FULL PAPER

BJC

British Journal of Cancer (2013) 109, 2014–2019 | doi: 10.1038/bjc.2013.427

Keywords: screening; lead time; overdiagnosis

Overestimated lead times in cancer screening has led to substantial underestimation of overdiagnosis

P-H Zahl*[,1], K J Jørgensen[2] and P C Gøtzsche[2]

[1]Norwegian Institute of Public Health, PO Box 4404 Nydalen, N-0403 Oslo, Norway and [2]The Nordic Cochrane Centre, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark

Background: Published lead time estimates in breast cancer screening vary from 1 to 7 years and the percentages of overdiagnosis vary from 0 to 75%. The differences are usually explained as random variations. We study how much can be explained by using different definitions and methods.

Methods: We estimated the clinically relevant lead time based on the observed incidence reduction after attending the last screening round in the Norwegian mammography screening programme. We compared this estimate with estimates based on models that do not take overdiagnosis into account (model-based lead times), for varying levels of overdiagnosis. Finally, we calculated overdiagnosis adjusted for clinical and model-based lead times and compared results.

Results: Clinical lead time was about one year based on the reduction in incidence in women previously offered screening. When overdiagnosed tumours were included, the estimates increased to 4–9 years, depending on the age at which screening begins and the level of overdiagnosis. Including all breast cancers detected in women long after the end of the screening programme dilutes the level of overdiagnosis by a factor of 2–3.

Conclusion: When overdiagnosis is not taken into account, lead time is substantially overestimated. Overdiagnosis adjusted for model-based lead time is a function tending to zero, with no simple interpretation. Furthermore, the estimates are not in general comparable, because they depend on both the duration of screening and duration of follow-up. In contrast, overdiagnosis adjusted for clinically relevant tumours is a point estimate (and interpreted as percentage), which we find is the most reasonable method.

Overdiagnosis and lead time. The natural history of tumours and their growth rates are difficult to study, because almost all patients are treated. We also know very little about when tumours metastasise, or whether all tumours have the capacity to disseminate.

It is a widely held belief that earlier diagnosis through screening leads to better prognosis, but this assumes that the tumours can be detected before they metastasise. In observational studies, screening generally failed to reduce the occurrence of advanced cancers, and in the randomised trials of screening for breast cancer there was bias, both in assessment of cause of death and in estimating the number of cancers with metastases, which favoured screening (Gøtzsche *et al*, 2012).

A fundamental concept in cancer screening is overdiagnosis, which is the detection of lesions that would never have been detected in a person's lifetime in the absence of screening. Cancer screening inevitably leads to overdiagnosis, because some people diagnosed with preclinical cancer will die from competing causes before the cancer would have been noticed clinically. Overdiagnosis is a major concern with screening, particularly for prostate, breast and lung cancer (Welch, 2010).

Overdiagnosis is mostly caused by the detection of slow-growing and dormant tumours. Moreover, spontaneous regression of screen-detected cancers occurs, and indeed much more commonly than previously imagined, which we demonstrated by comparing

*Correspondence: Dr P-H Zahl; E-mail: per-henrik.zahl@fhi.no

Received 22 March 2013; revised 3 July 2013; accepted 4 July 2013; published online 20 August 2013

breast cancer incidence in screened and non-screened cohorts in Norway and Sweden (Zahl *et al*, 2008, 2011). Spontaneous regression is well known also for other cancers. Neuroblastoma is an extra-cranial solid cancer in childhood, which can be detected at a preclinical stage by screening for catecholamines in the urine. A screening trial found that many cancers regressed spontaneously after the parents had refused treatment (Schilling *et al*, 2002). Another example is malignant melanoma, in which numerous cases report the existence of complete regression (Printz, 2001).

A fundamental concept in cancer screening is lead time, which is the time the diagnosis of a cancer is brought forward through screening, on average. Lead time can mean two different things: (1) clinical lead time is the lead time for clinically relevant tumours, that is, those that are not overdiagnosed. The longest possible lead time, limited only by the detection threshold of the screening technique under ideal conditions, is called the sojourn time. It can rarely, if ever, be observed in clinical practice. (2) Model-based lead time is a highly theoretical construct, which includes a hypothetical *post-mortem* scenario, where the time when the tumour would have caused symptoms is not limited by the person's death. It is the average time the diagnosis is brought forward for both clinically relevant and overdiagnosed cancers.

