Attachment 13
Sample Size & Power Calculations

Sample-Size Calculations

Sample size was calculated using the method suggested by Taylor and Fontana [125], modified to allow for arbitrary magnitude of screening impact, arbitrary sample-size ratio between screened and control arms, and arbitrary levels of compliance in the screened and control arms. Let N_C be the number of individuals randomized to the control arm and N_S be the number randomized to the screened arm, with $N_S = f N_C$, where f is a proportionality constant. For $0 < r < 1$, assume the trial is designed to detect a $(1 - r) \times 100\%$ reduction in the cumulative disease-specific death rate over the duration of the trial. Also let P_C be the proportion of individuals in the control arm who comply with the usual-care protocol and P_S be the proportion of individuals in the screened group who comply with the screening protocol. The total number of disease-specific deaths needed for a one-sided α-level significance test with power $1 - \beta$ is then:

$$D = \frac{[(Q_C + f\,Q_S)\,Z_{1-\alpha} - \sqrt{Q_C Q_S}\,(1 + f)\,Z_\beta]^2}{f\,(Q_C - Q_S)^2}$$

where $Q_C = r + (1 - r)P_C$ and $Q_S = 1 - (1 - r)P_S$. The number of participants required in the control arm is:

$$N_C = \frac{D}{(Q_C + f\,Q_S)\,R_C Y}$$

where Y is the duration of the trial from entry to end of follow-up in years and R_C is the average annual disease-specific death rate in the control arm expressed in deaths per person per year.

A one-sided hypothesis testing approach to sample-size calculation was employed based on the nature of the question being addressed. The PLCO trial is intended to provide definitive evidence of the effect of screening on cause-specific mortality compared to usual medical care, analogous to phase III placebo-controlled trials in the therapeutic setting. The focal question for each of the four cancers is whether screening reduces mortality. This is inherently a one-sided research question, implying a one-sided design and analysis approach. The question is not whether screening reduces *or* increases mortality. Determining whether screening increases mortality is not an objective of this trial. Furthermore, if the screening intervention has no effect or if it is harmful, the consequences in terms of a public health decision are the same—screening is not recommended. This further dictates a one-sided approach [126].

The estimation procedure is illustrated for prostate cancer for white males. Prostate cancer screening was the impetus for the trial and is the primary focus for sample-size calculations. Similar calculations can be done for the other sites using the data in Table 3. This illustration is based on calculations done for the original design prior to the pilot phase, when the eligible age range was 60–74 years and the trial duration was 10 years from randomization for each participant.

Table 3 Cancer Mortality Rates per Person per Year ($\times 10^{-5}$), Estimated Using 1983–1987 Data

	Prostate		Ovarian		Colorectal	Lung
Age (years)	White	Black	White	Black	White[a]	White[b]
Males						
50–54	3.5	11.1	—	—	20.8	88.4
55–59	11.5	31.9	—	—	39.6	165.4
60–64	30.4	80.4	—	—	64.6	252.6
65–69	71.1	174.1	—	—	104.4	367.6
70–74	137.8	332.3	—	—	156.1	470.2
75–79	244.8	515.7	—	—	216.0	543.9
80–84	402.8	838.7	—	—	296.0	555.0
85+	606.3	937.1	—	—	378.6	441.3
Females						
50–54	—	—	14.2	10.4	16.5	46.5
55–59	—	—	20.3	15.3	28.1	75.3
60–64	—	—	27.5	23.3	43.9	104.9
65–69	—	—	35.3	27.4	67.9	138.0
70–74	—	—	41.5	33.8	100.1	152.9
75–79	—	—	45.2	34.5	141.9	143.8
80–84	—	—	49.8	41.1	200.5	127.2
85+	—	—	44.7	35.0	289.2	103.5

[a] Rates for black males are very similar. Rates for black females in age group 65–79 years are about 15% higher.

[b] Average rate for black males in age group 65–79 years is about 13% higher. Average rate for black females in age group 65–79 years is about 20% lower.

Calculation of N_C requires an estimate of R_C. It was assumed that the trial would enroll an equal number of participants in each of three age strata: 60–64, 65–69, and 70–74 years. Because individuals recruited for screening trials are expected to be healthier than the general population, the usual cancer mortality rate obtained from national or registry data will overestimate the mortality rate of the participants, at least for the early part of the trial. Therefore, for a 10-year prostate cancer screening trial with men entered between the ages of 60–74, it was assumed that for the first 2 years the mortality rate in the control arm is 25% of the usual rate, for the next 3 years it is 50% of the usual rate, and for the last 5 years it equals the usual rate. The usual mortality rate was estimated by the unweighted average prostate cancer mortality rate for men ages 65–79 years. This age range was used to adjust for aging over the 10 years of the trial. The usual mortality rates from national data are shown in Table 3 [127]. The estimated rate for this example is $R_C = 103.763 \times 10^{-5}$.

