BMJ Open

Estimations of overdiagnosis in breast cancer screening vary between 0% and over 50%: why?

Dan Chaltiel [1,2] Catherine Hill [1]

To cite: Chaltiel D, Hill C. Estimations of overdiagnosis in breast cancer screening vary between 0% and over 50%: why? *BMJ Open* 2021;**11**:e046353. doi:10.1136/bmjopen-2020-046353

► Prepublication history and additional online supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-046353).

Received 27 October 2020
Accepted 10 May 2021

[1]Bureau de Biostatistique et d'Épidémiologie, Gustave Roussy, Université Paris-Saclay, Villejuif, France
[2]Oncostat U1018, Inserm, Équipe Labellisée Ligue Contre le Cancer, Université Paris-Saclay, Villejuif, France

Correspondence to
Dr Dan Chaltiel;
dan.chaltiel@gustaveroussy.fr

ABSTRACT

Background Published estimations of the extent of breast cancer overdiagnosis vary widely, and there have been heated debates around these estimations. Some high estimates have even been the basis of campaigns against national breast cancer screening programmes. Identifying some of the sources of heterogeneity between different estimates would help to clarify the issue.

Methods The simple case of neuroblastoma—a childhood cancer—screening is used to describe the basic principle of overdiagnosis estimation. The more complicated mechanism of breast cancer overdiagnosis is described based on data from Denmark, taking into account the type of data used, individual or aggregated.

Findings The type of data used in overdiagnosis studies has a meaningful effect on the estimation: no study based on individual data provides an estimate higher than 17%, while studies based on aggregated data often provide estimates higher than 40%. This is too systematic to be random. The analysis of two Danish studies, one of each kind, highlights the biases that come with the use of aggregated data and shows how they can lead to overdiagnosis.

Interpretation Many estimates of overdiagnosis associated with breast cancer screening programmes are serious overestimations.

INTRODUCTION

Many countries have a national breast cancer screening programme in which all women belonging to a specific age group are invited to have regular mammograms. These programmes have been criticised, with claims that their benefit has been overestimated and that the risk of overdiagnosis has been understated. Here, overdiagnosis is defined as the diagnosis, by a screening procedure, of a cancer that would never have become symptomatic during the life of the person.

Both in situ and invasive cancers will be included in the estimation of overdiagnosis, since an overdiagnosed in situ breast cancer leads to an unnecessary treatment, which can include a mastectomy, a reconstructive surgery and a cosmetic surgery on the other breast to restore symmetry.

The estimations of overdiagnosis in breast cancer screening vary between 0% and more than 50% (figure 1), and the variety of these estimations contributes to the vigorous debate on the usefulness of breast cancer screening programmes.[1] Since it is extremely unlikely that overdiagnosis varies to such a large extent from one programme to another, one needs to study possible causes for this observed heterogeneity.

ESTIMATION METHODS

The ideal approach to estimate the overdiagnosis rate would be to use data from randomised controlled trials on breast cancer screening in which the participants in the control group were not offered screening at the end of the trial. Using data from trials does not come without bias if the post screening follow-up is not long enough. The methodology of estimation itself can also be controversial, as different CI calculations could under or overestimate the uncertainty.[2] The only such trials are the two Canada trials and part of the Malmö trial and the performance of the Canada trials has been questioned.[3] Thus, we have to rely on observational studies, among which the best option is a cohort study with individual patient data.

SCREENING FOR NEUROBLASTOMA

We shall start by introducing some basic concepts about screening diagnosis, using the example of the screening for neuroblastoma, a paediatric cancer of neuroblasts (specialised nerve cells). The screening test is a measurement of urinary catecholamines, which are hormones produced by neuroblastoma cells. A study conducted in Germany compared the incidence of neuroblastoma in regions without screening and in experimental regions where screening of 1-year-old children was systematically offered.[4]