These relationships are shown in Figure 1. Many researchers calculate model-based lead times (Duffy *et al*, 1995; Draisma *et al*, 2003; Weedon Fekjær *et al*, 2008; Seigneurin *et al*, 2011), probably because pathologists cannot say which tumours are overdiagnosed and which are clinically relevant to detect at screening. This has caused major confusion and controversy in the scientific literature on overdiagnosis with mammography screening (Gøtzsche, 2012). The estimates vary from about 0 (Smith *et al*, 2004; Njor *et al*, 2013) to 75% (Zahl and Mæhlen, 2012).

It is often argued that overdiagnosis of breast cancer should be adjusted for model-based lead time (Seigneurin *et al*, 2011; Kalager *et al*, 2012). However, this is never done when calculating overdiagnosis of prostate cancer (Draisma *et al*, 2003). Model-based lead time includes many slow-growing, dormant (Spratt *et al*, 1993; Draisma *et al*, 2003) or regressing tumours (Zahl *et al*, 2008, 2011), and adjustment for model-based lead time will therefore inevitably 'adjust away' a large part of the overdiagnosis.

Overdiagnosis can be measured accurately in a randomised trial with lifelong follow-up if people are assigned to a screening or control group for as long as screening would be offered in practice. In the absence of overdiagnosis, the initial increase in cancers in the screened age groups would be fully compensated for by a similar decrease in cancers among older age groups, who are no longer offered screening. Overdiagnosed cancers are the absolute difference in the number of cancers detected during the lifetime of the two groups, provided the control group is not screened.

A good approximation to the ideal trial is studying any compensatory drop in the incidence of breast cancer among older, previously screened women. All extra cancers detected during screening (the incidence increase when screening multiplied by women at risk) minus the reduction of cancers during the compensatory drop (the incidence decline after screening has stopped multiplied by the women at risk) is the absolute number of overdiagnosed cancers. This number should be compared with all cancers detected during screening (Draisma *et al*, 2003), and in this way overdiagnosis can be estimated without using lead time. We used this method in a systematic review and found 52% overdiagnosis for screening mammography (Jørgensen and Gøtzsche, 2009). Alternatively, the life-time risk of overdiagnosis can be calculated by including all cancers, also all those detected after screening stopped. Most researchers add cancers detected in a 5–15 years period after screening has stopped, arguing that this is a sufficient adjustment for lead time (Smith *et al*, 2004; Seigneurin *et al*, 2011; Kalager *et al*, 2012; Njor *et al*, 2013). Here we discuss how clinical lead time can be estimated and how adjustment for clinical and model-based lead times gives fundamentally different estimates of overdiagnosis.

TUMOUR GROWTH

The two most natural units for measuring tumour growth are the number of cell divisions after the appearance of the first tumour cell and the tumour volume doubling time. Tumour volume doubling time is the same as the cell doubling time when there is

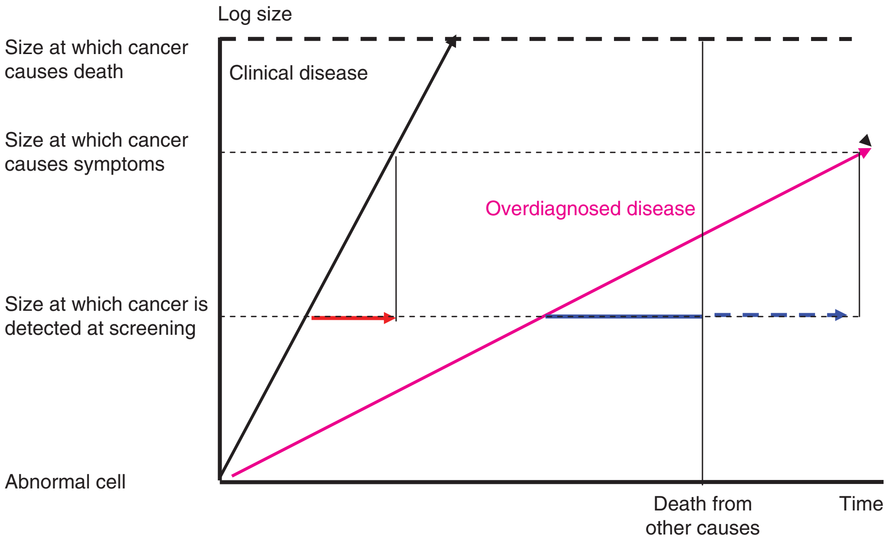
Log size
Size at which cancer causes death
Clinical disease
Size at which cancer causes symptoms
Overdiagnosed disease
Size at which cancer is detected at screening
Abnormal cell
Death from other causes
Time

Figure 1. Two different types of tumour growth indicated by the black (clinical disease) and pink arrows (overdiagnosed disease), respectively. Clinically relevant lead time is indicated by the red arrow. When overdiagnosed tumours are included, some researchers have defined lead time as time to death (solid blue arrow, so-called censored lead time) (Draisma *et al*, 2003), others have included the time to clinical appearance if there had been no competing causes of death (solid plus dotted blue arrow, so-called uncensored lead time) (Draisma *et al*, 2003; Weedon Fekjær *et al*, 2008).

no cell death, whereas tumour size (the diameter of the tumour) increases more slowly.

