BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-046353.R1
Article Type:	Communication
Date Submitted by the Author:	15-Nov-2020
Complete List of Authors:	Chaltiel, Dan; Gustave Roussy, Hill, Catherine; Gustave Roussy,
Primary Subject Heading:	Public health
Secondary Subject Heading:	Epidemiology, Diagnostics, Health policy, Oncology, Obstetrics and gynaecology
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Breast tumours < ONCOLOGY, PREVENTIVE MEDICINE, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

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

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

Findings

Analysis of the data from Denmark demonstrates the importance of using individual patients' data and identifies the use of aggregated data as the main reason for overestimation of overdiagnosis.

Interpretation

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

1. Introduction

Many countries have a national breast cancer screening program in which all women belonging to a specific age-group are invited to have regular mammograms. These programs have been criticized, 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 programs[1]. Since it is extremely unlikely that overdiagnosis varies to such a large extent from one program to another, one needs to study possible causes for this observed heterogeneity.

2. 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 (specialized 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 one-year-old children was systematically offered [2].

Such a screening program 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) [3]. In theory, the screening program 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 one 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 [2]).

This study showed that screening for neuroblastoma at one 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.

3. 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. [4]. The data used were individual data, i.e., 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 programs, 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 was a screening program, versus other regions, where there was not) and during several periods (at the time of the screening program versus 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 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 program started on November 1, 1993, and the whole female population aged 50 to 69 was invited.

Each screening round lasted two years; therefore, the first found spanned from November 1, 1993 to October 31, 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 31/12/2009 at the latest. Therefore, in order to have sufficient follow-up time, Njor et al. included only patients aged 59 to 70 on November 1, 1993, as younger patients would not have been followed for long enough [4]. 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 randomized 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 November 1, 1923-October 31, 1934), data from Funen before screening (generation November 1, 1912-October 31, 1923), and data from Danish regions without screening before the introduction of screening in Funen (generation November 1, 1912-October 31, 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 to 3 years, 4 to 7 years, and 8+ years from the end of invitation to screening, respectively [4]. 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 program. However, this is still only half of the solution, as it would not take into account the effect of geography.

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 one-year generations: 1) women born in 1922 who were 71 on 1/11/1993 and, hence, were never invited to screening; and 2) women born in 1932 who were 61 on 1/11/1993 and, hence, were invited to screening.

Table 1 and the corresponding Figure 4 show the incidence as a function of age in these two generations, in Funen versus 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. Using the totals in Table 1 would lead to a ratio of cumulative incidences equal to 1·03 [(1436x1065)/(1082x1371)]; hence, giving a 3% estimation of overdiagnosis.

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 [4].

As described in Table 2 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 program, 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 [4].

4. 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) [4].

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 to 69 and 70 to 79 [5]. Therefore, these four populations are represented by four rectangles in 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 to 69 age group during the screening period (solid orange rectangle) as compared to the risk in the same age group during the reference period (faded orange rectangle) [5]. 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

post-screening 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 yellow 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 "post-screening" follow-up overestimates the overdiagnosis rate. The second flaw is that the screened patients in the lower orange 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.

5. 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) [6–11]. 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 [6].

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 five years of the program (1996-2000), and could not measure the complete post-screening deficit [7]. 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 randomized 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 program. 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.

Jørgensen and Gotzsche estimated the expected incidence without screening by linearly extrapolating the pre-screening incidence and concluded that there was 30% to 40% overdiagnosis in Funen [9]. 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 screening, but the national registry data show that, in Norway just like in Denmark, breast cancer incidence was on the increase before screening started [10]. Taking this increasing trend into account reduces the estimation of overdiagnosis from 42% to 13%.

6. 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 program 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 post-menopausal 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.

Statements

Contributorship

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.

Competing interests

There are no competing interests for any author

Funding

This study has received no specific funding. DC is an employee of Institut Gustave Roussy, the largest cancer hospital in France, and CH has retired from the same Institute.

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**. doi:10.1093/jnci/djx060

2 Schilling FH, Spix C, Berthold F, *et al.* Neuroblastoma screening at one year of age. *N Engl J Med* 2002;**346**:1047–53. doi:10.1056/NEJMoa012277

3 Spix C, Michaelis J, Berthold F, *et al.* Lead-time and overdiagnosis estimation in neuroblastoma screening. *StatMed* 2003;**22**:2877–92. doi:10.1002/sim.1533

4 Njor SH, Olsen AH, Blichert-Toft M, *et al.* Overdiagnosis in screening mammography in Denmark: population based cohort study. *BMJ* 2013;**346**. doi:10.1136/bmj.f1064

5 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. doi:10.7326/M16-0270

6 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. doi:10.1002/ijc.31420

7 Zahl P-H, Strand BH, Maehlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. *BMJ* 2004;**328**:921–4. doi:10.1136/bmj.38044.666157.63

8 Jørgensen KJ, Zahl P-H, Gøtzsche PC. Overdiagnosis in organised mammography screening in Denmark. A comparative study. *BMC Womens Health* 2009;**9**:36. doi:10.1186/1472-6874-9-36

9 Jorgensen KJ, Gotzsche PC. *Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends*. Centre for Reviews and Dissemination (UK) 2009. https://www.ncbi.nlm.nih.gov/books/NBK78523/ (accessed 5 Nov 2020).

10 Zahl P-H, Mæhlen J. Overdiagnosis of breast cancer after 14 years of mammography screening. *Tidsskr Nor Laegeforen* 2012;**132**:414–7. doi:10.4045/tidsskr.11.0195

11 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. doi:10.7326/0003-4819-156-7-201204030-00005

Tables

Table 1: *Breast cancer incidence per 100,000 by age-group in Funen and in other regions for two generations, taken from Fig. 2 of Njor et al. 2018 [6].*

Age-group	Other regions, generation 1922	Other regions, generation 1932	Funen, generation 1922 (without screening)	Funen, generation 1932 (with screening)
50-54	153	181	166	210
55-59	178	216	152	202
60-65	204	281	236	402
65-69	237	325	251	373
70-74	293	368	278	249
Total	1065	1371	1083	1436

Table 2*: Breast cancer incidence per 100,000 observed in Funen for every generation, and model estimations in the case of no screening program. Adapted from Njor et al. 2013 [4].*

Period	Before screening	Invitation to screening		Follow-up post-screening			Cumulative over generations
		1st 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

* 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.

Figures captions

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 C. Hill*

Figure 2*: Overdiagnosis estimation, example of screening for neuroblastoma in Germany. Based on Schilling et al. and Spix et al. [2,3].*
Control and test regions have a comparable population size, with 1.1 and 1.5 million children, respectively. Incidence is expressed in arbitrary units.
2a: *Incidence is displayed as a function of age, and generalized neuroblastoma screening takes place at one year of age. There is logically no difference in incidence between control and test regions before screening age (<1-year-old). The screening program 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.*
2b: *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).*
2c: *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).*
2d: *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.*

Figure 3*: Lexis diagrams of the Funen overdiagnosis experiment, based on Njor et al. [4]. Generations can be followed on diagonals.*
3a: *Only women born between 01/11/1923 and 01/11/1943, who were 50 to 69 at the start of screening (01/11/1993), were invited to screening.*
3b: *In order to have sufficient follow-up time (follow-up ended on 31/12/2009), 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.*
3c: *When following the screened population (S), several periods can be identified: first screenings (red), later screenings (orange), and 3 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.*

Figure 4: *Incidence of breast cancer as a function of age in Funen with screening (**green**), compared to 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**). Adapted from Njor et al. 2018 [6].*

Figure 5: *Incidence of breast cancer in Funen, compared to control. The incidence of breast cancer observed in Funen during screening (in the same colours as in figure 3c) 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 (2009) to estimate overdiagnosis in Denmark [8]. 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 post-screening 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.*