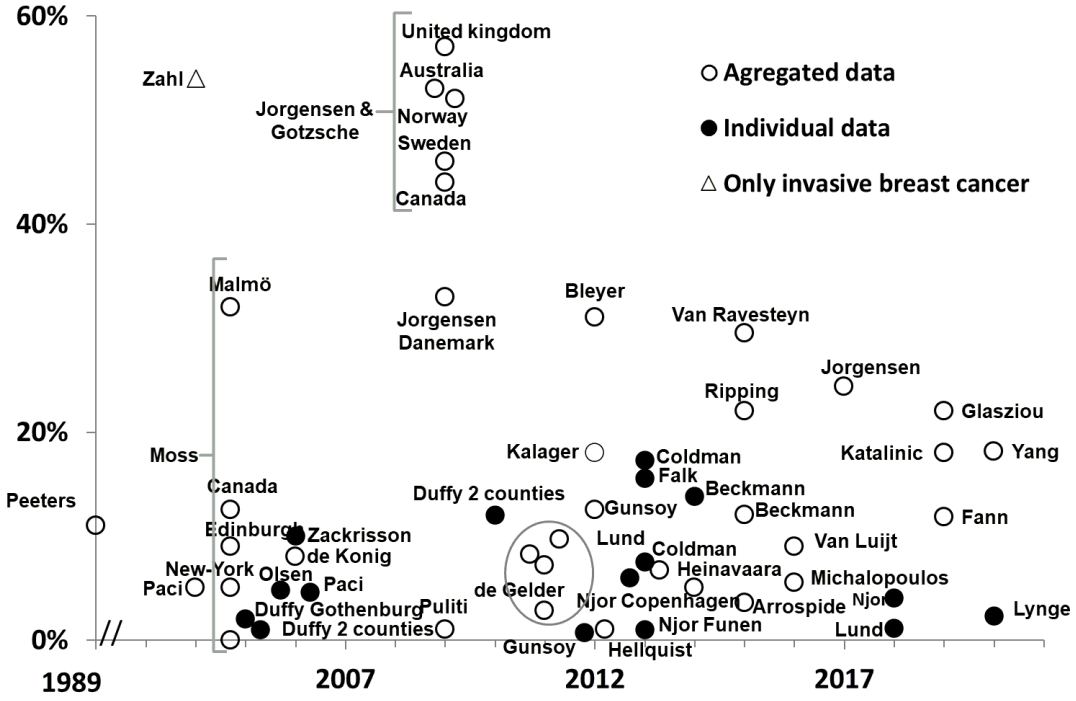

Figure 1 Published estimations of in situ and invasive breast cancer overdiagnosis (open symbols: two publications studying only invasive breast cancers quoted in the present text). Studies conducted on aggregated data give generally higher estimations of overdiagnosis than studies conducted on individual data. Source: Ripping *et al*[1] updated by Hill. A comprehensive list of these studies is provided in online supplemental data.

Such a screening programme causes an increased incidence of cases immediately after screening (age 1 for neuroblastoma), a decrease shortly afterwards, and a return to normal thereafter (around age 5 for neuroblastoma) (figure 2A).[5] In theory, the screening programme should allow the detection of the same number of cases, only earlier (figure 2B). Therefore, if there is no overdiagnosis, the number of cases additionally diagnosed during screening (solid green) is equal to the number of cases that would have been diagnosed later, if there was no screening. Thus, overdiagnosis is measured by the difference between these two numbers (figure 2C). In the German study, there were 7·3 and 14·2 cases per 100 000 children, respectively, in the control and experimental regions (figure 2D). Overdiagnosis is the difference between these cumulative incidences, generally expressed as a percentage. Here, it represented 49% ((14·2–7·3)/14·2) of the cases found in the population invited to screening.

This simple example shows the importance of the follow-up duration in correctly estimating the amount of overdiagnosis. In the most extreme case, one would compare the incidences observed at 1 year of age only, which would then attribute overdiagnosis to all cases with a diagnosis brought forward by screening. Figure 2B shows that the incidence of neuroblastoma at age 5 and over is again the same in the two populations, which is why overdiagnosis has been estimated by comparing the cumulative incidence with and without screening between 12 and 60 months of age (figure 2D, based on reference[4] Schilling *et al*).

This study showed that screening for neuroblastoma at 1 year of age identified many cases that would have regressed spontaneously. In the end, almost half of the diagnoses were unnecessary and and detrimental to the child and his/her family; therefore, this screening is no longer offered.

BREAST CANCER SCREENING: EXAMPLE OF THE FUNEN DATA

The estimation of overdiagnosis

To evaluate the amount of screening-induced overdiagnosis in breast cancer, we shall use data from Denmark, as studied by Njor *et al*.[6] The data used were individual data, that is, for each woman, her date of birth, history of mammography, and, where applicable, dates of breast cancer diagnosis and death.