Exponential growth, with a constant time between each volume doubling, is the simplest model for tumour growth. After some time, the growth will slow down, because the tumour outgrows its blood supply, limiting nourishment and causing necrosis (Spratt *et al*, 1993). Other statistical models for growth, such as the Gompertz or logistic model, can be used to adjust for declining growth rates. Beyond a certain size, the host cannot sustain further growth of the primary tumour, but metastases may keep growing. The tumour volume doubling time is calculated from two or more volumes estimated by X-ray (Spratt *et al*, 1993).

Clinical data show that the frequency distribution of doubling times is skewed with some tumours having very long doubling times (Spratt *et al*, 1993; Aoki *et al*, 2000). Most of these slow-growing tumours are highly differentiated and are therefore relatively benign tumours (Aoki *et al*, 2000). Spratt *et al* (1993) studied 448 women with breast cancer and estimated that 90% of the doubling times were between 69 and 1622 days, with a median of 260 days. This was a selected group of women with mammography-detected cancers, which for various reasons were not treated.

The relationship between the number of cell doublings, volume and tumour diameter for breast cancer is presented in Table 1, assuming a spherical size of the tumour. It has been argued that dissemination starts when the tumour diameter is 1 mm, as further growth requires a dedicated blood supply (Folkman *et al*, 1989), which is also needed for dissemination.

Useful data on invasive breast tumour sizes at the time of diagnosis are available from five randomised trials (Gøtzsche *et al*, 2012). The average tumour size was 16 mm in the screened groups and 21 mm in the control groups. A cancer of 16 mm will grow up to 21 mm after 1.3 volume doubling times or after about 260 × 1.3 = 340 days (Gøtzsche *et al*, 2012). However, the window of opportunity for mammography screening to work is considerably less than 340 days, because overdiagnosed tumours are generally small and screening predominantly catches small and slow-growing tumours (length bias). This means that the average tumour size in the screened groups would have been larger than 16 mm (and clinical lead time less than 340 days), if we had been able to identify and subtract the overdiagnosed tumours.

Statistical models for estimating model-based lead time are not based on observations of tumour growth but are calculated indirectly from the observed incidence increase when screening is introduced, assuming that all increase is caused by earlier diagnosis of cases that would have become detected clinically later on. The fundamental problem with these lead time models is that they include overdiagnosed cases. Other problems are that the growth rates and the screening test sensitivity are unknown and must be assumed (Walter and Day, 1984). Sensitivity depends on tumour size, mammographic density, number of pictures and the use of digital reading, and the estimates vary grossly (Weedon Fekjær *et al*, 2008). The seminal statistical work on lead time started with a postulated lead time distribution, which gave the best fit to the observed incidence increase when screening was introduced (Zelen and Feinleib, 1969; Prorok, 1976; Walter and Day, 1984). If the rate of cancers in the first screening round is much higher than rates in later screening rounds, this means that there are many slow-growing tumours (and that the mean lead time is therefore long). Tumours detected between two screening rounds—so-called interval cancers—are those that grow so fast that they were too small to be detected at the previous screening round, or those that were simply missed because of low sensitivity.

In these models, tumours grow with varying speeds, but none are dormant or regress. Lead-time models are sometimes combined with multistage Markov models, for example, the MIcro-simulation SCreening ANalysis model (Draisma *et al*, 2003). The problems with these complicated models are that few people can understand them, and that there are additional, important assumptions (such as calibrating to stage distributions and underlying incidence increases) that cannot be verified. Duffy *et al* (1995), Weedon Fekjær *et al* (2008) and Draisma *et al* (2003) have used such models to calculate model-based lead time. However, these models cannot be used to calculate overdiagnosis when there are dormant tumours or regression.

Fryback *et al* (2006) avoided such problems, as they allowed tumours to stop growing or to regress and calibrated their model to real data on tumour growth (Spratt *et al*, 1993). They included both tumours and ductal carcinoma *in situ* in their study and noted that the best model fit was achieved when assuming that 42% of the lesions regressed.