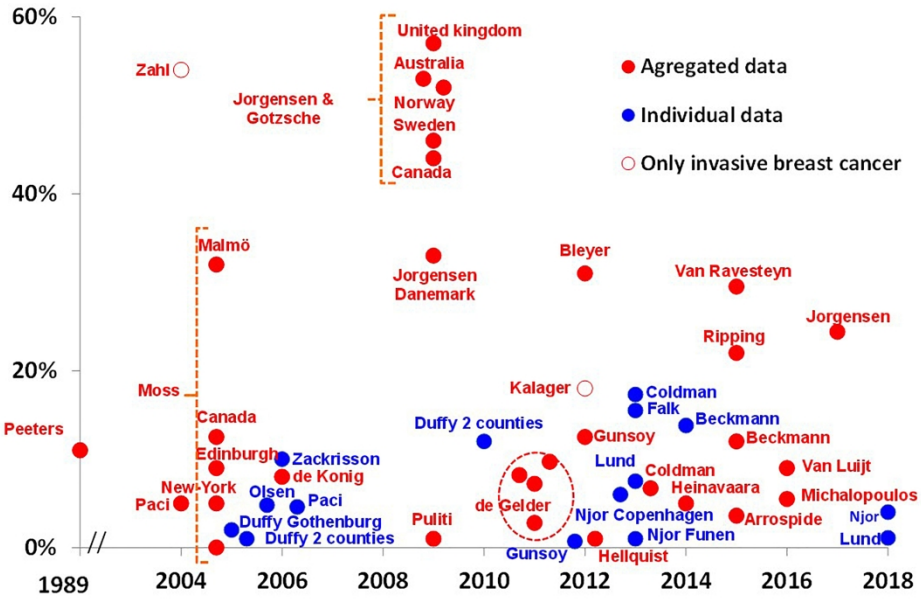

Figure 1: Published estimations of in situ and invasive breast cancer overdiagnosis (open symbols: two publications studying only invasive breast cancers quoted in the present text). Studies conducted on aggregated data give generally higher estimations of overdiagnosis than studies conducted on individual data. Source: Ripping et al. [1]. Updated by C. Hill

308x194mm (300 x 300 DPI)

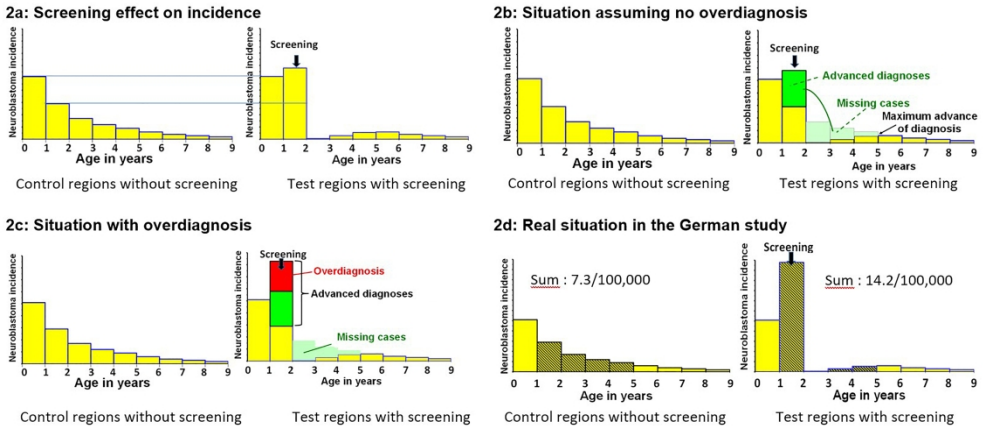

Figure 2: Overdiagnosis estimation, example of screening for neuroblastoma in Germany. Based on Schilling et al. and Spix et al. [2,3].
Control and test regions have a comparable population size, with 1.1 and 1.5 million children, respectively. Incidence is expressed in arbitrary units.
2a: Incidence is displayed as a function of age, and generalized neuroblastoma screening takes place at one year of age. There is logically no difference in incidence between control and test regions before screening age (<1-year-old). The screening program causes an increased incidence of cases immediately after screening at age 1, a decrease shortly afterward, and a return to normal at around age 5.
2b: 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).
2c: 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).
2d: 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.

435x187mm (300 x 300 DPI)

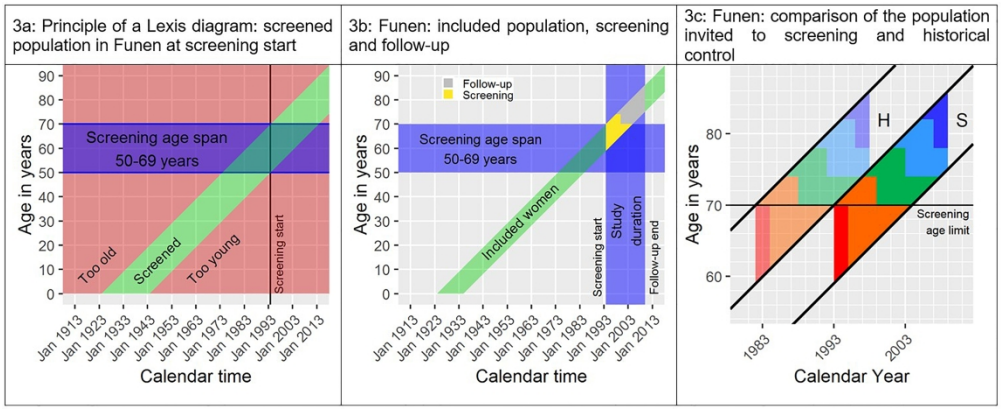

Figure 3: Lexis diagrams of the Funen overdiagnosis experiment, based on Njor et al. [4]. Generations can be followed on diagonals.
3a: Only women born between 01/11/1923 and 01/11/1943, who were 50 to 69 at the start of screening (01/11/1993), were invited to screening.
3b: In order to have sufficient follow-up time (follow-up ended on 31/12/2009), 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.
3c: When following the screened population (S), several periods can be identified: first screenings (red), later screenings (orange), and 3 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.

400x162mm (300 x 300 DPI)

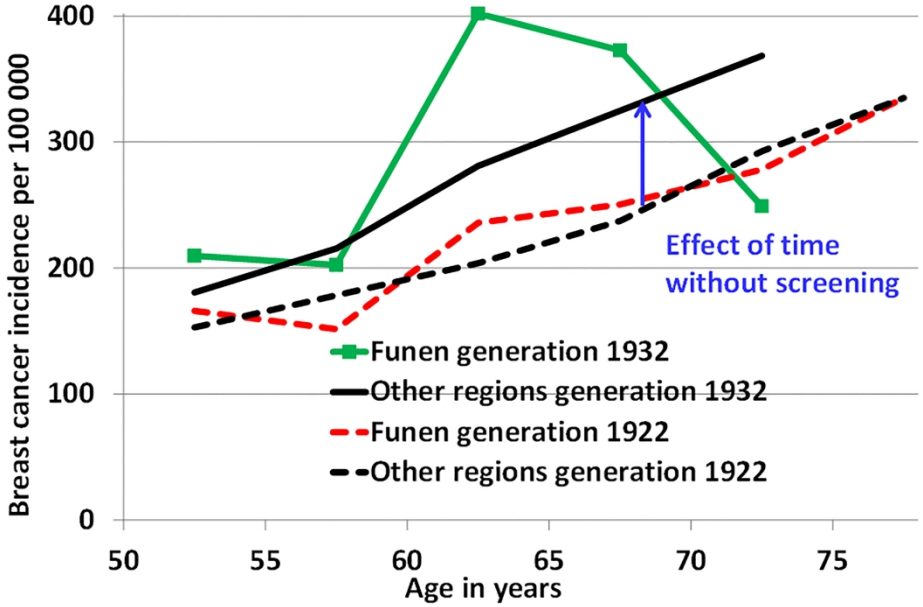

Figure 4: Incidence of breast cancer as a function of age in Funen with screening (green), compared to a historical control group (Funen in a different period, red), to a national control group (other regions, same period, solid black line), and to a historical national control group (other regions, different period, dashed black line). Adapted from Njor et al. 2018 [6].