This type of screening is a very different situation: in breast cancer screening programmes, the same woman may be invited several times, at different ages, whereas children in the neuroblastoma study were all screened only once at 12 months old. Thus, while age was sufficient to evaluate overdiagnosis in neuroblastoma, one needs to take both age and calendar time into account to understand overdiagnosis in breast cancer, which adds a layer of complexity. This breast cancer study measured overdiagnosis by comparing the incidence of breast cancer in several places in Denmark (Funen Island, where there

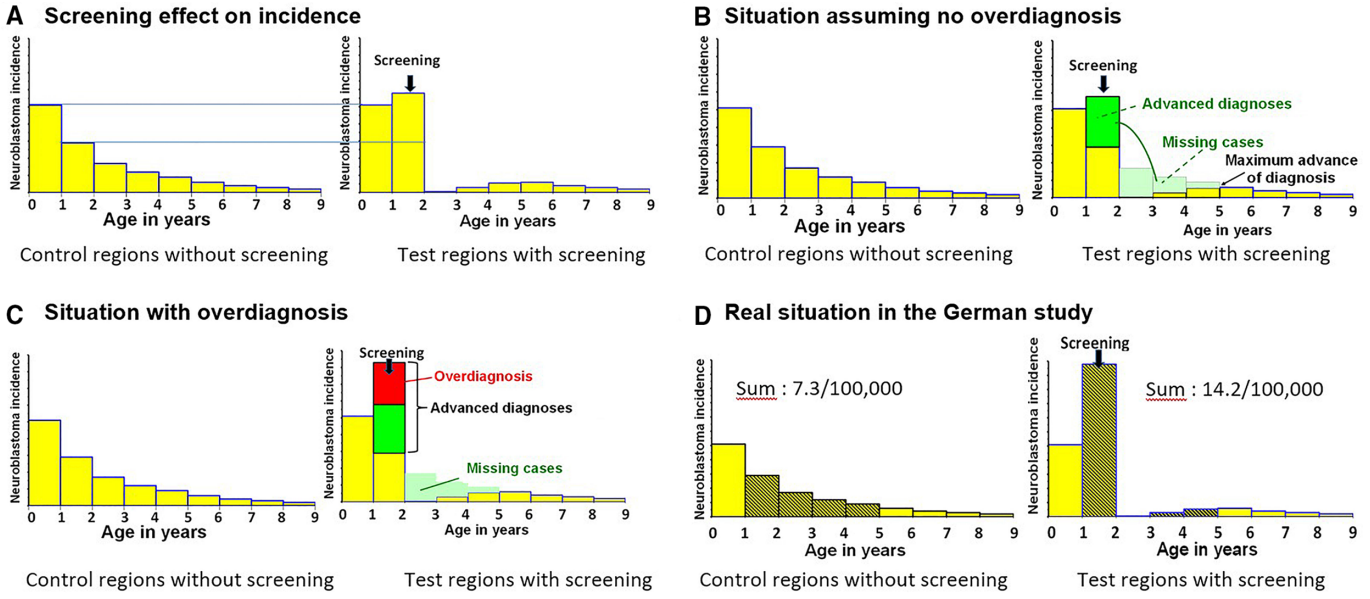

Figure 2 Overdiagnosis estimation, example of screening for neuroblastoma in Germany. based on Schilling *et al* and Spix *et al*.[4 5] Control and test regions have a comparable population size, with 1.1 and 1.5 million children, respectively. incidence is expressed in arbitrary units. (A) Incidence is displayed as a function of age, and generalised neuroblastoma screening takes place at 1 year of age. There is logically no difference in incidence between control and test regions before screening age (<1 year old). The screening programme causes an increased incidence of cases immediately after screening at age 1, a decrease shortly afterwards, and a return to normal at around age 5. (B) If there is no overdiagnosis, the number of cases additionally diagnosed during screening (solid green) should be equal to the sum of the number of missing cases, which would have been diagnosed later if there had been no screening (faded green). (C) In the case of overdiagnosis, screening reveals an additional number of cases that would never have been clinically important enough to be diagnosed otherwise (red). (D) The actual difference between the regions with and without screening was estimated to be 6.9/100 000, which translates to an overdiagnosis of 49%. According to this estimation, around half of neuroblastoma diagnosed during screening would have regressed spontaneously or would, at least, never have become clinical enough to be diagnosed, leading to unnecessary and potentially invasive treatment.

was a screening programme, vs other regions, where there was not) and during several periods (at the time of the screening programme vs beforehand).