The assumption that all tumours progress can be tested by comparing the observed incidence after screening has stopped to the expected incidence (Boer *et al*, 1994). If there is no overdiagnosis, all extra cancers detected by screening should be compensated by a corresponding decline in the number of cancers after screening has stopped. The fact that there is very little compensatory decline in incidence when women are no longer screened, in Norway as well as in other countries (Jørgensen and Gøtzsche, 2009), proves that model-based lead-time models are seriously misleading.

Our primary aim was to estimate lead time for clinically relevant tumours and to compare this with model-based estimates. The secondary purpose was to compare overdiagnosis adjusted for model-based lead time with overdiagnosis adjusted for clinically relevant tumours only.

MATERIALS AND METHODS

Calculation of clinical lead time. It is possible to estimate clinical lead time by studying how fast the incidence returns to the background incidence after a screening round. We only need to

Table 1. Number of cell doublings, cells, volume and tumour diameter for breast cancer

No. of cell doublings	No. of cells	Volume (mm^3)	Diameter (mm)	Feature relevant to size
0	1	0.000001	0.012	First malignant cell
19	524 288	0.52	1.0	Potential for metastasis (Folkman *et al*, 1989; Weedon Fekjær *et al*, 2008)
28	268 435 456	268	8.0	Mammographic detection threshold (Gøtzsche *et al*, 2012)
29	536 870 912	536	10.0	Clinical detection threshold (minimum palpable size) (Gøtzsche *et al*, 2012)
31	2 147 483 648	2148	16.0	Average size at screen detection in RCT (Gøtzsche *et al*, 2012)
32	4 294 967 296	4295	20.1	Average size at clinical detection in RCT (Gøtzsche *et al*, 2012)

Abbreviation: RCT = randomised controlled trial.

assume that the maximum clinical lead time is, for example, 4 years. This assumption is very reasonable and has been tested by studying how fast the incidence rate returns to the background level (Jørgensen and Gøtzsche, 2009; Zahl and Mæhlen, 2012).

We call the annual per cent incidence reductions compared with the background incidence in a control group pr_1, pr_2, pr_3 and pr_4, respectively. It is largest in the first year after screening, that is, $pr_1 > pr_2 > pr_3 > pr_4$. The clinical lead time (in years) is then calculated approximately as the weighted average:

$$(0.5\times pr_1+1.5\times pr_2+2.5\times pr_3+3.5\times pr_4)/S \qquad (1)$$

where $S=(pr_1+pr_2+pr_3+pr_4)$. Note that this estimate is not inflated by including overdiagnosed tumours and this is a novel method.

Calculation of model-based lead time. Next, we compare model-based lead times with clinical lead times. To simplify the comparisons, we assume that the average lead time for overdiagnosed tumours is 10 years (scenario 1) or 25 years (scenario 2). Such long lead times are expected for overdiagnosed tumours, because the compensatory decline in incidence when women are no longer screened owing to advanced age is very small or absent (Jørgensen and Gøtzsche, 2009; Zahl and Mæhlen, 2012). We assume the average clinical lead time is 1 year (which we calculate in the results by using equation 1 above), and for each scenario we vary the proportion of overdiagnosis from 10 to 70%. Model-based lead time (T_M) is calculated as the weighted average of clinical lead time (T_C) and lead time for overdiagnosed cases (T_O):

$$T_M = T_C\times(1-p)+T_O\times p \qquad (2)$$

where p is the proportion of overdiagnosed cases out of all diagnosed cases. Note that this is different from the way we usually describe overdiagnosis. When we talk about 50% overdiagnosis, we mean that overdiagnosis adds 50% more diagnoses to those that would have been found if screening had not been performed (Jørgensen and Gøtzsche, 2009). In equation 2, we start from the total number of diagnoses, which means that for 50% overdiagnosis, p is 0.33 (as 50% of 150% is 0.33). Furthermore, note that we assume all women are screened.

Diluted estimates of overdiagnosis introduced by long-term follow-up after screening is stopped. The Malmö randomised mammography screening trial reported 10% overdiagnosis after 15 years of follow-up, after the initial 10 years randomised phase (Zackrisson *et al*, 2006). Kalager *et al* (2012) reported 15–25% overdiagnosis in the Norwegian Breast Cancer Screening Program, with a 10-year follow-up period. We shall therefore study by how much the estimated level of overdiagnosis is diluted, by adding cancers diagnosed after the end of a screening programme to both the screened and the non-screened cohort. We pretend that the Norwegian breast cancer data from 1991 is the unscreened cohort (data from the Norwegian Cancer Registry). There were 152, 157, 200, 221 and 245 breast cancers in the age groups 50–54, 55–59, 60–64, 65–69 and 70–74 years. We made a screening cohort with a 55% incidence increase during screening from 50 to 59 years, followed by a reduction of 15 cancers in the age group 60–64 years so that there is 50% overdiagnosis. Then we calculated the cumulative rates and the cumulative incidence rate ratio for the screening and the control cohort during 15 years of follow-up. We also studied what happened when both groups were screened after age 59 years, because this is also an adjustment for lead time (Zahl *et al*, 2008, 2011).