258x167mm (300 x 300 DPI)

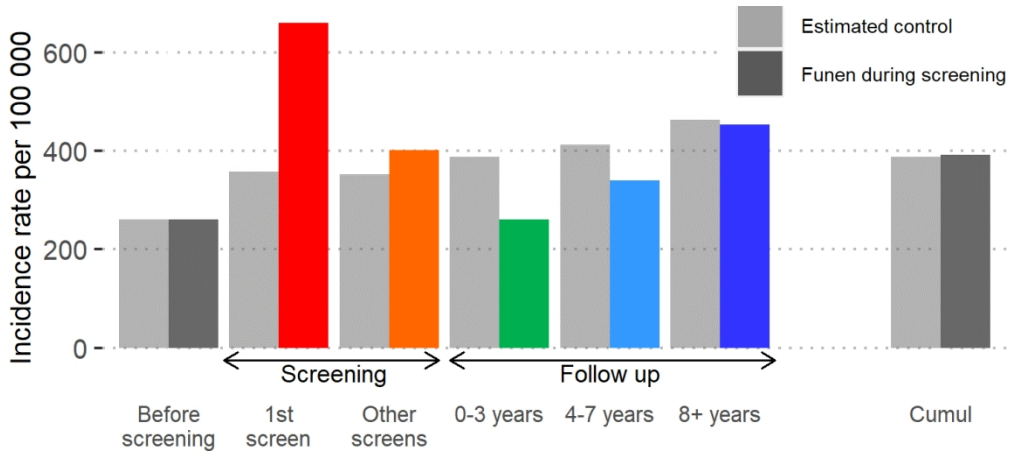

Figure 5: Incidence of breast cancer in Funen, compared to control. The incidence of breast cancer observed in Funen during screening (in the same colours as in figure 3c) is compared in each period to the incidence in Funen estimated by the model in the absence of screening (in grey).

125x55mm (300 x 300 DPI)

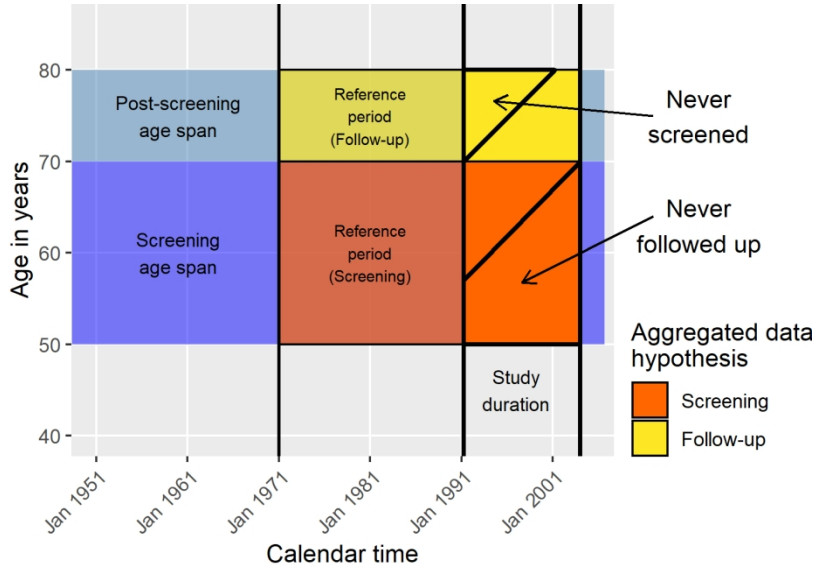

Figure 6: Data analysed by Jørgensen (2009) to estimate overdiagnosis in Denmark [8]. Under the aggregated data hypothesis, some women who were over 70 years of age at the beginning of screening, and therefore have never been screened, are included in the post-screening follow-up. These women are older 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.

159x89mm (300 x 300 DPI)

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-046353.R2
Article Type:	Communication
Date Submitted by the Author:	20-Feb-2021
Complete List of Authors:	Chaltiel, Dan; Gustave Roussy, Hill, Catherine; Gustave Roussy
Primary Subject Heading:	Public health
Secondary Subject Heading:	Epidemiology, Diagnostics, Health policy, Oncology, Obstetrics and gynaecology
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Breast tumours < ONCOLOGY, PREVENTIVE MEDICINE, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

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

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

Correspondence to:
Dr Dan Chaltiel
dan.chaltiel@gustaveroussy.fr
Service de Biostatistique et d'Epidémiologie
Gustave Roussy
114 rue Edouard Vaillant
94805 Villejuif cedex
Tel. : +33 (0)1 42 11 42 11
Fax. : +33 (0)1 42 11 42 17

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 programs. 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.

Findings

Analysis of the data from Denmark demonstrates the importance of using individual patients' data and identifies the use of aggregated data as the main reason for overestimation of overdiagnosis.

Interpretation

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

1. Introduction

Many countries have a national breast cancer screening program in which all women belonging to a specific age-group are invited to have regular mammograms. These programs have been criticized, 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 programs[1]. Since it is extremely unlikely that overdiagnosis varies to such a large extent from one program to another, one needs to study possible causes for this observed heterogeneity.

2. Estimation Methods

The ideal approach to estimate the overdiagnosis rate would be to use data from randomized 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 confidence interval 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.

3. 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 (specialized 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 one-year-old children was systematically offered [4].

Such a screening program 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 program 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 one 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]).

This study showed that screening for neuroblastoma at one 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.

4. 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, i.e., 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 programs, 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 was a screening program, versus other regions, where there was not) and during several periods (at the time of the screening program versus 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 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 program started on November 1, 1993, and the whole female population aged 50 to 69 was invited.

Each screening round lasted two years; therefore, the first found spanned from November 1, 1993 to October 31, 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 31/12/2009 at the latest. Therefore, in order to have sufficient follow-up time, Njor et al. included only patients aged 59 to 70 on November 1, 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 randomized 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 November 1, 1923-October 31, 1934), data from Funen before screening (generation November 1, 1912-October 31, 1923), and data from Danish regions without screening before the introduction of screening in Funen (generation November 1, 1912-October 31, 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 to 3 years, 4 to 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 program. However, this is still only half of the solution, as it would not take into account the effect of geography.

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 one-year generations: 1) women born in 1922 who were 71 on 1/11/1993 and, hence,

were never invited to screening; and 2) women born in 1932 who were 61 on 1/11/1993 and, hence, were invited to screening.

Table 1 and the corresponding Figure 4 show the incidence as a function of age in these two generations, in Funen versus 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. Using the totals in Table 1 would lead to a ratio of cumulative incidences equal to 1·03 [(1436x1065)/(1082x1371)]; hence, giving a 3% estimation of overdiagnosis.

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 2 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 program, 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].

5. 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 to 69 and 70 to 79 [7]. Therefore, these four populations are represented by four rectangles in 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 to 69 age group during the screening period (solid orange rectangle) as compared to 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 post-screening 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 yellow 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 "post-screening" follow-up overestimates the overdiagnosis rate. The second flaw is that the screened patients in the lower orange 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.

6. 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 five years of the program (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 randomized 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 program. 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.

Jørgensen and Gotzsche estimated the expected incidence without screening by linearly extrapolating the pre-screening incidence and concluded that there was 30% to 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 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%.

7. 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 program 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 post-menopausal 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.

Statements

Contributorship

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.

Competing interests

There are no competing interests for any author

Funding

This study has received no specific funding. DC is an employee of Institut Gustave Roussy, the largest cancer hospital in France, and CH has retired from the same Institute.

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**. doi:10.1093/jnci/djx060

2 Baker SG, Prorok PC. Breast cancer overdiagnosis in stop-screen trials: More uncertainty than previously reported. *J Med Screen* 2020;:969141320950784. doi:10.1177/0969141320950784

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. doi:10.1007/s00330-015-3849-2

4 Schilling FH, Spix C, Berthold F, *et al.* Neuroblastoma screening at one year of age. *N Engl J Med* 2002;**346**:1047–53. doi:10.1056/NEJMoa012277

5 Spix C, Michaelis J, Berthold F, *et al.* Lead-time and overdiagnosis estimation in neuroblastoma screening. *StatMed* 2003;**22**:2877–92. doi:10.1002/sim.1533

6 Njor SH, Olsen AH, Blichert-Toft M, *et al.* Overdiagnosis in screening mammography in Denmark: population based cohort study. *BMJ* 2013;**346**. doi:10.1136/bmj.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. doi:10.7326/M16-0270

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. doi:10.1002/ijc.31420

9 Zahl P-H, Strand BH, Maehlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. *BMJ* 2004;**328**:921–4. doi:10.1136/bmj.38044.666157.63

10 Jørgensen KJ, Zahl P-H, Gøtzsche PC. Overdiagnosis in organised mammography screening in Denmark. A comparative study. *BMC Womens Health* 2009;**9**:36. doi:10.1186/1472-6874-9-36

11 Jorgensen KJ, Gotzsche PC. *Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends*. Centre for Reviews and Dissemination (UK) 2009. 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. doi:10.4045/tidsskr.11.0195

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. doi:10.7326/0003-4819-156-7-201204030-00005

Tables

Table 1: *Breast cancer incidence per 100,000 by age-group in Funen and in other regions for two generations, taken from Fig. 2 of Njor et al. 2018.*

Age-group	Other regions, generation 1922	Other regions, generation 1932	Funen, generation 1922 (without screening)	Funen, generation 1932 (with screening)
50-54	153	181	166	210
55-59	178	216	152	202
60-65	204	281	236	402
65-69	237	325	251	373
70-74	293	368	278	249
Total	1065	1371	1083	1436

Table 2*: Breast cancer incidence per 100,000 observed in Funen for every generation, and model estimations in the case of no screening program. Adapted from Njor et al. 2013.*

Period	Before screening	Invitation to screening		Follow-up post-screening			Cumulative over generations
		1st 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

* 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.