To describe the screening experience of a population over time, a Lexis diagram is often used. An example is presented in figure 3A: the horizontal axis represents the

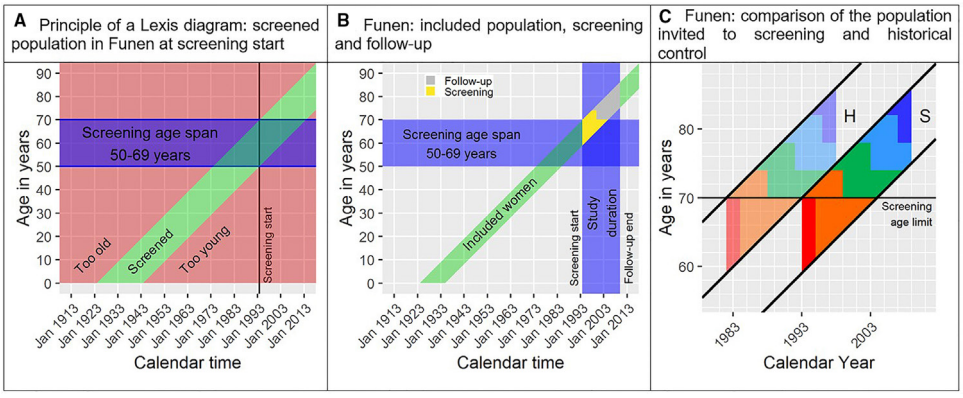

Figure 3 Lexis diagrams of the Funen overdiagnosis experiment, based on Njor *et al*.[6] Generations can be followed on diagonals. (A) Only women born between 11 January 1923 and 11 January 1943, who were 50–69 at the start of screening (11 January 1993), were invited to screening. (B) In order to have sufficient follow-up time (follow-up ended on 31 December 2009), screened women born after 1933 were not included in the study. The screening area is shown in yellow and the follow-up area in grey. In the second and third rounds (1993–1999), women were invited again, even if they were over 70; hence, the extra upper trapezoid in the 'screening' area. (C) When following the screened population (S), several periods can be identified: first screenings (red), later screenings (orange), and three follow-up periods: 0–3 years (green), 4–7 years (light blue), and ≥8 years (dark blue) from the end of invitation to screening. The comparison between the screened (S) and the historical control population (H) is performed within each period.

calendar time, and the vertical axis represents the age of the person. Thus, the trajectory of a given woman is a diagonal, starting at age 0 on her date of birth. A generation can therefore be represented by a parallelogram. In Funen Island, the screening programme started on 1 November 1993, and the whole female population aged 50–69 was invited.

Each screening round lasted 2 years; therefore, the first round spanned from 1 November 1993 to 31 October 1995. During the first three rounds, women were invited again, even if they were over age 70. Figure 3B shows the study inclusion design on a Lexis diagram. The study followed all patients from screening start until 31 December 2009 at the latest. Therefore, in order to have sufficient follow-up time, Njor *et al* included only patients aged 59–70 on 1 November 1993, as younger patients would not have been followed for long enough.[6] In the figure, the intersection of the 'study duration' area, the 'screening age span' area and the 'included women' area identifies the screened population during screening (yellow) and during follow-up (grey).

Since Funen was not the experimental arm of a randomised trial, there was no obvious control population allowing direct estimation of overdiagnosis. Thus, to evaluate the extent of overdiagnosis, one needs to estimate the incidence expected in Funen without screening.

Two types of potential control populations can be considered: (1) the population of a region without screening at the time when screening was offered in the experimental region, allowing a comparison between 'here with screening' and 'elsewhere without screening' and (2) the population in the experimental region before screening, allowing a comparison between 'before without screening' and 'after with screening'.

In the study of Funen, the control data available were data from Danish regions without screening at the time of screening in Funen (generation 1 November 1923–31 October 1934), data from Funen before screening (generation 1 November 1912–31 October 1923), and data from Danish regions without screening before the introduction of screening in Funen (generation 1 November 1912–31 October 1923).