RESULTS

In Norway, the incidence reduction in screened women aged 50–69 years (compared with the expected incidence rate for unscreened women) was 70% (compared with the previous year) in the first year after a mammogram and 30% in the second year, and the size of these reductions were not related to the number of times the women had been screened previously (Zahl and Mæhlen, 2012). After the last screening round at age 68–69 years, the incidence rate returned to the expected incidence level over a 5-year period (Zahl and Mæhlen, 2012). Thus, it is reasonable to assume that the reductions in the third and fourth year are 10 and 5%, respectively (and maximum lead time is 4 years). Using equation 1, we calculate that the clinical lead time is 1.06 years. Including 5% tumours with 5 years clinical lead time increased the estimated clinical lead time by 0.14 years; assuming a 1% annual underlying incidence increase added only 0.01 year to the estimate; assuming a 50% higher incidence reduction after screening increased the estimate by 0.06 year, and combining all three extreme assumptions increased the estimate by 0.18 year.

The model-based lead times are much longer. In Table 2, we have calculated the model-based lead times for different levels of overdiagnosis (10–70%), using 10 and 25 years lead times for overdiagnosed tumours, and 1 year for clinically relevant tumours, which is realistic (see above). An overdiagnosed mammography detected tumour at age 60 years must have a minimum expected lead time of 25 years, because this is the expected remaining life time for these women. As expected, the bias inherent in calculating model-based lead time increases with the level of overdiagnosis, and even for moderate levels of overdiagnosis the bias is huge. For an overdiagnosis of 50%, for example, the model-based lead time is 4 and 9 years when the clinical lead time is 1 year.

The dilution effect is studied in Figure 2. If we screen for 10 years and add new cancers to both groups during another 10 years of follow-up, adjustment for model-based lead time gives 20% overdiagnosis, whereas adjustment for clinical lead time gives 50% overdiagnosis. Note that when both groups are screened after age 59 years, we get a similar curve; however, this is a test of how many cancers that would go in regression during a 10-year period and not a long-time risk of overdiagnosis (in this case there is 100% more overdiagnosed cancers than in the other scenario).

DISCUSSION

It is a widely held misunderstanding that cancer screening can move time of diagnosis 5–10 years. The clinical lead time we estimated by using the reductions in breast cancer incidence after a

Table 2. Calculated model-based lead time for a combination of clinical tumours (all with lead time of 1 year) and overdiagnosed tumours

	Overdiagnosis	Model-based lead time
Scenario 1	10%	1.8
Lead time for overdiagnosed	30%	3.1
tumours is 10 years	50%	4.0
	70%	4.7
Scenario 2	10%	3.2
Lead time for overdiagnosed	30%	6.5
tumours is 25 years	50%	9.0
	70%	10.9

The level of overdiagnosis varies from 10 to 70% and the lead times for overdiagnosed tumours are 10 and 25 years, respectively.

