Should we expect all-cause mortality reductions in large screening studies?

INTRODUCTION

Randomised controlled trials of cancer screening are generally designed with disease-specific mortality (DSM) as the primary outcome. All-cause mortality (ACM) is often not reported, or reported only as a secondary outcome.[1,2] Implicit in the choice of DSM as the primary outcome is that the screening intervention will not have an adverse effect on other causes of death, or at most that effect will be small in comparison with the DSM benefit. For example, the radiation exposure from mammography likely causes some cancers, but this number is small in comparison with the number of breast cancer deaths prevented.[3] Harms due to overdiagnosis and overtreatment are more difficult to estimate, and may be substantial.[4–6] Commentators have therefore argued that ACM is the preferred outcome for cancer screening trials, because DSM is a biased outcome due to incorrect assignment of the cause of death and failure to fully account for harms.[7,8]

RELATIONSHIPS BETWEEN DSM AND ACM

If DSM decreases under screening, then there are three possible relationships between DSM and ACM in a cancer screening trial. First, both DSM and ACM may decrease by approximately the same absolute number of deaths, suggesting no important harms of screening. Second, DSM may decrease but ACM does not change significantly, suggesting that any benefits of screening are offset by harms. Third, ACM may increase while DSM decreases, suggesting that unintended harms of screening are greater than the benefits. An important limitation of determining which of these patterns has occurred in an individual screening trial is the difference in sample sizes needed to demonstrate a reduction in DSM as opposed to ACM.

In a recent randomised controlled trial of screening for ovarian cancer (the United Kingdom Controlled Trial Ovarian Cancer Screening or UKCTOCS), a post-hoc analysis concluded that screening reduced DSM.[1] Although not directly reported by the investigators, a review of data in the appendices revealed that ACM did not decrease, and actually increased slightly (although non-significantly). This is in contrast with recent randomised controlled trials of screening for breast cancer and lung cancer, where both DSM and ACM were reduced in the screened groups (Table 1).[9,10] Here we point out one possible explanation for these types of discrepancies between results for DSM and ACM in screening studies.

Table 1. Disease-specific mortality and all-cause mortality rates from three large, contemporary trials of cancer screening: UKCTOCS (ovarian cancer), UK AGE (breast cancer), and NLST (lung cancer)[a]

		Deaths/100 000 persons					
Disease	Outcome	Screened	Unscreened	Reduction in deaths (95% CI)	SE	*Z*	*P*-value
Ovarian cancer[1]	Disease-specific mortality	292	342	50 (–9 to 109)	30	1.67	0.10
	All-cause mortality	6667	6569	–98 (–363 to 167)	135	–0.73	0.47
Breast cancer[9]	Disease-specific mortality	338	385	47 (–14 to 108)	31	1.50	0.13
	All-cause mortality	3947	4039	92 (–110 to 294)	103	0.89	0.37
Lung cancer[10]	Disease-specific mortality	1308	1620	312 (106 to 518)	105	2.98	0.003
	All-cause mortality	7024	7482	457 (18 to 896)	224	2.04	0.04

[a] *Results have been standardised to deaths/100 000 persons over the duration of the study. The standard errors are calculated as* $SE(R_1 - R_2) = \sqrt{Var(R_1) + Var(R_2)}$ *where* $Var(R_1) = m^2 (p_1(1 - p_1)/n_1)$. *An R markdown file showing the calculations is available from the authors on request.*

EXPLAINING THE DISCREPANCIES

Because the deaths in DSM are a subset of deaths in ACM, the rate of DSM is smaller than the rate of ACM. Indeed, in the UKCTOCS study of ovarian cancer screening, only 5% of the observed deaths from all causes were attributed to ovarian cancer. This creates a problem when trying to detect differences in ACM: the standard error (SE) of a mortality rate estimate $\hat{p}$ is $s.e.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$, where p is the true mortality rate and n is the sample size. The SE is largest when $p = 1/2$ and approaches 0 as p approaches 0. For example, if $n = 50\,000$ are screened, then the SE for a DSM of 1% is 0.0004, whereas for an ACM of 20% the SE is 0.0018 — more than four times as large. Thus, there is always greater uncertainty about ACM estimates than there is about DSM estimates, leading to a larger required sample size to detect a significant difference between rates.