Results of sample-size calculations for the trial are given in Table 4. These calculations assume a 10-year trial using a one-sided, 0.05-level test, $P_C = P_S = 1$, and possible mortality reductions as shown in a screened group compared to an equal-sized, usual care group ($f = 1$). The sample sizes are based on mortality rates for whites. Including blacks in the trial does not substantially alter sample size. A sample size of 37,000 (rounded up from 36,221 in Table 4) screened and 37,000 controls of each gender was chosen on the following basis. A high power of at least 90% is mandatory to yield a meaningful negative result, should that happen, and to achieve a high level of scientific validity

Table 4 Number of Participants Ages 60–74 Years at Entry Needed in Each Arm of the Trial

		Mortality Reduction (%)			
Site	Power	10	20	30	35
Prostate	0.9	153,577	36,221	15,078	
(males)	0.8	110,906	26,182	10,920	
Lung	0.9	76,721	18,095		
(males and females)	0.8	55,404	13,080		
Colorectum	0.9	177,208	41,794	17,397	
(males and females)	0.8	127,971	30,211	12,600	
Ovary	0.9		134,697	56,069	39,733
(females)	0.8		97,365	40,606	28,817

because a trial of this magnitude addressing these questions is not likely to be repeated. In addition, it was felt that for an effect of prostate cancer or colorectal cancer screening to be of public health importance, it must be at least 20% or greater. Given the magnitude of the lung cancer problem, it was felt that a screening effect of 10% or greater would be very important. To estimate whether a 20% effect for prostate cancer screening was realistic, two calculations were performed. The first used plausible stage shifts due to screening and survival by stage to project possible improved outcome for screen-detected cancers. The second used projections from a computer model [128]. Both gave mortality reduction estimates in the range of 25% with perfect compliance.

Power calculations are displayed in Table 5. With 37,000 men and women in each arm, the trial has 91% power to detect a 20% mortality reduction in prostate cancer mortality and 89% power to detect a 10% lung cancer mortality reduction. The power is nearly 90% to detect a 15% colorectal cancer mortality reduction and 99% for a 20% effect. For ovarian cancer, the power is nearly 90% to detect a 35% mortality reduction.

It was recognized that compliance will not be perfect in either randomized group. Contamination or drop-in will occur in the control arm ($P_C < 1$) and noncompliance or dropout is to be anticipated in the screened arm ($P_S < 1$). The target mortality reductions of 20% for prostate and colorectal cancers and 10% for lung cancer therefore are to be interpreted as effects that the trial seeks

Table 5 Power by Percent Reduction in Mortality with 37,000 Men and 37,000 Women in Each Arm

		Mortality Reduction (%)						
Site	Gender	5	10	15	20	25	30	35
Prostate	male	—	—	0.71	0.91	0.98	—	—
Lung	both genders	0.41	0.89	0.997	—	—	—	—
	female	0.17	0.41	0.69	—	—	—	—
	male	0.34	0.81	0.985	—	—	—	—
Colorectum	both genders	—	—	0.89	0.99	0.999	—	—
	female	—	—	0.56	0.79	0.93	—	—
	male	—	—	0.72	0.92	0.99	—	—
Ovary	female	—	—	—	0.45	0.62	0.77	0.88

Table 6 Percent Mortality Reduction Required When Compliance Is 100% in Both Groups, Based on a Mortality Reduction of 20% in the Presence of Noncompliance, as a Function of P_S and P_C

Compliance in the Control Group (P_C)	Compliance in the Screened Group (P_S)					
	0.5	0.6	0.7	0.8	0.9	1.0
0.5	—	100	67	50	40	33
0.6	90	71	53	42	34	29
0.7	77	56	43	36	30	26
0.8	59	45	37	31	27	24
0.9	48	39	32	28	24	22
1.0	40	33	29	25	22	20

to detect in the presence of whatever noncompliance and contamination exist in the populations. This implies that if there were perfect compliance, the mortality reductions would be greater since they would not be diminished by noncompliance.