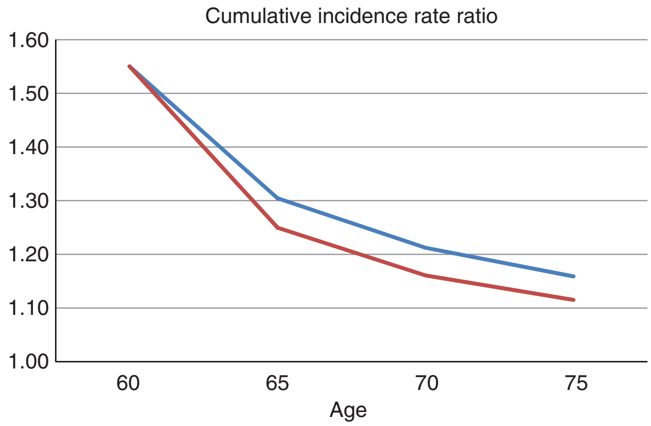

Figure 2. Here the incidence increase when screening one cohort from age 50–59 years is 55%, and overdiagnosis adjusted for clinical lead time it is 50%. The blue curve is the incidence rate ratio for this cohort followed up in 15 more years (up to age 75 years) with no screening compared with a control cohort that is not screened at all. The incidence rate ratio declines as more and more cancers diagnosed after age 60 years are included in the calculations. An incidence rate ratio of 1.1 is interpreted as 10% overdiagnosis adjusted for model-based lead time. The red curve is the same example, but now both cohorts are screened from age 60–75 years – this is also adjustment for lead time. Curves are similar, but the interpretations are totally different. The blue curve is tending to the life-time risk of being overdiagnosed. The red curve is a test of cancer regression. Furthermore, screening from age 40 to 59 years gives different curves (not presented here). When the incidence rate ratio depends on both how long you are screened and how long you follow up individuals after screening has ended, this method generates results that are not generally comparable.

screening round is about 1 year. This is similar to the estimate we derived from the observed tumour diameters in the screened groups and in the control groups in the randomised trials (Gøtzsche *et al*, 2012). A clinical lead time of 1 year also corresponds to a little more than one volume doubling time, as reported by Spratt *et al* (1993).

The model-based lead time is many times larger than the clinical lead time (Table 2) and is excessively inflated by including overdiagnosed cases. Model-based lead time estimates in the literature vary from 2 to 7 years for screen-detected breast cancer (Duffy *et al*, 1995; Weedon Fekjær *et al*, 2008) and from 3 to 12 years for screen-detected prostate cancer (Draisma *et al*, 2003, 2009). Our study shows that the theory and the assumption that you need 10–15 years of follow-up to adjust for long lead times when calculating overdiagnosis of breast cancer is flawed. It is enough with 5 years follow-up.

Another way to illustrate that 5 years are enough is to look at the interval cancer rate. The cancer rate in the second year after a screening is 140 per 100 000 in Norway, which is approximately the same as the cancer rate before mammography screening started (Zahl and Mæhlen, 2012). This has also been demonstrated by Vaidya (2004) who observed that the cancer rate in Scotland in the second and third year after a screening was similar to the rate before screening started.

The validity of the lead time models can also be tested by studying whether constant high incidence rates during screening can be explained by the detection of many tumours with long lead times. First, we can study whether attending 3 screening rounds leads to more breast cancer cases compared with attending only the last screening round (Zahl *et al*, 2008, 2011). Second, we can study whether the incidence increase during 20 or 30 years with mammography screening would have accumulated and would have been detectable at a prevalence screening at age 70 years, and this has actually been done by Smith-Bindman *et al* (2003). Third, we can study whether there is a compensatory decline in the incidence rate when women are no longer invited to screening (Zahl *et al*, 2004; Jørgensen and Gøtzsche, 2009; Zahl and Mæhlen, 2012). All these 3 analyses invalidate the theory that there is a large reservoir of breast cancers with long lead times. Overdiagnosis of breast cancer should primarily be explained by the detection of tumours that normally would have regressed spontaneously. Finally, if model-based estimates of lead time were correct, then it should prevent many metastases; however, screening does not prevent the occurrence of metastases (Esserman *et al*, 2009).

Calculating overdiagnosis by adding cancers diagnosed many years after screening has stopped results in substantial underestimation as shown in Figure 2. This method defines overdiagnosis as a function tending to the life-time risk. Estimates depend on both the duration of screening and the follow-up period, and are therefore not generally comparable. The dilution method has been used by several research groups (Zackrisson *et al*, 2006; Seigneurin *et al*, 2011; Kalager *et al*, 2012; Marmot *et al*, 2012), and when diluted estimates are compared with undiluted estimates, it causes a major confusion (Welch, 2006; Marmot *et al*, 2012). For example, Kalager *et al* (2012) estimated 25% overdiagnosis, whereas Zahl *et al* (2004) and Zahl and Mæhlen (2012) estimated 51% overdiagnosis (and 75% when premalignant lesions were included). However, the absolute numbers of overdiagnoses were 87 and 93 per 100 000, respectively. If both groups are screened simultaneously after the end of the screening period (this is also adjustment for lead time) and curves never meet, it is a formal test of cancer regression (Zahl *et al*, 2008, 2011). It is not the life-time risk of overdiagnosis as some argue (Seigneurin *et al*, 2011; Marmot *et al*, 2012).

