. The different options result in different power and possible biases. The theoretical development for these designs may be found in Ref. (22) and will not be duplicated here.

The UFD is defined by giving each subject an initial exam at time t_0. The subjects may be randomized before or after the initial exam to either a control or a study group. The initial exam may be a different exam than those offered in the study group. There are two options in carrying out the trial and subsequent analyses. One could drop all diagnosed cases from the initial examination and only analyze subjects having a negative up-front exam outcome. Alternatively, all individuals may be included in the analysis regardless of the initial exam outcome. These two options are referred to as the 'drop' and 'keep' options.

If the drop option is used, then the scientific question answered by the study is to evaluate the benefits of screening conditional on the prevalent cases being eliminated (prevalent cases are defined as being in the pre-clinical state at t_0).

There are two ways of implementing the 'keep option' as outlined by Figure 3a and b. Figure 3a represents the keep option if randomization is carried out after the initial

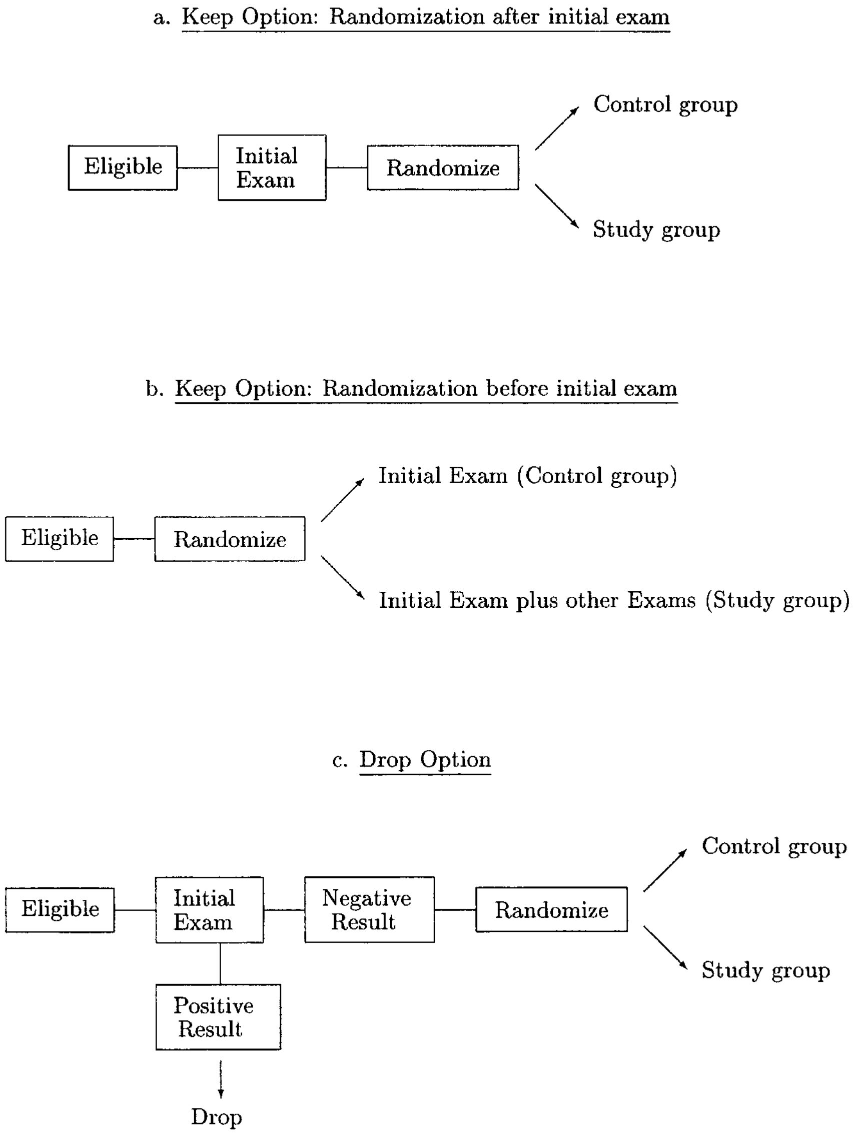

Figure 3 Up-front experimental design: keep and drop options.

exam, whereas, Figure 3b shows the keep option if the randomization is carried out before the initial examination. Note that, if the prevalent cases are dropped in the analysis, the trial is equivalent to Figure 3c, which illustrates the drop option. This equivalence is independent of when the randomization was carried out. In some applications, the outcome of the initial exam may not be known for some period of time, but the randomization is carried out at t_0. Hence, the clinical trial would be carried out according to Figure 3b (randomize all subjects before the initial exam).