Figures captions

***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 C. Hill. A comprehensive list of these studies is provided in Supplementary Data.*

***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.
***2a:** Incidence is displayed as a function of age, and generalized neuroblastoma screening takes place at one year of age. There is logically no difference in incidence between control and test regions before screening age (<1-year-old). The screening program 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.*
***2b:** 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).*
***2c:** 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).*
***2d:** 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.*

***Figure 3**: Lexis diagrams of the Funen overdiagnosis experiment, based on Njor et al. [6]. Generations can be followed on diagonals.*
***3a:** Only women born between 01/11/1923 and 01/11/1943, who were 50 to 69 at the start of screening (01/11/1993), were invited to screening.*
***3b:** In order to have sufficient follow-up time (follow-up ended on 31/12/2009), 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.*
***3c:** When following the screened population (S), several periods can be identified: first screenings (red), later screenings (orange), and 3 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.*

Figure 4: *Incidence of breast cancer as a function of age in Funen with screening (**green**), compared to 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 five-year age group (e.g., a dot between 50 and 55 represents the age group 50-54).*
Adapted from Njor et al. 2018 [8].

Figure 5: *Incidence of breast cancer in Funen, compared to 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 (2009) 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 post-screening 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.*

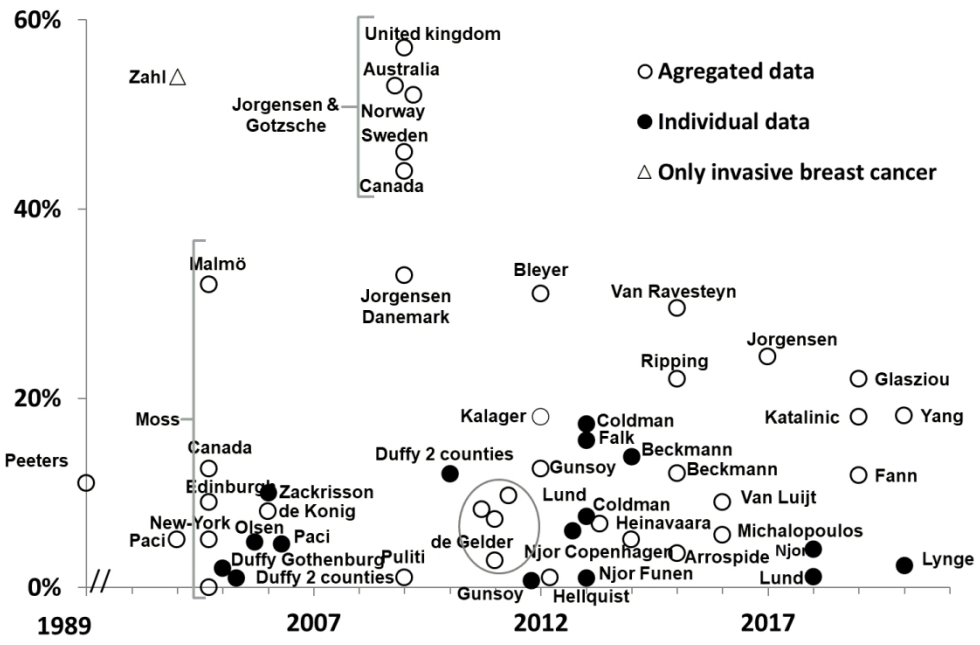

Figure 1: Published estimations of in situ and invasive breast cancer overdiagnosis (open symbols: two publications studying only invasive breast cancers quoted in the present text). Studies conducted on aggregated data give generally higher estimations of overdiagnosis than studies conducted on individual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

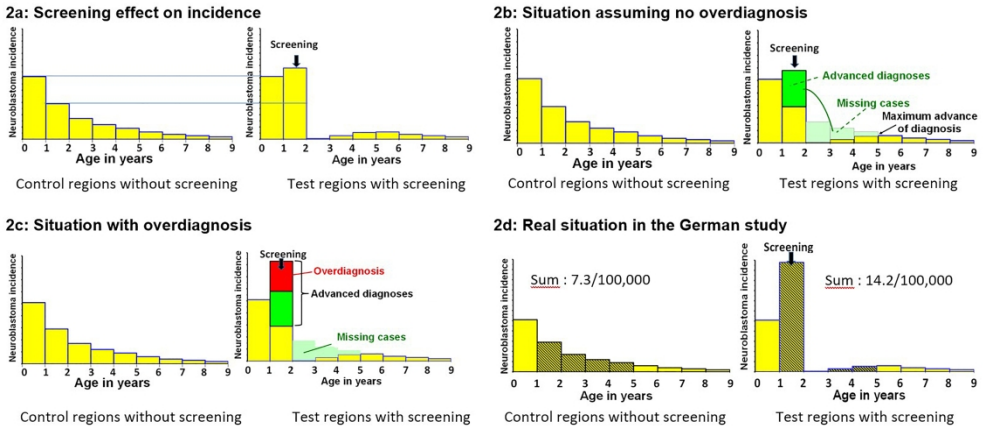

Figure 2: Overdiagnosis estimation, example of screening for neuroblastoma in Germany. Based on Schilling et al. and Spix et al. [2,3].
Control and test regions have a comparable population size, with 1.1 and 1.5 million children, respectively. Incidence is expressed in arbitrary units.
2a: Incidence is displayed as a function of age, and generalized neuroblastoma screening takes place at one year of age. There is logically no difference in incidence between control and test regions before screening age (<1-year-old). The screening program causes an increased incidence of cases immediately after screening at age 1, a decrease shortly afterward, and a return to normal at around age 5.
2b: 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).
2c: 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).
2d: 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.

435x187mm (300 x 300 DPI)

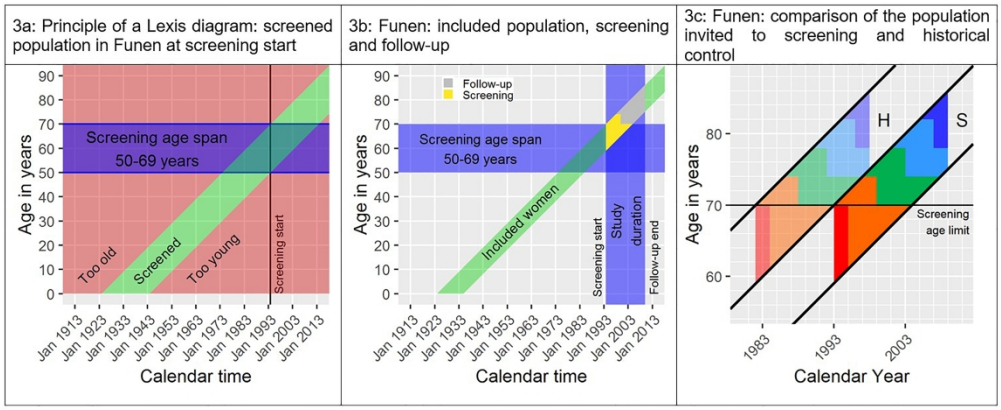

Figure 3: Lexis diagrams of the Funen overdiagnosis experiment, based on Njor et al. [4]. Generations can be followed on diagonals.
3a: Only women born between 01/11/1923 and 01/11/1943, who were 50 to 69 at the start of screening (01/11/1993), were invited to screening.
3b: In order to have sufficient follow-up time (follow-up ended on 31/12/2009), 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.
3c: When following the screened population (S), several periods can be identified: first screenings (red), later screenings (orange), and 3 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.