We searched PubMed with the terms ‘breast cancer’, ‘mammography’, ‘screening’ and ‘overdiagnosis’, and found 115 papers published after 2002 (date of search 5 February 2013). Forty-five letters and 21 papers were considered not relevant. We studied how papers reviewed the literature on overdiagnosis, and found 7 papers pointing out that differences in the percentages of overdiagnosis were mostly due to differences in which denominators were used in the calculations. But we also found 22 papers comparing percentages that are not comparable without caveats. We found no papers commenting that overdiagnosis adjusted for model-based lead time depends on both the duration of screening and the duration of follow-up after screening, and only one paper actually estimating the life-time risk (Falk *et al*, 2013). The duration of screening varied from 5 to 20 years and the duration of follow-up varied from 5 to 30 years. All observational studies were estimating overdiagnosis adjusted for annual underlying incidence increases. Bleyer and Welch (2012) reported the highest level of overdiagnosis: 114 overdiagnoses per 100 000 women over age 40 years. The corresponding number in Norway is 68 per 100 000 (Zahl *et al*, 2004; Zahl and Mæhlen, 2012). The annual underlying incidence increase was about 0.5% in Norway and in the United States. Papers reporting lower rates of overdiagnosis did not report the underlying incidence increase.

For prostate cancer, the cumulative number of cancers in the US randomised trial was 2541 (PSA screened group) and 1997 (control group) after 6 years with annual screening (Andriole *et al*, 2012). In the next 7 years of follow-up (without screening), the initial difference of 544 cancers declined to 435; that is, about 80% (435 out of 544) of the observed incidence increase appeared to be overdiagnosis and not earlier diagnosis as assumed when calculating model-based lead time (Draisma *et al*, 2009). Furthermore, there was no spike in the prevalence screening round. When there is no prevalence peak, then almost all clinical lead times are 1 year or less (Zelen and Feinleib, 1969; Prorok, 1976; Walter and Day, 1984). It should also be noted that as there was much opportunistic screening in the control group, the initial difference in number of tumours was underestimated.

Estimates of overdiagnosis when screening with PSA vary from 23 to 62% (Draisma *et al*, 2009). Overdiagnosis when screening with mammography is calculated differently – women not attending screening are usually also included in the calculations – and the percentages vary from 52% (Jørgensen and Gøtzsche, 2009) to 75% (Zahl and Mæhlen, 2012). It is higher if you only include those who actually attend regular screening (Welch, 2006). Thus, assuming 70% overdiagnosis in Table 2 is not an unrealistic estimate, neither for PSA screening nor for mammography screening. Note that the estimates used in Table 2 must not be compared with life-time risks.

We believe it is misleading to calculate life-time risk of overdiagnosis, or adding cancers to both groups many years after screening stopped. This method also draws attention from the fact that many more women are overdiagnosed at an earlier time, and therefore also ignores the harm caused by turning healthy women into cancer patients relatively early on in their life span.

REFERENCES

Andriole GL, Crawford ED, Grubb III RL, 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 (2012) Prostate cancer screening in the randomized prostate, lung, colorectal and ovarian cancer screening trial: mortality results after 13 years follow-up. *JNCI* **104**: 125–132.

Aoki T, Nakata H, Watanabe H, Nakamura K, Kasai T, Hashimoto H, Yasumoto K, Kido M (2000) Evolution of peripheral lung adenocarcinomas. CT findings correlated with histology and tumor doubling time. *Am J Roentgenol* **74**: 763–876.

Bleyer A, Welch GH (2012) Effect of three decades of screening mammography on breast-cancer incidence. *N Engl J Med* **367**: 1998–2005.

Boer R, Mamerdam P, de Koning H, van Oortmarssen G (1994) Extra incidence caused by mammographic screening. *Lancet* **343**: 979.

Draisma G, Boer R, Otto SJ, van der Cruijsen I, Damhuis RAM, Schröder FH, de Koning HJ (2003) Lead time and overdetection due to prostate-specific antigen screening: estimates from the European randomized study on screening for breast cancer. *JNCI* **95**: 868–878.

Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Guliati R, Feuer E, de Koning H (2009) Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. *JNCI* **101**: 374–383.

Duffy SW, Hill C, Chen H-H, Tabar L, Day NE (1995) Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Stat Med* **14**: 1531–1543.

Esserman L, Shieh Y, Thompson I (2009) Rethinking screening for breast cancer and prostate cancer. *JAMA* **302**: 1685–1692.

Falk RS, Hofvind R, Skaane S, Haldorsen T (2013) Overdiagnosis among women attending a population-based mammography screening program. *Int J Cancer* **133**(3): 705–712.

Folkman J, Watson K, Ingber D, Hanahan D (1989) Induction of angiogenesis during the transition from hyperplasia to neoplasia. *Nature* **339**: 58–61.