Figure 3C is the Lexis diagram of the study period for women aged 60 and higher, representing a comparison of the studied screened population (S, generation 1923–1934) to the local historical control population (H, generation 1912–1923). Njor *et al* identified five periods of observation in the screened population: the first screening round (prevalence screening), the later screening rounds (incidence screening), which included women aged 70+ for the first three rounds and three periods corresponding to follow-up 0–3 years, 4–7 years and 8+ years from the end of invitation to screening, respectively.[6] By comparing each period of observation to its historical situation, it is possible to estimate the number of cases that would have been diagnosed in Funen if there was no screening programme. However, this is still only half of the solution, as it would not take into account the effect of geography.

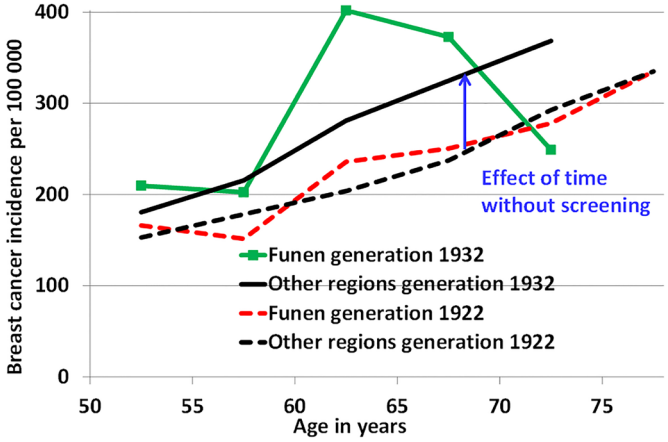

Figure 4 Incidence of breast cancer as a function of age in Funen with screening (green), compared with a historical control group (Funen in a different period, (red), to a national control group (other regions, same period, solid black line), and to a historical national control group (other regions, different period, dashed black line). Each dot represents a 5-year age group (eg, a dot between 50 and 55 represents the age group 50–54). Adapted from Njor *et al*.[8]

Simplified presentation of overdiagnosis estimation in Funen

To understand the estimation of the breast cancer incidence that would be expected if screening did not occur in Funen at the time of screening, let's focus on two 1-year generations: (1) women born in 1922 who were 71 on 11 January 1993 and, hence, were never invited to screening; and (2) women born in 1932 who were 61 on 11 January 1993 and, hence, were invited to screening.

Figure 4 shows the incidence as a function of age in these two generations, in Funen vs in other regions. Before screening (1922 generation), the incidence was rather similar in Funen (dashed red line) and in other regions (dashed black line), the data being more erratic in Funen due to its population being eight times lower than in the other regions. In the other regions, where there was no screening, the breast cancer incidence increased at all ages between the 1922 generation (black dashed line) and the 1932 generation (black solid line). This can be explained by the improvement in imaging and diagnostic techniques, among other things, during these 10 years.

Therefore, a simple estimation of the incidence that would be expected in Funen if there was no screening in the 1932 generation can be obtained by applying this estimation of the effect of time to the incidence observed in Funen in the 1922 generation. In practice, this is done by increasing the incidence observed in Funen in the 1922 generation, or a smoothed version of it, by the linear increase observed in the other regions. This is a partial view of the data from Denmark, which is shown here just to illustrate the principle of the method.

Table 1 Breast cancer incidence per 100 000 observed in Funen for every generation, and model estimations in the case of no screening programme

Period	Before screening	Invitation to screening		Follow-up postscreening			Cumulative over generations
		First screen	Further screens	0–3 years after	4–7 years after	8+ years after	
Observed	260	659	402	260	340	453	392
Expected*	260	358	352	388	411	462	387
RR	1.00	1.84	1.14	0.66	0.82	0.97	1.01

Adapted from Njor *et al*.
*The expected case number is calculated from model estimations, which take into account screening invitation (yes/no), period (before/after screening), region (other/Funen) and generations, along with interactions between periods and generations.
RR, Relative Risk.

Overdiagnosis estimation in Funen

The analysis by Njor *et al* is actually more complete and relies on a mathematical model including screening invitation (yes/no), period (before/after screening), region (other/Funen) and generations, along with interactions between periods and generations.[6]

As described in table 1 and the corresponding figure 5, this model allows an estimation of the incidence of breast cancer in Funen in the case where there was no screening programme, taking into account all the above-mentioned factors. It is then possible to compare the incidence of breast cancer in both populations, separately in each generation. By analogy, this is the equivalent of figure 2D, for which it was possible to place the age directly on the x-axis, as the screening was performed at the same age for everyone.