400x162mm (300 x 300 DPI)

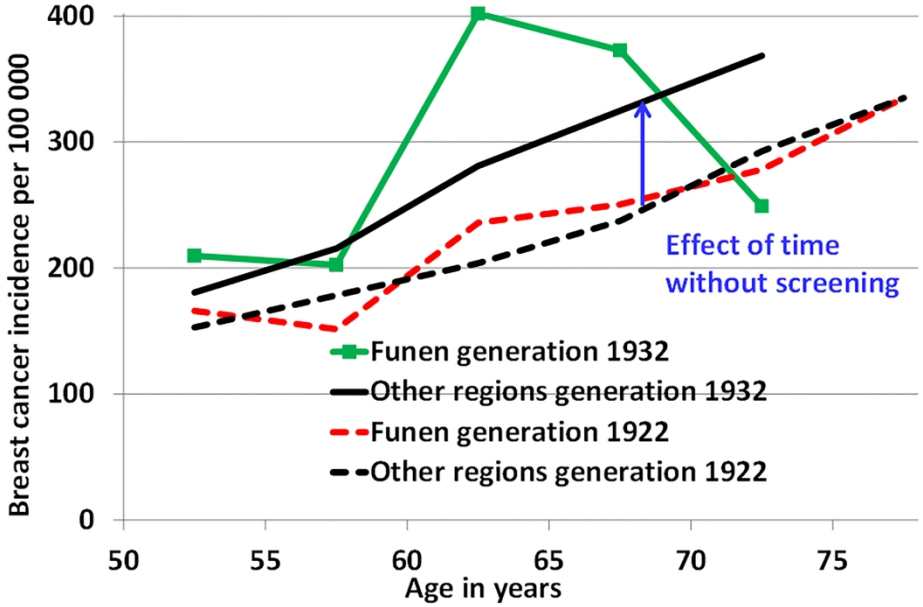

Figure 4: Incidence of breast cancer as a function of age in Funen with screening (green), compared to a historical control group (Funen in a different period, red), to a national control group (other regions, same period, solid black line), and to a historical national control group (other regions, different period, dashed black line). Adapted from Njor et al. 2018 [6].

258x167mm (300 x 300 DPI)

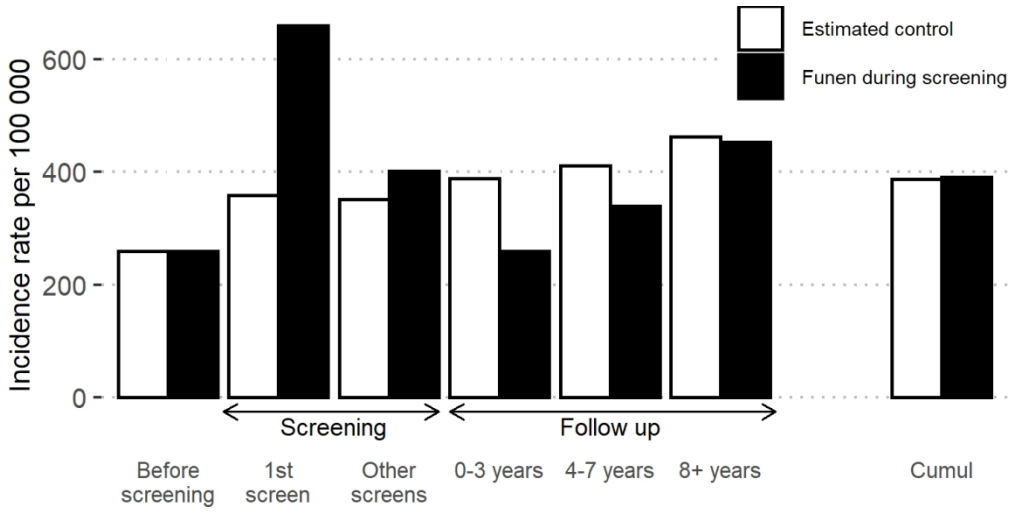

Figure 5: Incidence of breast cancer in Funen, compared to control. The incidence of breast cancer observed in Funen during screening (in black) is compared in each period to the incidence in Funen estimated by the model in the absence of screening (in grey).

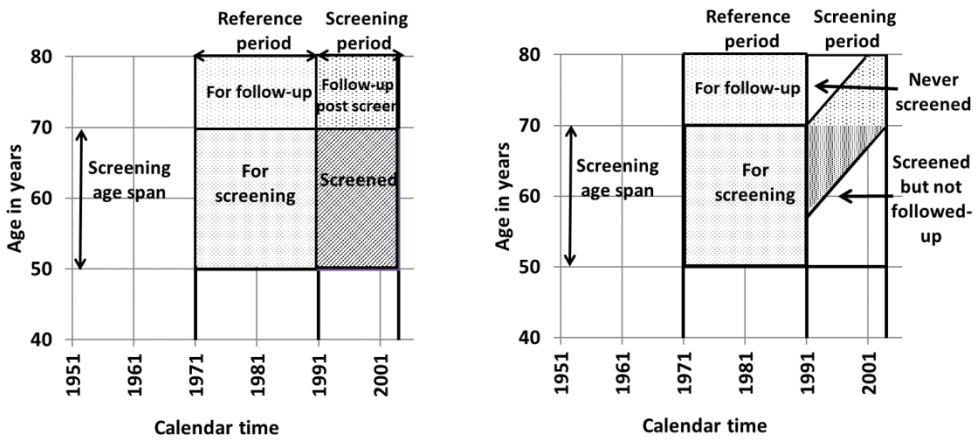

Figure 6: Data analysed by Jørgensen (2009) to estimate overdiagnosis in Denmark [9]. Under the aggregated data hypothesis, some women who were over 70 years of age at the beginning of screening, and therefore have never been screened, are included in the post-screening follow-up. These women are older 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.

Supplementary data

Comprehensive list of studies used to generate Figure 1:

1. Beckmann K, Duffy SW, Lynch J, Hiller J, Farshid G, Roder D. Estimates of over-diagnosis of breast cancer due to population-based mammography screening in South Australia after adjustment for lead time effects. J Med Screen. 2015;22(3):127–135.
2. Paci E, Miccinesi G, Puliti D, et al. Estimate of overdiagnosis of breast cancer due to mammography after adjustment for lead time. A service screening study in Italy. Breast Cancer Res. 2006;8(6):R68.
3. Paci E, Warwick J, Falini P, Duffy SW. Overdiagnosis in screening: Is the increase in breast cancer incidence rates a cause for concern? J Med Screen. 2004;11(1):23–27.
4. Morrell S, Barratt A, Irwig L, Howard K, Biesheuvel C, Armstrong B. Estimates of overdiagnosis of invasive breast cancer associated with screening mammography. Cancer Causes Control. 2010;21(2):275–282.
5. Hellquist BN, Duffy SW, Nystrom L, Jonsson H. Overdiagnosis in the population-based service screening programme with mammography for women aged 40 to 49 years in Sweden. J Med Screen. 2012;19(1):14–19.
6. Jonsson H, Johansson R, Lenner P. Increased incidence of invasive breast cancer after the introduction of service screening with mammography in Sweden. Int J Cancer. 2005;117(5):842–847.
7. Kalager M, Adami H-O, Bretthauer M, Tamimi RM. Overdiagnosis of invasive breast cancer due to mammography screening: Results from the Norwegian Screening Program. Ann Intern Med. 2012;156(7):491–499.
8. Martinez-Alonso M, Vilaprinyo E, Marcos-Gragera R, Rue M. Breast cancer incidence and overdiagnosis in Catalonia (Spain). Breast Cancer Res. 2010;12(4): R58.
9. Svendsen AL, Olsen AH, Von Euler-Chelpin M, Lynge E. Breast cancer incidence after the introduction of mammography screening: What should be expected? Cancer. 2006;106(9):1883–1890.
10. Duffy SW, Agbaje O, Tabar L, et al. Overdiagnosis and overtreatment of breast cancer: Estimates of overdiagnosis from two trials of mammographic screening for breast cancer. Breast Cancer Res. 2005;7(6):258–265.
11. Olsen AH, Agbaje OF, Myles JP, Lynge E, Duffy SW. Overdiagnosis, sojourn time, and sensitivity in the Copenhagen Mammography Screening Program. Breast J. 2006;12(4):338–342.
12. Arrospide A, Rue M, van Ravesteyn NT, et al. Evaluation of health benefits and harms of the breast cancer screening programme in the Basque Country using discrete event simulation. BMC Cancer. 2015;15(1):671.
13. De Gelder R, Fracheboud J, Heijnsdijk EAM, et al. Digital mammography screening: Weighing reduced mortality against increased overdiagnosis. Prev Med. 2011;53(3):134–140.
14. De Gelder R, Heijnsdijk EAM, Van Ravesteyn NT, Fracheboud J, Draisma G, De Koning HJ. Interpreting overdiagnosis estimates in population-based mammography screening. Epidemiol Rev. 2011;33(1):111–121.