Fryback DG, Stout NK, Rosenberg MA, Trentham-Dietz A, Kuruchittham V, Remington PL (2006) The Wisconsin breast cancer epidemiology simulation model. *J Natl Cancer Inst Monogr* **36**: 37–47.

Gøtzsche PC. (2012) *Mammography Screening: Truth, Lies and Controversy*. Radcliffe Publishing: London.

Gøtzsche PC, Jørgensen KJ, Zahl P-H, Mæhlen J (2012) Why mammography screening has not lived up to expectations from the randomised trials. *Cancer Causes Control* **23**: 15–21.

Jørgensen KJ, Gøtzsche PC (2009) Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *BMJ* **339**: b2587.

Kalager M, Adami H-O, Bretthauer M, Tamimi RM (2012) Overdiagnosis of invasive breast cancer due to mammography screening: results from the Norwegian Screening program. *Ann Intern Med* **156**: 491–499.

Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. Independent UK Panel on Breast Cancer Screening (2012) The benefits and harms of breast cancer screening: an independent review. *Lancet* **380**: 1778–1786.

Njor SH, Olsen AH, Blichert-Toft M, Schwartz W, Vejborg I, Lynge E (2013) Overdiagnosis in screening mammography in Denmark: population based cohort study. *BMJ* **346**: f1064.

Printz C (2001) Spontaneous regression of melanoma may offer insight into cancer immunology. *JNCI* 1047–1048.

Prorok PC (1976) The theory of periodic screening I: lead time and proportion detected. *Adv Appl Probability* **8**: 127–143.

Schilling FH, Spix C, Berthold F, Erttmann R, Fehse N, Hero B, Klein G, Sander J, Schwartz K, Treuner J, Zorn U, Michaelis J (2002) Neuroblastoma screening at one year of age. *N Engl J Med* **346**: 1047–1053.

Seigneurin A, Francois O, Labarere J, Oudeville P, Monlong J, Colonna M (2011) Overdiagnosis from non-progressive cancer detected by screening mammography: stochastic simulation study with calibration to population based registry data. *BMJ* **343**: d7017.

Smith RA, Duffy SW, Gabe R, Tabar L, Yen AM, Chen TH (2004) The randomized trials of breast cancer screening: what have we learned? *Radiol Clin North Am* **42**: 793–806.

Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash R, Bobo JK, Lee NC, Wallis MG, Patnick J, Kerlekowske K (2003) Comparison of screening mammography in the United States and the United Kingdom. *JAMA* **290**: 2129–2137.

Spratt JA, von Fournier D, Spratt JS, Weber EE (1993) Decelerating growth and human breast cancer. *Cancer* **71**: 2013–2019.

Vaidya JS (2004) Women undergoing screening mammography experience a higher incidence of invasive breast cancer, without a corresponding reduction in symptomatic breast cancer. *BMJ*. Available at http://www.bmj.com/content/339/bmj.b2587?tab=responses (accessed 20 March 2012).

Walter SD, Day NE (1984) Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. *Biometrics* **40**: 1–14.

Weedon Fekjær H, Lindquist BH, Allen OO, Vatten LJ, Tretli S (2008) Estimating sojourn time and screening sensitivity using questionnaire data on time since previous screening. *J Med Screen* **15**: 83–90.

Welch HG (2006) Ramifications of screening for breast cancer: 1 in 4 cancers detected at mammography screening are pseudocancers. *BMJ*. Available at http://www.bmj.com/content/332/7543/689?tab=responses (accessed 20 March 2012).

Welch HG (2010) Overdiagnosis in cancer. *JNCI* **102**: 605–613.

Zackrisson S, Andersson I, Janzon L, Manjer J, Garne JP (2006) Rate of overdiagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study. *BMJ* **332**: 689–691.

Zahl P-H, Gøtzsche PC, Mæhlen J (2011) Spontaneous regression of screening-detected cancers in the Swedish breast screening program. *Lancet Oncol* **12**: 1118–1124.

Zahl P-H, Mæhlen J (2012) [Overdiagnosis of breast cancer after 14 years with screening]. *Tidsskr Nor Lægeforen* **132**: 414–417.

Zahl P-H, Mæhlen J, Welch HG (2008) The natural history of invasive breast cancers detected by screening mammography. *Arch Intern Med* **168**: 2311–2316.

Zahl P-H, Strand BH, Mæhlen J (2004) Breast cancer incidence in Norway and Sweden during introduction of nation-wide screening: prospective cohort study. *BMJ* **328**: 921–924.

Zelen M, Feinleib M (1969) On the theory of screening for chronic diseases. *Biometrika* **56**: 601–613.