One can assess the relationship between true effect size and level of noncompliance during the screening period by examining Table 6, which shows what the mortality reductions with perfect compliance would have to be to realize a 20% mortality reduction for various levels of noncompliance in the screened and control groups. For example, if 90% of participants in the screened group undergo a PSA test ($P_S = 0.9$) while 20% of controls are so screened ($P_C = 0.8$), then the prostate cancer mortality reduction from such screening would have to be 27% with perfect compliance for there to be a 20% effect in the presence of noncompliance. The 27% figure corresponds very closely to the modeling estimate. Thus, compliance of at least 90% and contamination of no greater than 20% for prostate cancer screening, particularly with PSA, were chosen as the target values for these parameters.

Inquiries into potential screening compliance and screening contamination for the four cancer sites being studied in this trial indicated that the ranges of reasonable target values at the time of initiation of recruitment were as shown in Table 7. In addition to direct contact with health maintenance organizations and existing SCs, published data from the 1987 National Health Interview Survey were used to gauge these effects [129, 130]. These numbers were necessarily somewhat subjective. Additional estimates were obtained directly from the trial population during the pilot phase, and further assessment will occur as the trial progresses, possibly leading to sample-size adjustment.

In the context of these levels of contamination and compliance, the required true levels of mortality effect (effect size) with perfect compliance are, approximately, lung 20%, colon 25%, and prostate 25%. These requirements are consistent with expected effects based on modeling efforts [74, 75, 103].

Regarding the ovarian cancer objectives of this trial, if the mortality reduction from screening for ovarian cancer were 35%, this design would have almost a 90% power to demonstrate this effect. However, if the mortality effect were only 25%, 84,000 screened women and an equal number of controls would be required to achieve 90% power. Thus, the ovarian component of this trial is to be viewed as a two-step process. Near the end of the screening phase of the trial, sufficient cases of ovarian cancer should accrue to provide good estimates

Table 7 Design Contamination and Compliance Ranges Projected by Modality

	Compliance (%)	Contamination (%)
Digital rectal exam	>90	<20
Prostate-specific antigen	>90	<20
CA125	>90	<10
Ovarian palpation	>85	<10
Transvaginal ultrasound	>85	<10
Sigmoidoscopy	>85	<15
Chest X-ray	>85	<40

of sensitivity for each screening modality. Specificity and predictive value can also be estimated. If as a result any one or combination of the tests appears sufficiently promising to justify a full mortality study, the female population base of this trial could be supplemented or a meta-analysis of data from this trial and other relevant studies could be done to increase power.

As noted above, in January 1996 the lower age limit for trial participation was reduced from 60 to 55 years. Given the lower mortality rates in the 55–59 age stratum, this would ordinarily imply the need for an increase in the sample size. However, this protocol change took place after the April 1995 eligibility criterion change, also noted above, to exclude men who had prior repeat PSA screening, thereby reducing the contamination level. Sample-size estimates for prostate cancer screening for the age range 55–74 years are shown in Table 8. For compliance of 90% and a revised estimate of contamination of 10–15%, a sample of 37,000 men (and therefore 37,000 women) in each trial arm is still appropriate. A similar conclusion holds for the other cancer sites as well. As mentioned, this estimate is monitored regularly during the enrollment phase of the trial to determine if adjustment is required.

Based on the monitoring of design parameters, further protocol modifications were adopted in December 1998. These were to change from a 3-year to a 5-year interval for flexible sigmoidoscopy for individuals who had not yet had their second exam, and at the same time to add year 4 and 5 PSA and CA125 tests. Also, the remaining third annual chest X-ray exams are offered only to current or former smokers, and follow-up is extended 3 years, so that all participants will be followed at least 13 years from randomization. A final change was that the ovarian palpation exam, which had been part of the original protocol, was eliminated.

Table 8 Number of Males Required in Each Arm to Achieve 90% Power with Age at Entry Range 55–74 Years, as a Function of P_S and P_C

	P_S		
P_C	0.85	0.90	0.95
0.80	53,057	45,338	39,134
0.85	46,087	39,787	34,650
0.90	40,440	35,225	30,918

The interval between flexible sigmoidoscopy was lengthened to coincide with recommendations in the community and was based on preliminary information suggesting that sigmoidoscopy at 3 years finds polyps, but very few are likely to be of any significance. A delay of 2 years was expected to yield more polyps and cancers, leading to a greater potential for mortality reduction. The addition of 2 extra years of PSA and CA125 blood tests and at least 3 additional years of follow-up were adopted to provide assurance of sufficient screening effect and statistical power in the event that initial design assumptions were incorrect. The final round of chest X-ray testing for individuals who never smoked was eliminated because of the very low yield of this exam. Finally, the ovarian palpation exam was deleted because of very low yield and the fact that a very high proportion of women participating in the trial regularly underwent pelvic examination, thereby diluting any possible effect of the palpation exam.