This leads to an estimation of overdiagnosis of 1%, based on the data observed in Funen, addressing possible differences in the incidence between periods, between Funen and the other regions, and between generations, as well as possible interactions.

In their article, Njor *et al* also present a 5% estimation based on the data observed in Copenhagen, where screening started on January 1 1991, and concluded with a global estimation of 4% overdiagnosis, based on all the data available in Denmark.[6]

ANALYSIS OF AGGREGATED DATA

The data presented by Njor *et al* are individual data, allowing the follow-up of each woman, invited to screening or not, residing in Funen or in another region, including the relevant dates (of birth, of screening invitation, of actual screening, of diagnosis and of death).[6]

However, a large number of overdiagnosis estimations rely on aggregated data. These aggregated data are incidences observed by periods and by age groups, which are publicly available for breast cancer in many countries, hence the popularity of their analysis.

To understand the difference between aggregated and individual data, the Lexis diagram is again useful. Jørgensen *et al* estimated breast cancer overdiagnosis in Denmark using aggregated data from two periods, 1971–1990 (without screening) and 1991–2003 (with screening) in two age groups: 50–69 and 70–79.[7] Therefore, these four populations are represented by four rectangles in

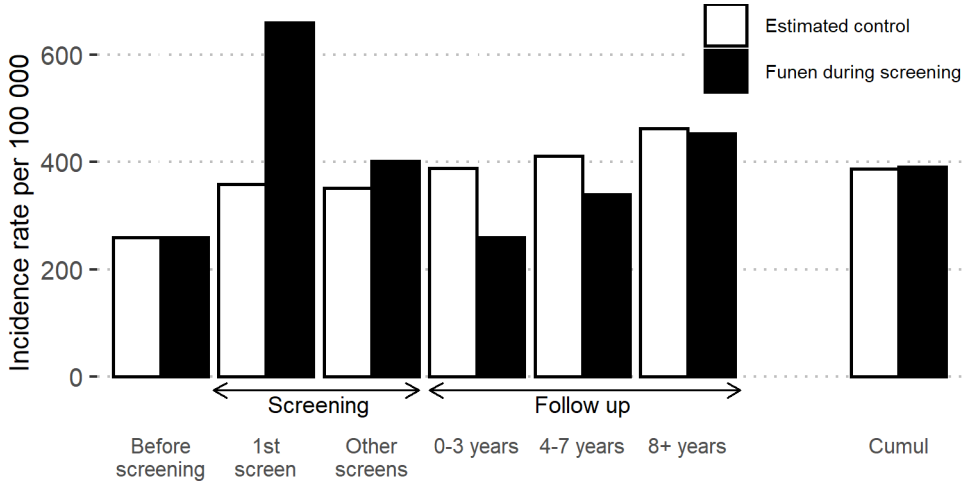

Figure 5 Incidence of breast cancer in Funen, compared with control. The incidence of breast cancer observed in Funen during screening (in black) is compared in each period to the incidence in Funen estimated by the model in the absence of screening (in grey).

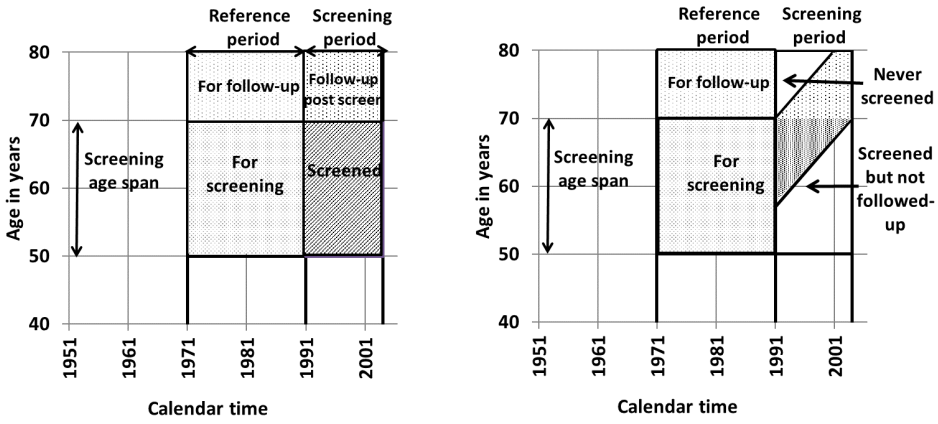

Figure 6 Data analysed by Jørgensen *et al* to estimate overdiagnosis in Denmark.[10] Under the aggregated data hypothesis, some women who were over 70 years of age at the beginning of screening, and therefore have never been screened, are included in the postscreening follow-up. These women are older and therefore at greater risk of cancer; hence, this leads to an overestimation of risk. Similarly, some women were not followed up so no hypothesis on their future incidence can be explored.