Turning to the DSM endpoint of the UKCTOCS study first (Table 1), there was an estimated rate of 292 ovarian cancer deaths/100 000 women in the screened arm and 342 ovarian cancer deaths/100 000 women in the unscreened arm, based on sample sizes of 50 640 and 101 359 respectively.[1] The estimated change in the DSM rate is 50 fewer deaths/100 000 in the screening arm with a standard error of 30. On the other hand, the estimated difference in ACM is an increase of 98 deaths/100 000 in the screening group, with a standard error of 135. Even though the change is almost twice as large in the ACM arm as it is in the DSM arm (and in the wrong direction), the standard error is four times larger for the estimate of ACM. The result is that a two-sided *P*-value for DSM is $P = 0.10$ whereas for ACM the corresponding *P*-value is $P = 0.47$. For a fixed sample size, there is more statistical information in the DSM estimate than the ACM estimate. Importantly, the difference we are noting here is purely statistical and results from the fact that the variance for a binomial random variable, unlike a normal random variable, changes with its mean.

More generally, we can define an inflation factor as the number by which the ACM sample size must be multiplied to achieve the power of the DSM sample size when trying to detect a common difference between two rates ($\Delta = r_1 - r_2$), and assuming equal sample sizes for screened and unscreened. Results are shown in Figure 1. Note that, in the ovarian cancer study, the ratio of ACM to DSM is estimated

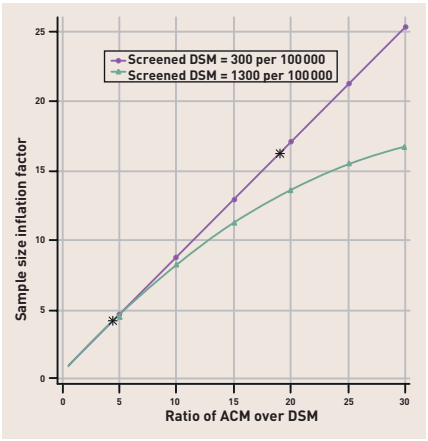

Figure 1. Sample-size inflation factors based on 90% power to detect a difference at the 0.05 level. To achieve equivalent power, the ACM sample size must be multiplied by the inflation factor. For the DSM = 300 case, the detectable difference in rates is 50 per 100 000 — corresponding to the UKCTOCS trial. For the DSM = 1300, the detectable difference in rates is 300 per 100 000 — roughly corresponding to the NLST trial. The '*' indicate the actual ratios observed in the unscreened arms of the UKCTOCS and NLST trials.

to be about 20. Put another way, to have similar statistical power for ACM as for DSM, a study would have to be 20 times larger.

Returning to our original three possibilities for the relationship between DSM and ACM, the AGE and NLST trials both found a statistically significant reduction in DSM and no significant reduction in ACM.[9,10] Although at first glance this might put these studies in the second category (DSM decreases while ACM remains the same), the absolute reduction in death rates was similar for DSM and ACM in both studies, and the reduction in ACM was in fact a bit higher than the DSM reduction (Table 1). To clinicians, it provides some reassurance that ACM is moving in the same direction as DSM, even if statistical significance cannot be demonstrated. In the UKCTOCS results, though, ACM increases while DSM decreases.[1] Such discrepancies may represent random variation or a real effect due to harms of interventions and surgeries, although the confidence interval around the estimate of ACM is broad, and ranges from a reduction of 167 deaths/100 000 to an increase of 363 deaths per 100 000. We cannot conclude that ovarian cancer screening increases ACM, and the groups appeared to be balanced at baseline and the ratio of false positive to true positive surgeries was admirably low. However, it highlights the need for careful follow-up and ascertainment of the causes of death in study participants.