15. de Koning HJ, Draisma G, Fracheboud J, de Bruijn A. Overdiagnosis and overtreatment of breast cancer: Microsimulation modelling estimates based on observed screen and clinical data. Breast Cancer Res. 2006;8(1):202.
16. Gunsoy NB, Garcia-Closas M, Moss SM. Modelling the overdiagnosis of breast cancer due to mammography screening in women aged 40 to 49 in the United Kingdom. Breast Cancer Res. 2012;14(6):R152.
17. Gunsoy NB, Garcia-Closas M, Moss SM. Estimating breast cancer mortality reduction and overdiagnosis due to screening for different strategies in the United Kingdom. Br J Cancer. 2014;110(10):2412–2419.
18. Seigneurin A, Francois O, Labarere J, Oudeville P, Monlong J, Colonna M. Overdiagnosis from non-progressive cancer detected by screening mammography: Stochastic simulation study with calibration to population based registry data. BMJ. 2011;343:d7017.
19. van Ravesteyn NT, Stout NK, Schechter CB, et al. Benefits and harms of mammography screening after age 74 years: Model estimates of overdiagnosis. J Natl Cancer Inst. 2015;107(7): djv103.
20. Beckmann KR, Lynch JW, Hiller JE, et al. A novel case-control design to estimate the extent of over-diagnosis of breast cancer due to organised population-based mammography screening. Int J Cancer. 2015;136(6): 1411–1421.
21. Coldman A, Phillips N. Incidence of breast cancer and estimates of overdiagnosis after the initiation of a population-based mammography screening program. CMAJ. 2013;185(10):E492–E498.
22. Duffy SWT, Olsen L, Vitak AH, et al. Absolute numbers of lives saved and overdiagnosis in breast cancer screening, from a randomized trial and from the Breast Screening Programme in England. J Med Screen. 2010;17(1):25–30.
23. Falk RS, Hofvind S, Skaane P, Haldorsen T. Overdiagnosis among women attending a population-based mammography screening program. Int J Cancer. 2013;133(3):705–712.
24. Heinavaara S, Sarkeala T, Anttila A. Overdiagnosis due to breast cancer screening: Updated estimates of the Helsinki service study in Finland. Br J Cancer. 2014;14(10):413.
25. Jørgensen KJ, Gøtzsche PC. Overdiagnosis in publicly organised mammography screening programmes: Systematic review of incidence trends. BMJ. 2009;339(7714):206–209.
26. Jorgensen KJ, Zahl PH, Gotzsche PC. Overdiagnosis in organised mammography screening in Denmark. A comparative study. BMC Womens Health. 2009; 9:36.
27. Lund E, Mode N, Waaseth M, Thalabard J-C. Overdiagnosis of breast cancer in the Norwegian Breast Cancer Screening Program estimated by the Norwegian Women and Cancer cohort study. BMC Cancer. 2013;13(1):614.
28. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: An independent review. Br J Cancer. 2013;108(11):2205–2240.
29. Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year followup for breast cancer incidence and mortality of the Canadian National Breast Screening Study: Randomised screening trial. BMJ. 2014;348:g366.
30. Moss S. Overdiagnosis and overtreatment of breast cancer: Overdiagnosis in randomised controlled trials of breast cancer screening. Breast Cancer Res. 2005;7(5):230–234.
31. Njor SH, Olsen AH, Blichert-Toft M, Schwartz W, Vejborg I, Lynge E. Overdiagnosis in screening mammography in Denmark: Population based cohort study. BMJ. 2013;346:f1064.

32. Peeters PHM, Verbeek ALM, Straatman H, et al. Evaluation of overdiagnosis of breast cancer in screening with mammography: Results of the Nijmegen Programme. Int J Epidemiol. 1989;18(2):295–299.
33. Puliti D, Miccinesi G, Zappa M, Manneschi G, Crocetti E, Paci E. Balancing harms and benefits of service mammography screening programs: A cohort study. Breast Cancer Res. 2012;14(1):R9.
34. Puliti D, Zappa M, Miccinesi G, Falini P, Crocetti E, Paci E. An estimate of overdiagnosis 15 years after the start of mammographic screening in Florence. Eur J Cancer. 2009;45(18):3166–3171.
35. Ripping TM, Verbeek AL, Fracheboud J, de Koning HJ, van Ravesteyn NT, Broeders MJ. Overdiagnosis by mammographic screening for breast cancer studied in birth cohorts in The Netherlands. Int J Cancer. 2015;137(4):921–929.
36. Waller M, Moss S, Watson J, Møller H. The effect of mammographic screening and hormone replacement therapy use on breast cancer incidence in England and Wales. Cancer Epidemiol Biomarkers Prev. 2007;16(11):2257–2261.
37. Zackrisson S, Andersson I, Janzon L, Manjer J, Garne JP. Rate of overdiagnosis of breast cancer 15 years after end of Malmo¨ mammographic screening trial: Follow-up study. BMJ. 2006;332(7543):689–692.
38. Zahl PH, Strand BH, Maehlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: Prospective cohort study. BMJ. 2004;328(7445):921–924.
39. Yen AM, Duffy SW, Chen TH, et al. Long-term incidence of breast cancer by trial arm in one county of the Swedish Two-County Trial of mammographic screening. Cancer. 2012;118(23):5728–5732.
40. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. N Engl J Med. 2012;367(21):1998–2005.
41. Michalopoulos D, Duffy SW. Estimation of overdiagnosis using short-term trends and lead time estimates uncontaminated by overdiagnosed cases: Results from the Norwegian Breast Screening Programme. J Med Screen 2016, Vol. 23(4) 192–202
42. Jørgensen KJ, Gøtzsche PC, Kalager M and Zahl PH. Breast Cancer Screening in Denmark A Cohort Study of Tumor Size and Overdiagnosis Ann Intern Med. 2017;166:313-323. doi:10.7326/M16-0270
43. Lund E, Nakamura A, Thalabard JC. No overdiagnosis in the Norwegian Breast Cancer Screening Program estimated by combining record linkage and questionnaire information in the Norwegian Women and Cancer study. European Journal of Cancer 2018; 89: 102e112
44. Njor SH, Paci E and 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 : 1287–1294
45. Glasziou PP, Jones MA, Pathirana T, et al. Estimating the magnitude of cancer overdiagnosis in Australia. Med J Aust 2020; doi: 10.5694/mja2.50455.
46. Lynge E, Beau AB, von Euler-Chelpin M, Napolitano G, Njor S, Olsen AH, Schwartz W, Vejborg I. Breast cancer mortality and overdiagnosis after implementation of population-based screening in Denmark. Breast Cancer Research and Treatment (2020) 184:891–899. https://doi.org/10.1007/s10549-020-05896-9.
47. Yang L, Wang S, Huang Y. An exploration for quantification of overdiagnosis and its effect for breast cancer screening. Chin J Cancer Res 2020;32(1):26-35
48. Katalinic, A., Eisemann, N., Kraywinkel, K., Noftz, M.R. and Hübner, J. (2020), Breast cancer incidence and mortality before and after implementation of the German mammography screening program. Int. J. Cancer, 147: 709-718. https://doi.org/10.1002/ijc.32767

49. Fann JC, Chang KJ, Hsu CY, Yen AM, Yu CP, Chen SL, Kuo WH, Tabár L, Chen HH. Impact of Overdiagnosis on Long-Term Breast Cancer Survival. Cancers (Basel). 2019 Mar 7;11(3):325. doi: 10.3390/cancers11030325. PMID: 30866499; PMCID: PMC6468420.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-046353.R3
Article Type:	Communication
Date Submitted by the Author:	07-May-2021
Complete List of Authors:	Chaltiel, Dan; Gustave Roussy, Hill, Catherine; Gustave Roussy
Primary Subject Heading:	Public health
Secondary Subject Heading:	Epidemiology, Diagnostics, Health policy, Oncology, Obstetrics and gynaecology
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Breast tumours < ONCOLOGY, PREVENTIVE MEDICINE, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

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

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

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

Correspondence to:
Dr Dan Chaltiel
dan.chaltiel@gustaveroussy.fr
Service de Biostatistique et d'Epidémiologie
Gustave Roussy
114 rue Edouard Vaillant
94805 Villejuif cedex
Tel. : +33 (0)1 42 11 42 11
Fax. : +33 (0)1 42 11 42 17

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 programs. 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 programs are serious overestimations.