A classical experimental design consists of a control and a study group and can generate information on the benefit of the screening exam program; that is exam program versus usual medical care. However, the UFD, with either the drop or keep option, does not directly generate data which can unambiguously show the benefit of early detection exams. If the major benefit of the early detection exam is with prevalent cases, then a conclusion of no difference with the drop option does not necessarily mean that early detection is without benefit. Similarly, if the keep option results in a conclusion of no difference, this also does not necessarily mean that early detection is without benefit. One interpretation is that the greatest benefit may be with the prevalent cases. Additional exams may be without significant benefit.

We will illustrate the effect on power of these two experimental designs using parameter values for breast screening. The parameters used in the up-front illustrations are summarized in Table 4

The first illustration shows the consequences of the UFD compared with a classical design for three exams spaced two years apart. The calculations are summarized in Table 5. The classical design having 24 000 subjects, each in the control and study groups, has a power of 0.80 with the parameters in Table 4. The drop option generally has superior power than the keep option. In order for either option to have the same power (0.80) as the classical design, it is necessary to increase the sample size of the study group to 70 000 for the drop option and 82 000 for the keep option. With each option, the power of the UFD is severely decreased relative to a classical design.

The next illustration explores the role of the number of examinations. The previous example had three exams, two years apart. In this example, we consider the same situation, but with two and four exams. The results are summarized in Table 6. Fewer exams (two) results in lower power and more exams (four) results in higher power. Again, we note the greatly reduced power of the UFD relative to the classical design.

Our next illustration concerns the Canadian National Breast Cancer Study for women aged 40–49, which used an UFD with the keep option.[3,4,20] It is the largest randomized trial in the world to evaluate the benefit of mammography for women 40–49 years of age. This trial randomized subjects to a study group receiving annual mammograms and physical exams whereas a control group received usual medical care. However, all subjects entered into the study received an initial physical examination.

Table 4 Summary of parameter values: up-front exam design

Median survival (control group), not prevalent at time t_0	10 years
Median survival (control group), having disease at time t_0	11 years
Median survival (study group), not prevalent at time t_0	20.0 years
Median survival (study group), having disease at time t_0	22.0 years
Prevalence	6.4 per 1000/year
Transition probability from S_0 to S_p	2.9 per 1000/year
Sensitivity of a single exam	0.90
Time interval between exams (study group)	2 years

Note: (a) Median survival for the control interval cases, prevalence and transition probability are estimated from SEER 1988–1996 data. (b) 25% reduction in mortality at nine years of follow-up from entry is used for the study group (this results in median survival of 20 years).

Table 5 Maximum power with three screening exams spaced two years apart for the study group (classical design versus up-front exam design)

Option	Drop			Keep		
	$N_c = N_s$	Power	Follow-up (yrs)	$N_c = N_s$	Power	Follow-up (yrs)
A	24 000	0.80	11–14	24 000	0.80	11–14
B	24 000	0.42	12–15	24 000	0.38	12–15
	48 000	0.66	12–14	48 000	0.60	13–15
	70 000	0.80	12–14	72 000	0.75	13–15
	–	–	–	82 000	0.80	13–15

Note: A: Classical experimental design. B: Up-front exam design. Follow-up: Time is dated from entry.

The investigators found no statistically significant difference between the groups after 11–16 years of follow-up.[20]

At the time of randomization, the outcome of the initial physical exam was not known. Hence, the keep option was adopted in the analysis. We will calculate the power of this trial. Table 7 summarizes the values of the parameters used to calculate the power. The clinical trial was planned to detect a 40% reduction in mortality with a power of 0.80. According to our calculations this reduction is achieved at nine years of follow-up if the median survival of those found on screening is increased from 10 years to 40 years. Of course this is an idealized survival with no other causes of death. We regard this median survival to be unrealistic. To be more realistic we have calculated the power using a mortality reduction which ranges from 25% to 40%. The power calculations are summarized in Table 8 for both the drop and keep options.

The maximum power is achieved after 12–13 year of follow-up. Power decreases with longer follow-up. If a mortality reduction of 30% is considered realistic, the power for the keep and drop option are 0.79 and 0.91, respectively.