Because screening studies are sized to detect differences in DSM rather than ACM, changes in the direction of DSM and ACM are to be expected due to random variation. The probability of observing changes in direction will increase as the ratio ACM:DSM increases. If this ratio is ACM:DSM = 1, then all deaths are due to the disease in question; in this case the probability of a change in direction of DSM and ACM is zero. The two must agree. As the ratio increases, however, the stochastic variation in ACM will increase and the probability of a change in direction will approach 50%. This assumes the study is powered to detect DSM. If the study is powered instead to detect ACM, then the sample size will be much larger, and the probability of a change in direction will be small. For example, we can look at the likelihood of the discrepancy observed in the UKCTOCS, where the ACM/DSM ratio was around 20. Under some simplifying assumptions listed in the appendix (available from the authors on request), the probability of observing a discrepancy as large as the –98 per 100 000 (observed value) using $n = 151\,999$ is 14%. Doubling the sample size (to $n = 303\,998$) decreases the probability to 6%; tripling the sample size reduces the probability further to 3%.

CONCLUSION

In conclusion, both DSM and ACM are important, and both should be reported in all randomised controlled trials of screening. However, the failure to detect a statistically significant reduction in ACM, even in very large studies, is not surprising. The focus should be on the absolute magnitude of mortality reduction, and understanding that finding consistency in the direction and absolute magnitude of DSM and ACM is reassuring. We report a method for calculating the likelihood that these outcomes would move in opposite directions, and propose the ACM/DSM ratio as a way to understand the danger of over-interpreting ACM comparisons in studies powered to detect changes in DSM.

Kevin K Dobbin,
Associate Professor of Biostatistics, College of Public Health, University of Georgia, Athens, GA, US.

Mark Ebell,
Professor of Epidemiology, College of Public Health, University of Georgia, Athens, GA, US.

ADDRESS FOR CORRESPONDENCE

Mark H Ebell
233 Miller Hall, University of Georgia Health Sciences Campus, Athens, GA 30602, US.

Email: ebell@uga.edu

Provenance
Freely submitted; not externally peer reviewed.

DOI: https://doi.org/10.3399/bjgp18X696545

REFERENCES

1. Jacobs IJ, Menon U, Ryan A, *et al.* Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet* 2016; **387(10022):** 945–956.
2. Schröder FH, Hugosson J, Roobol MJ, *et al.* Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet* 2014; **384(9959):** 2027–2035.
3. Miglioretti DL, Lange J, van den Broek JJ, *et al.* Radiation-induced breast cancer incidence and mortality from digital mammography screening: a modeling study. *Ann Intern Med* 2016; **164(4):** 205–214.
4. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *New Engl J Med* 2012; **367(21):** 1998-2005.
5. Patz EF Jr, Pinsky P, Gatsonis C, *et al.* Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med* 2014; **174(2):** 269–274.
6. Etzioni R, Gulati R, Mallinger L, Mandelblatt J. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Ann Intern Med* 2013; **158(11):** 831–838.
7. Black WC, Haggstrom DA, Welch HG. All-cause mortality in randomized trials of cancer screening. *J Natl Cancer Inst* 2002; **94(3):** 167–173.
8. Penston J. Should we use total mortality rather than cancer specific mortality to judge cancer screening programmes? Yes. *BMJ* 2011; **343:** d6395.
9. Moss SM, Wale C, Smith R, *et al.* Effect of mammographic screening from age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a randomised controlled trial. *Lancet Oncol* 2015; **16(9):** 1123–1132.
10. National Lung Screening Trial Research Team; Aberle DR, Adams AM, Berg CD, *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *New Engl J Med* 2011; **365(5):** 395–409.
11. Chow S-C, Liu J-P. *Design and analysis of clinical trials.* 2nd edn. Hoboken, NJ: Wiley, 2004.