1. Introduction

Many countries have a national breast cancer screening program in which all women belonging to a specific age-group are invited to have regular mammograms. These programs have been criticized, 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 programs[1]. Since it is extremely unlikely that overdiagnosis varies to such a large extent from one program to another, one needs to study possible causes for this observed heterogeneity.

2. Estimation Methods

The ideal approach to estimate the overdiagnosis rate would be to use data from randomized 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 confidence interval 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.

3. 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 (specialized 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 one-year-old children was systematically offered [4].

Such a screening program 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 program 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 one 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]).

This study showed that screening for neuroblastoma at one 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.

4. 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, i.e., 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 programs, 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 was a screening program, versus other regions,

where there was not) and during several periods (at the time of the screening program versus 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 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 program started on November 1, 1993, and the whole female population aged 50 to 69 was invited.

Each screening round lasted two years; therefore, the first found spanned from November 1, 1993 to October 31, 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 31/12/2009 at the latest. Therefore, in order to have sufficient follow-up time, Njor et al. included only patients aged 59 to 70 on November 1, 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 randomized 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 November 1, 1923-October 31, 1934), data from Funen before screening (generation November 1, 1912-October 31, 1923), and data from Danish regions without screening before the introduction of screening in Funen (generation November 1, 1912-October 31, 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 to 3 years, 4 to 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 program. However, this is still only half of the solution, as it would not take into account the effect of geography.

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 one-year generations: 1) women born in 1922 who were 71 on 1/11/1993 and, hence, were never invited to screening; and 2) women born in 1932 who were 61 on 1/11/1993 and, hence, were invited to screening.

Figure 4 shows the incidence as a function of age in these two generations, in Funen versus 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.

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 program, 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].

5. 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 to 69 and 70 to 79 [7]. Therefore, these four populations are represented by four rectangles in 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 to 69 age group during the screening period (solid orange rectangle) as compared to 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 post-screening 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 yellow 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 "post-screening" follow-up overestimates the overdiagnosis rate. The second flaw is that the screened patients in the lower orange 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.

6. 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 five years of the program (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 randomized 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 program. 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.

Jørgensen and Gotzsche estimated the expected incidence without screening by linearly extrapolating the pre-screening incidence and concluded that there was 30% to 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 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%.

7. 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 program 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 post-menopausal 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.

Statements

Contributorship

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.

Competing interests

There are no competing interests for any author

Funding

This study has received no specific funding. DC is an employee of Institut Gustave Roussy, the largest cancer hospital in France, and CH has retired from the same Institute.

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**. doi:10.1093/jnci/djx060

2 Baker SG, Prorok PC. Breast cancer overdiagnosis in stop-screen trials: More uncertainty than previously reported. *J Med Screen* 2020;:969141320950784. doi:10.1177/0969141320950784

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. doi:10.1007/s00330-015-3849-2

4 Schilling FH, Spix C, Berthold F, *et al.* Neuroblastoma screening at one year of age. *N Engl J Med* 2002;**346**:1047–53. doi:10.1056/NEJMoa012277

5 Spix C, Michaelis J, Berthold F, *et al.* Lead-time and overdiagnosis estimation in neuroblastoma screening. *StatMed* 2003;**22**:2877–92. doi:10.1002/sim.1533

6 Njor SH, Olsen AH, Blichert-Toft M, *et al.* Overdiagnosis in screening mammography in Denmark: population based cohort study. *BMJ* 2013;**346**. doi:10.1136/bmj.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. doi:10.7326/M16-0270

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. doi:10.1002/ijc.31420

9 Zahl P-H, Strand BH, Maehlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. *BMJ* 2004;**328**:921–4. doi:10.1136/bmj.38044.666157.63

10 Jørgensen KJ, Zahl P-H, Gøtzsche PC. Overdiagnosis in organised mammography screening in Denmark. A comparative study. *BMC Womens Health* 2009;**9**:36. doi:10.1186/1472-6874-9-36

11 Jorgensen KJ, Gotzsche PC. *Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends*. Centre for Reviews and Dissemination (UK) 2009. 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. doi:10.4045/tidsskr.11.0195

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. doi:10.7326/0003-4819-156-7-201204030-00005

Tables

Table 1: *Breast cancer incidence per 100,000 observed in Funen for every generation, and model estimations in the case of no screening program. Adapted from Njor et al. 2013.*

Period	Before screening	Invitation to screening		Follow-up post-screening			Cumulative over generations
		1st 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

* 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.

Figures captions

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 C. Hill. A comprehensive list of these studies is provided in Supplementary Data.*

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.
2a: *Incidence is displayed as a function of age, and generalized neuroblastoma screening takes place at one year of age. There is logically no difference in incidence between control and test regions before screening age (<1-year-old). The screening program 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.*
2b: *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).*
2c: *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).*
2d: *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.*

Figure 3: *Lexis diagrams of the Funen overdiagnosis experiment, based on Njor et al. [6]. Generations can be followed on diagonals.*
3a: *Only women born between 01/11/1923 and 01/11/1943, who were 50 to 69 at the start of screening (01/11/1993), were invited to screening.*
3b: *In order to have sufficient follow-up time (follow-up ended on 31/12/2009), 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.*
3c: *When following the screened population (S), several periods can be identified: first screenings (red), later screenings (orange), and 3 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.*

Figure 4: *Incidence of breast cancer as a function of age in Funen with screening (**green**), compared to 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 five-year age group (e.g., a dot between 50 and 55 represents the age group 50-54).
Adapted from Njor et al. 2018 [8].

Figure 5: *Incidence of breast cancer in Funen, compared to 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 (2009) 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 post-screening 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.*

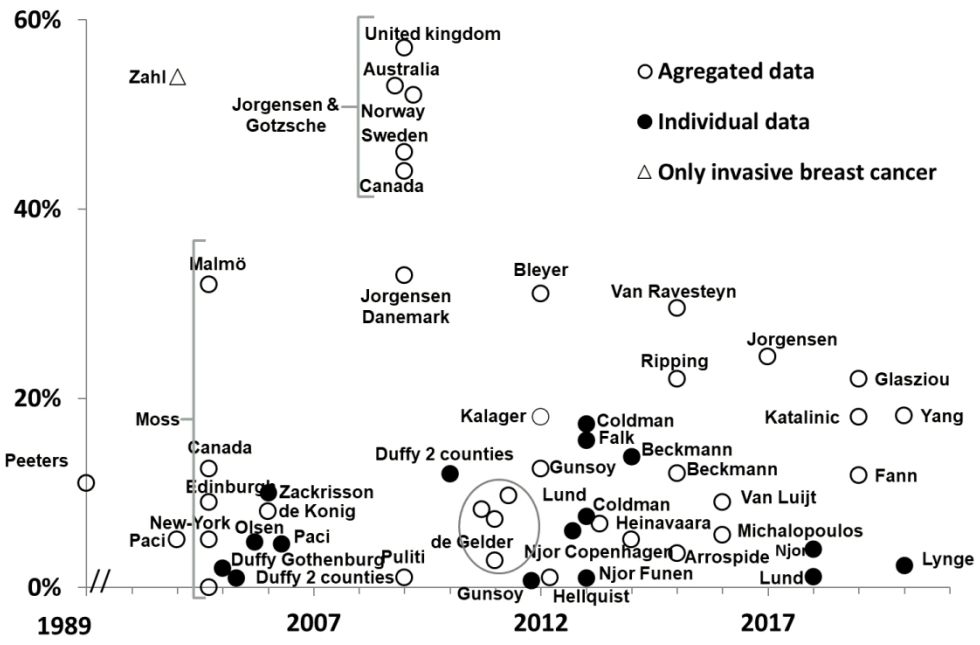

Figure 1: Published estimations of in situ and invasive breast cancer overdiagnosis (open symbols: two publications studying only invasive breast cancers quoted in the present text). Studies conducted on aggregated data give generally higher estimations of overdiagnosis than studies conducted on individual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