We next consider the COD. The COD is defined by offering the control group a single exam which coincides with the timing of the last exam given to the study group. This design was used in a breast cancer early detection trial began in 1981 in Stockholm. It is called the Stockholm Mammography Breast Screening Trial.[1,2] The trial had two exams in the study group and a delayed exam in the control group which

Table 6 Up-front design: effect of different numbers of exams (spaced two years apart) in the study group ($n = 24\,000$ in each group)

No. of exams in study group	Design	Power	Follow-up (years)
2	Classical	0.61	9–12
	Drop	0.20	10–12
	Keep	0.18	10–14
3	Classical	0.80	11–14
	Drop	0.42	12–15
	Keep	0.38	12–15
4	Classical	0.91	12–16
	Drop	0.65	14–16
	Keep	0.60	15–16

Table 7 Summary of parameter values used in the CNBSS study (age 40–49)

Parameter values	Women (40–49)
Median survival (control group), not prevalent at time 0	10 years[a]
Median survival (control group), having disease at time 0	10 years[a]
Median survival (study group) not prevalent at time 0	18.5/23/30/40 years[b]
Median survival (study group) having disease at time 0	18.5/23/30/40 years[b]
Sensitivity (study group)	0.81[c]
Sensitivity (control group)	0.56[c]
Prevalence	4.2 per 1000/year[a]
Transition from S_0 to S_p	2.0 per 1000/year[c]

[a] Median survival for the control interval and prevalence cases are estimated from the results of CNBSS trials.
[b] 25%, 30%, 35% and 40% reduction in mortality at nine years of follow-up from entry is used for the study group (these result in median survival of 18.5, 23, 30 and 40 years, respectively).
[c] Sensitivity and transition probabilities are estimated by Shen and Zelen.[23] Note that sensitivity for control group refers to physical exam.

was given, approximately, at the same time as the second exam in the study group. A comparison of the control versus study group answers the question of the magnitude of the benefit of a screening program compared to a delayed single exam. It does not provide unambiguous information on the benefit of early detection versus usual care.

Under the COD design, data may be analyzed in two ways: i) compare all deaths in each group or ii) only use deaths for cases diagnosed up to and including

Table 8 The statistical power in the CNBSS study (age 40–49) as a function of mortality reduction and follow-up time

Follow-up year[a]	Power (%)	$N_c = 25\,216$				$N_s = 25\,214$		
	Mortality reduction (keep)				Mortality reduction (drop)			
	25%	30%	35%	40%	25%	30%	35%	40%
4	32	38	45	51	49	54	58	62
5	41	50	60	68	61	68	74	78
6	49	61	72	81	70	78	84	89
7	55	68	80	88	76	84	90	94
8	60	73	85	92	80	88	93	96
9	62	76	87	94	82	89	94	97
10	64	78	89	95	83	90	95	98
11	64	79	89	95	83	91	95	98
12	64	79	90	95	83	91	96	98
13	64	79	90	95	82	90	95	98
14	63	78	89	95	81	90	95	98
15	62	77	88	95	80	89	95	97
16	60	76	88	94	79	88	94	97
17	59	74	87	94	77	87	93	97
18	57	72	85	93	76	86	93	96
19	55	71	84	92	74	84	92	96
20	53	69	83	91	73	83	91	95

[a] Follow-up time is dated from entry.

Table 9 Maximum power with three screening exams (spaced two years apart) in the study group (classical design versus close-out exam design, follow all cases)

Option	N_c	N_s	Power	Follow-up (years)	Comments
A	24 000	24 000	0.80	11–14	Classical
B	24 000	24 000	0.53	7–9	All cases followed
C	50 000	50 000	0.80	8–9	Increased equal sample size

Note: A: Classical experimental design. B and C: Close-out exam design – follow all cases. Follow-up: Time is dated from entry.

the last exam. Etzioni *et al.*[21] called the latter analysis a 'limited mortality analysis'. However, the COD design with the 'limited mortality analysis' will yield a biased result unless there is perfect sensitivity, no refusers and both groups have the last exam at the same time. This bias will favor the study group, but will not be present when all deaths are included in the analysis. A problem with this design is that there 'can be' or 'is to likely to be' slippage in the timing of the last exams which can introduce a bias.

Our first illustration of the COD compares the COD with the classical design for a study in which the study group has three exams, spaced 2 years apart and each group has 24 000 subjects (Table 9). The parameter values are the same as in Table 4. Calculations in Table 10 are made for a) including all deaths for the optimal follow-up time and b) including all diagnosed cases in the analysis when the number of diagnosed cases is the same in both groups. We have calculated the power for nine and 14 years as well as for the follow-up time to reach maximum power. Calculation was made for exam sensitivities ranging from 0.80 to 1.00. Generally, the analysis which includes all cases up to the 'catch-up' time will have higher power than including all cases. As expected, power decreases with lower sensitivity.