the Lexis diagram (figure 6), instead of parallelograms corresponding to the follow-up of generations.

Jørgensen *et al* first estimated the relative risk of breast cancer in the 50–69 age group during the screening period (solid orange rectangle) as compared with the risk in the same age group during the reference period (faded orange rectangle).[7] They used this relative risk to estimate the initial excess of cases, due to screening. They then estimated the same relative risk in the 70–79 age group (solid and faded yellow rectangles), and used it to estimate the post-screening deficit, the number of cases that would have been diagnosed later if there was no screening. By subtracting the postscreening deficit from the initial excess, they estimated the number of 'falsely' diagnosed breast cancers, which was translated to a 33% rate of overdiagnosis.

Two major flaws with this design are shown on figure 6. The first is that a fraction of the patients, shown in the upper-right triangle, were never screened, because they were older than the upper age limit for screening at the beginning of the screening period. The inclusion of these unscreened older patients in the 'postscreening' follow-up overestimates the overdiagnosis rate. The second flaw is that the screened patients in the lower-right trapezoid were never followed up, so there is no information on a possible compensatory drop in later incidence. Moreover, this design cannot adjust for the evolution of medical techniques and imaging over time.

DISCUSSION

Another paper by Njor *et al* reviewed five of the most quoted studies, which had produced high estimates of overdiagnosis (some of these studies considered only invasive breast cancers).[8–13] The data and the method used in each of these studies were identified, and each method was then applied to data from Denmark, adapting the timing to correspond to the timing of screening in Funen. Njor *et al*'s 2018 study shows that using these methods leads to mistakenly high estimates of overdiagnosis, explained essentially by a too short duration of follow-up and by an inadequate estimation of the incidence expected without screening in the population invited to screening.[8]

Follow-up duration

The first problem is a too short follow-up duration in the populations that are being compared. Similar to the neuroblastoma example, if one wants to compare the number of breast cancers in a screened and an unscreened population, the two populations must be followed up long enough after the end of screening to avoid attributing the excess incidence observed by the screening to overdiagnosis.

Zahl *et al*, for instance, studied the incidence in a population invited to screening only during the first 5 years of the programme (1996–2000), and could not measure the complete post-screening deficit.[9] They assumed it to be negligible based on the trend in breast cancer incidence in the population aged 70 or over. However, it is not the largely unscreened population aged 70 and over who should be considered: what is needed is the breast cancer incidence in the screened population at age 70 or over. Zahl *et al* attributed the total excess incidence in the screened group to overdiagnosis, without taking into account the diagnoses brought forward by screening and therefore unobserved later. This explains the mistakenly high estimate of overdiagnosis.

The incidence expected in the absence of screening

In this case, where there are no data from randomised trials, one needs to estimate the breast cancer incidence that would be expected without screening in the population invited to screening. This is generally estimated on the basis of the observed incidence at the same time in an unscreened population geographically close to the population invited to screening, or in the population invited to screening before the start of the screening programme. This requires some assumptions on the variation in breast cancer incidence with

space and with time. The validity of the estimation depends on the validity of these assumptions.

Jorgensen and Gotzsche estimated the expected incidence without screening by linearly extrapolating the prescreening incidence and concluded that there was 30%–40% overdiagnosis in Funen.[11] The same linear extrapolation performed in regions without screening would lead to an increase in the expected incidence between 12% and 17%. They have therefore attributed part of the increase, which was unrelated to screening but simply the effect of time, to overdiagnosis.

Similarly, Zahl and Maehlen assumed the breast cancer incidence to have remained stable in Norway before and during the screening, but the national registry data show that, in Norway just like in Denmark, breast cancer incidence was on the increase before screening started.[12] Taking this increasing trend into account reduces the estimation of overdiagnosis from 42% to 13%.

CONCLUSION

These analyses show empirically the diversity of estimations that can be obtained on the basis of the same data, using different methods. The estimations vary between 0% and 55%, but some rely on data observed on the same women; hence, they cannot all be correct.

An important difference between studies is the use of individual versus aggregated data. Figure 1 shows that all the studies providing estimates above 17% were based on aggregated data; conversely, none of the studies based on individual data provided estimates above 17%. However, some studies of aggregated data obtain estimations below 17%; some of these use the simulation programme MISCAN and others were done by the Euroscreen working group.

In conclusion, the estimation of overdiagnosis is a difficult exercise. The analysis of individual data is generally less biased. The screened population must be followed up for several years after the end of screening, and the adequacy of the estimated incidence expected without screening in the screened population must be discussed. The exposure of the population to different breast cancer risk factors (age at first pregnancy, number of children, alcohol consumption and hormonal treatment for menopause…) may have varied with time, and some of these factors have different effects according to age. Some exposures may also vary with area. For instance, a reduced use of hormonal treatment for menopause over time will lead to a reduction in the incidence of postmenopausal breast cancer only, and the use of hormonal treatment for menopause may have been reduced earlier in some parts of a country than in others.

In the end, any overdiagnosis estimation is an arithmetic combination of observed data. The selection of the data and the way to combine them are more or less judicious, depending on what the investigators have understood of the problem.

Contributors Both DC and CH were involved in the conceptualisation of the overall paper and successive drafts, and contributed to the planning, conduct and reporting of the work described in the article. CH was responsible of the design of the paper.

Funding DC is an employee of Institut Gustave Roussy, the largest cancer hospital in France, and CH has retired from the same Institute.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Dan Chaltiel http://orcid.org/0000-0003-3488-779X
Catherine Hill http://orcid.org/0000-0002-0004-090

REFERENCES

1 Ripping TM, ten Haaf K, Verbeek ALM, *et al*. Quantifying overdiagnosis in cancer screening: a systematic review to evaluate the methodology. *J Natl Cancer Inst* 2017;109.
2 Baker SG, Prorok PC. Breast cancer overdiagnosis in stop-screen trials: more uncertainty than previously reported. *J Med Screen* 2020;969141320950784:096914132095078.
3 Heywang-Köbrunner SH, Schreer I, Hacker A, *et al*. Conclusions for mammography screening after 25-year follow-up of the Canadian National breast cancer screening study (CNBSS). *Eur Radiol* 2016;26:342–50.
4 Schilling FH, Spix C, Berthold F, *et al*. Neuroblastoma screening at one year of age. *N Engl J Med* 2002;346:1047–53.
5 Spix C, Michaelis J, Berthold F, *et al*. Lead-time and overdiagnosis estimation in neuroblastoma screening. *Stat Med* 2003;22:2877–92.
6 Njor SH, Olsen AH, Blichert-Toft M, *et al*. Overdiagnosis in screening mammography in Denmark: population based cohort study. *BMJ* 2013;346:f1064.
7 Jørgensen KJ, Gøtzsche PC, Kalager M, *et al*. Breast cancer screening in Denmark: a cohort study of tumor size and overdiagnosis. *Ann Intern Med* 2017;166:313–23.
8 Njor SH, Paci E, Rebolj M. As you like it: how the same data can support manifold views of overdiagnosis in breast cancer screening. *Int J Cancer* 2018;143:1287–94.
9 Zahl P-H, Strand BH, Mæhlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. *BMJ* 2004;328:921–4.
10 Jørgensen KJ, Zahl P-H, Gøtzsche PC. Overdiagnosis in organised mammography screening in Denmark. A comparative study. *BMC Womens Health* 2009;9:36.
11 Jorgensen KJ, Gotzsche PC. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. centre for reviews and dissemination (UK), 2009. Available: https://www.ncbi.nlm.nih.gov/books/NBK78523/ [Accessed 5 Nov 2020].
12 Zahl P-H, Mæhlen J. Overdiagnosis of breast cancer after 14 years of mammography screening. *Tidsskr Nor Laegeforen* 2012;132:414–7.
13 Kalager M, Adami H-O, Bretthauer M, *et al*. Overdiagnosis of invasive breast cancer due to mammography screening: results from the Norwegian screening program. *Ann Intern Med* 2012;156:491–9.