The COD may have a bias which depends on the exam sensitivity and the timing of the last exam. Our next illustration investigates this bias as a function of exam sensitivity assuming the timing of both exams are close.

We illustrate the effect of the bias by calculating the significance levels when the survival distribution is unaffected by the early detection of disease (null hypothesis)

Table 10 Statistical power with three screening exams offered to the study group (spaced two years apart). Comparison of classical design with COD under two options

Sensitivity	Power $N_c = N_s = 24\,000$					Catch-up year
	Classical	COD[a]	COD[b]			
1.00	0.86 (11–13)[c]	0.59 (7–9)	0.77 (9)	0.88 (14)	0.94 (23–41)[c]	4
0.98	0.85 (11–13)[c]	0.58 (7–9)	0.75 (9)	0.86 (14)	0.93 (24–39)[c]	5
0.95	0.83 (11–14)[c]	0.56 (7–9)	0.68 (9)	0.79 (14)	0.90 (29–39)[c]	7
0.90	0.80 (11–14)[c]	0.53 (7–9)	0.53 (9)	0.70 (14)	0.84 (31–45)[c]	9
0.80	0.74 (12–13)[c]	0.47 (8–9)	0.47 (9)	0.58 (14)	0.76 (36–49)[c]	11

[a] Close-out design: follow all participants. Maximum power is achieved with 7–9 years of follow-up.
[b] Close-out design: follow cases to the 'catch-up' time. Analysis is done at nine, 14 and 23+ years of follow-up.
[c] Maximum power. Numbers in parentheses refer to follow-up years dated from entry.

Table 11 COD (only include cases diagnosed up to and including last exam): Levels of significance for a range of sensitivity values ($\beta = 0.90$, 0.75, 0.50) and follow-up times $T = 2(2)20$ with $n = 55\,000$ in each group (for a nominal level of significance equal to 0.05). Departures from the nominal level (0.05) indicate bias

Year	Sensitivity = 0.90			Sensitivity = 0.75			Sensitivity = 0.50		
	No. of exams (study) versus single (control)			No. of exams (study) versus single (control)			No. of exams (study) versus single (control)		
	2	3	4	2	3	4	2	3	4
2	0.07	–	–	0.09	–	–	0.13	–	–
4	0.07	0.08	–	0.11	0.13	–	0.17	0.23	–
6	0.07	0.08	0.08	0.11	0.16	0.16	0.17	0.29	0.29
8	0.07	0.08	0.09	0.10	0.15	0.18	0.16	0.29	0.36
10	0.06	0.08	0.09	0.09	0.14	0.18	0.14	0.27	0.35
12	0.06	0.08	0.08	0.09	0.13	0.16	0.13	0.24	0.32
14	0.06	0.07	0.08	0.08	0.12	0.15	0.12	0.21	0.29
16	0.06	0.07	0.08	0.08	0.11	0.14	0.11	0.19	0.26
18	0.06	0.07	0.07	0.08	0.10	0.12	0.10	0.17	0.23
20	0.06	0.07	0.07	0.07	0.10	0.11	0.09	0.16	0.20

Table 12 Summary of parameter values

Parameter	Value
Median survival (control = study)	9 years
Prevalence at entry	14 per 1000/year
Transition from S_0 to S_p	3.7 per 1000/year
Sensitivity	0.90, 0.75, 0.50
Number of exams (study group)	2–4
Time interval between exams	2 years

Notes: a) Median survival for the control interval cases and transition probability are estimated from SEER 1988–1996 data for women age 50+. b) Prevalence is estimated from the results of CNBSS trial for women 50–59 (Shen and Zelen).[23]

Table 13 Summary of parameter values used in Stockholm example

Parameter	Women (50–64)
Median survival (control group), not prevalent at time 0	10 years[a]
Median survival (control group), having disease at time 0	10 years[a]
Median survival (study group), not prevalent at time 0	18.5/23/30/40 years[b]
Median survival (control group), having disease at time 0	18.5/23/30/40 years[b]
Prevalence	5.9 per 1000/year[c]
Transition from S_0 to S_p	1.9 per 1000/year[c]
Sensitivity (control = study)	0.89[d]

[a] Median survival for the control interval cases are estimated from SEER 1988–1996 data.
[b] The same assumptions are made as those in CNBSS example (Table 7).
[c] Prevalence and transition probability are estimated from the results for Stockholm trial.
[d] Sensitivity is estimated by Shen and Zelen.[23]

Table 14 The statistical power in the Stockholm trial: women aged 50–64 compared to classical design

Follow-up year[a]	Power (%) $N_c = 12\,840$				$N_s = 24\,789$			
	Classical				Close-out[b]			
	18.5	23	30	40	18.5	23	30	40
4	12	14	17	20	11	13	15	17
5	16	21	26	31	16	21	26	31
6	21	28	36	43	20	26	33	39
7	26	35	45	54	21	28	36	43
8	30	41	52	62	21	28	37	45
9	33	45	58	69	21	28	37	45
10	35	49	62	73	20	27	36	44
11	37	51	65	76	19	26	35	43
12	38	52	67	78	18	25	33	41
13	38	53	69	80	18	24	31	39
14	38	54	69	80	17	23	30	38
15	38	54	69	81	16	21	28	36
16	38	53	69	81	15	20	27	34
17	37	52	69	81	14	19	26	32
18	36	52	68	80	14	18	24	31
19	35	50	67	79	13	18	23	29
20	34	49	66	78	13	17	22	28

Notes: Power is shown for four values of median survival (18.5, 23, 30, 40 year) for those diagnosed at a scheduled exam.

[a] Follow-up time is dated from entry.

[b] All participants are followed.

and cases are only included in the analysis for those cases diagnosed up to and including the timing of the last exam. The calculations are summarized in Table 11 using the parameter values in Table 12. The significance level was chosen to be 5%. Values departing from 5% show a bias. It is clear that the bias increases as the sensitivity decreases. Shorter follow-up times have a slightly higher bias than long follow-up times. In addition, the bias increases with more exams in the study group. The overall effect of the bias is that the actual type I error is larger than the nominal 5% type I error. Consequently, it will result in a higher false positive rate.

We have used our model to calculate the power of the Stockholm trial which used a COD. The power calculations are summarized in Table 14, based on the parameter values of Table 13. The calculations in Table 14 are compared to a classical design. The power has been calculated assuming a median survival for those diagnosed from a screening exam of 18.5, 23, 30 and 40 years. The median survival of the control group dated from diagnosis is 10 years. The columns labeled 'close-out' refer to the power of the Stockholm trial assuming all deaths are included in the analysis. The power is very low over the range of alternatives being considered. Note that the maximum power for the close-out trial is achieved with eight years of follow-up, whereas for a classical design the maximum power is achieved at about 15 years of follow-up.

4 Discussion

In this article, we discussed several issues for planning randomized early detection trials. Our theoretical results enable one to plan early detection trials to be optimized as a function of the number of exams, spacings of the exams and follow-up time. Furthermore, our theoretical results allow one to plan such experimental designs such that the control group may receive some potential benefit. However, such trials must have considerably larger sample sizes than a classical design in order to have acceptable power.

We have illustrated the theoretical results using several examples. In general, we have shown that the power may be increased by having larger intervals between examinations. However, with larger intervals it is necessary to have longer follow-up time to reach the maximal power. If the intervals between examinations are relatively small, long-term follow-up may not result in much of an increase in power. Eventually, the power will decrease if the follow-up period is too long. The need for long follow-up time to achieve acceptable power diseases with increasing sample size.

Furthermore, if the major costs of a trial are related to the number of exams, it is possible to have increased power by having fewer subjects in the study group but with a larger number of examinations.

The COD and the UFD offer ways in which the control group may possibly benefit. Both design have very much reduced power compared with the classical design of a study versus a control group. Non-significance in the UFD does not necessarily mean the screening is without benefit. This holds true for both the keep and drop options. Actually, all of the breast trials conducted in Sweden eventually offered the control groups mammogram examinations. This change in the experimental design may introduce biases in favor of the study group which are a function of the examination sensitivity and the timing of the control group examinations.

We have not discussed issues of cluster randomization which was used in four of the eight breast cancer clinical trials. Cluster randomization requires that the mortality associated with the study and control groups be calculated using the cluster as the basic unit. Cluster randomization generally results in lower power than randomizing individuals. However, because the mortality rate is so low, the power may only be marginally reduced.

It is our impression that there have been many lost opportunities because of the inadequate planning of early detection trials. Perhaps the insights provided by our modeling will alert investigators as to how the trials can be better planned. We hope that further trials will take advantage of our theoretical results and plan these expensive and high impact trials in an optimal way.

Acknowledgements

This research was partially supported by a Public Service Grant CA-88270 from the National Cancer Institute. We thank Dr Phillip Prorok for a careful reading of the article and making suggestions which improved the exposition. All figures and tables are reproduced with the permission of Oxford University Press and are taken from reference 17 and reference 20.

References

1 Frisell J, Eklund G, Hellstrom L, Lidbrink E, Rutqvist LE, Somell A. Randomized study of mammography screening: preliminary report on mortality in the Stockholm trial. *Breast Cancer Research and Treatment* 1991; **18**: 49–56.

2 Frisell J, Lidbrink E, Hellstrom L, Rutqvist LE. Followup after 11 years—update of mortality results in the Stockholm mammographic screening trial. *Breast Cancer Research and Treatment* 1997; **45**: 263–70.

3 Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Canadian Medical Assocociation Journal* 1992a; **147**: 1459–76.

4 Miller AB, Baines CJ, To T, Wall C. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Canadian Medical Association Journal* 1992b; **147**: 1477–88.

5 Nystrom L, Rutqvist LE, Wall S, Lindgren A. *et al*. Breast cancer screening with mammography: overview of Swedish randomised trials. *Lancet* 1993; **341**: 973–78.

6 Nystrom L, Andersson I, Bjurstam N, Frisell J, Nordenskjolb B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. *Lancet* **359**: 909–19.

7 Roberts MM, Alexander FE, Anderson TJ, Chetty U, Donnan PT, Forrest P, Hepburn W, Huggins A, Kirkpatrick AE, Lamb J, Muir BB, Prescott RJ. Edinburgh trial of screening for breast cancer: mortality at seven years. *Lancet* 1990; **335**: 241–46.

8 Shapiro S, Venet W, Strax PH, Venet L, Rosser R. Ten to fourteen year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute* 1982; **69**: 349–55.

9 Shapiro S, Venet W, Strax PH, Venet L, Rosser R. Selection, follow-up, and analysis in the Health Insurance Plan Study: a randomized trial with breast cancer screening. *National Cancer Institute Monograph* 1985; **67**: 65–74.

10 Shapiro, S. Periodic screening for breast cancer: the HIP randomized controlled trial. Health insurance plan. *National Cancer Institute Monograph* 1997; **22**: 27–30.

11 Tabar L, Fagerberg CJG, Gad A, Baldetorp L, Holmberg LH, Grontoft O, Ljungquist U, Lundstrom B, Manson JC, Kopparberg County Project Group, Ostergotland County Project Group, Eklund G, Day NE, Pettersson F. Reduction in mortality from breast cancer after mass screening with mammography. *Lancet* 1985; **1**: 829–32.

12 Tabar L, Vitak B, Chen HH, Duffy SW, Yen MF, Chiang CF, Krusemo UB, Tot T. Smith RA. The Swedish two-county trial twenty years later., Updated mortality results and new insights from long-term follow-up. *Radiologic Clinics of North America* 2000; **38**: 625–51.

13 Gotzsche PC, Olsen, O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000; **355**: 129–34.

14 Prorok P, Connor R, Baker S. Statistical considerations in cancer screening programs. *Urologic Clinics of North America* 1990; **17**: 699–708.

15 Zelen M, Feinleib M. On the theory of screening for chronic diseases. *Biometrika* 1969; **56**: 601–14.

16 Zelen M. Optimal scheduling of examination for the early detection of disease. *Biometrika* 1993; **80**: 279–93.

17 Hu P, Zelen M. Planning clinical trials to evaluate early detection programmes. *Biometrika* 1997; **84**: 817–29.

18 Walter SD, Day NE. Estimation of the duration of a pre-clinical disease state using screening data. *American Journal of Epidemiology* 1983; **118**: 865–86.

19 Stomper PC, Gelman RS. Mammography in symptomatic and asymptomatic patients. *Hematology/Oncology Clinics of North America* 1989; **3**: 611–40.

20 Miller AB, To T, Baines CJ, Wall C. The Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up. A randomized screening trial of mammography in women age 40 to 49 years. *Annals of Internal Medicine* 2002; **137**: 305–12.

21 Etzioni RD, Connor RJ, Prorok PC, Self SG. Design and analysis of cancer screening trials. *Statistical Methods in Medical Research* 1995; **4**: 3–17.

22 Hu P, Zelen M. Experimental design issues for the early detection of disease: novel designs *Biostatistics* 2002; **3**: 299–313.

23 Shen Y, Zelen M. Parametric estimation procedures for screening programmes: stable and nonstable disease models for multimodality case finding. *Biometrika* 1999; **86**: 503–15.