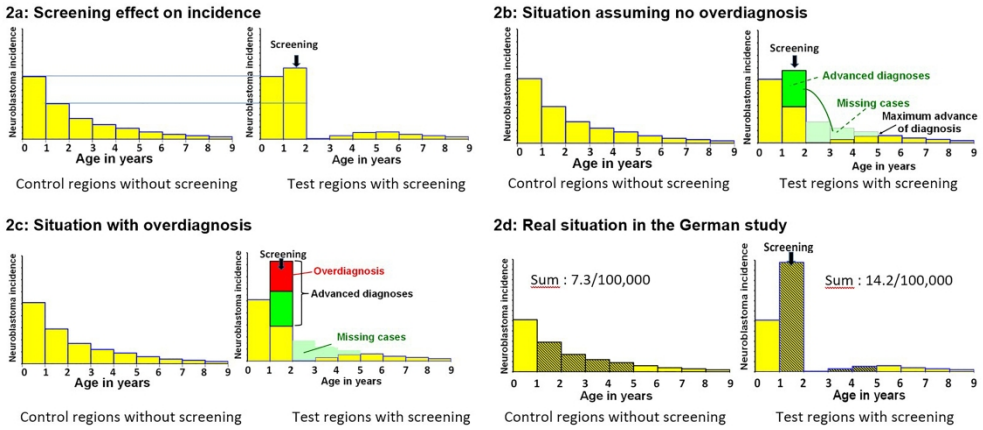

Figure 2: Overdiagnosis estimation, example of screening for neuroblastoma in Germany. Based on Schilling et al. and Spix et al. [2,3].
Control and test regions have a comparable population size, with 1.1 and 1.5 million children, respectively. Incidence is expressed in arbitrary units.
2a: Incidence is displayed as a function of age, and generalized neuroblastoma screening takes place at one year of age. There is logically no difference in incidence between control and test regions before screening age (<1-year-old). The screening program causes an increased incidence of cases immediately after screening at age 1, a decrease shortly afterward, and a return to normal at around age 5.
2b: 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).
2c: 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).
2d: 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.

435x187mm (300 x 300 DPI)

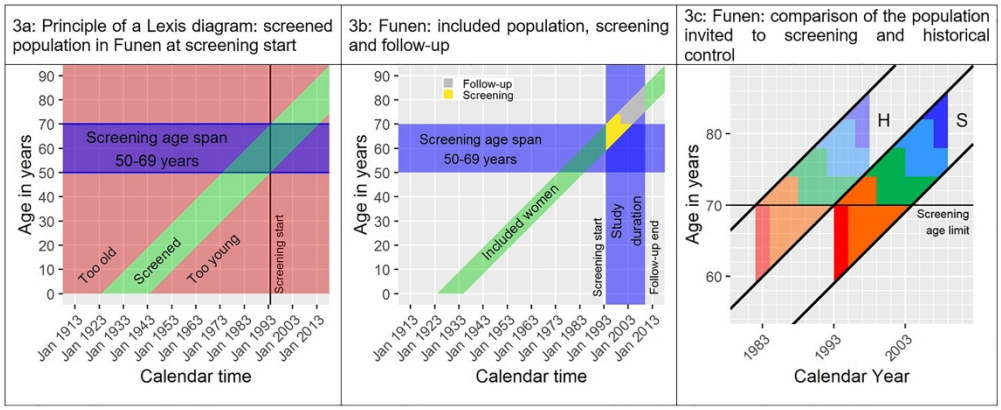

Figure 3: Lexis diagrams of the Funen overdiagnosis experiment, based on Njor et al. [4]. Generations can be followed on diagonals.
3a: Only women born between 01/11/1923 and 01/11/1943, who were 50 to 69 at the start of screening (01/11/1993), were invited to screening.
3b: In order to have sufficient follow-up time (follow-up ended on 31/12/2009), 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.
3c: When following the screened population (S), several periods can be identified: first screenings (red), later screenings (orange), and 3 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.

400x162mm (300 x 300 DPI)

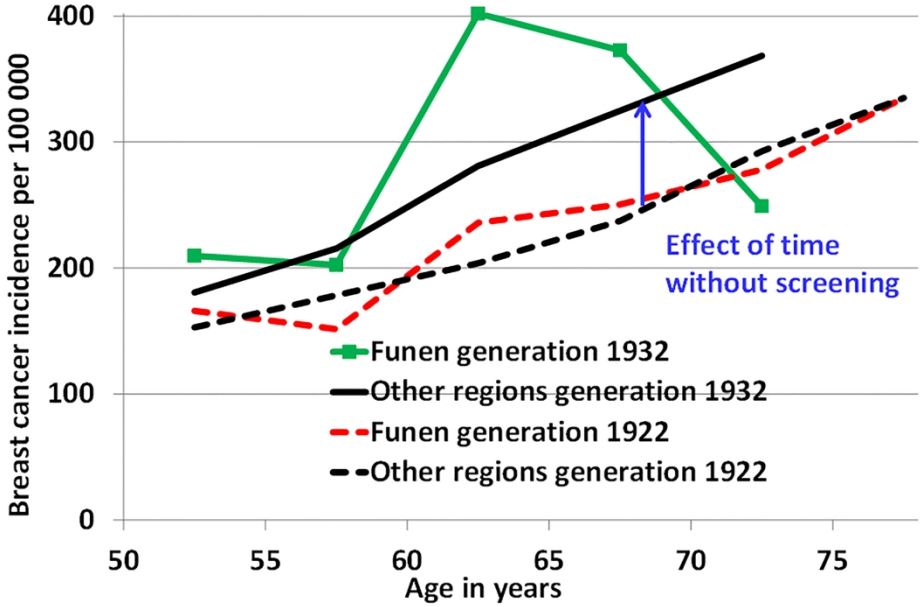

Figure 4: Incidence of breast cancer as a function of age in Funen with screening (green), compared to a historical control group (Funen in a different period, red), to a national control group (other regions, same period, solid black line), and to a historical national control group (other regions, different period, dashed black line). Adapted from Njor et al. 2018 [6].

258x167mm (300 x 300 DPI)

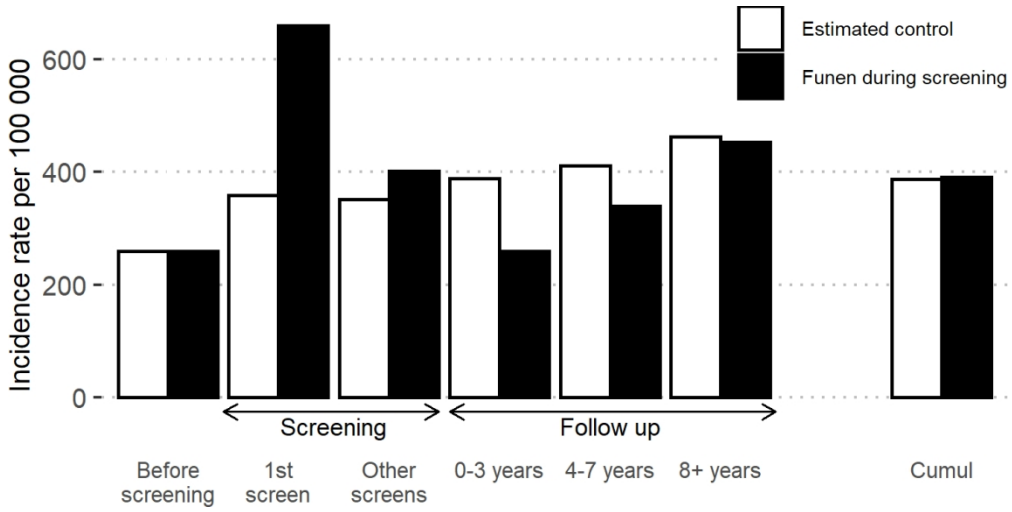

Figure 5: Incidence of breast cancer in Funen, compared to control. The incidence of breast cancer observed in Funen during screening (in black) is compared in each period to the incidence in Funen estimated by the model in the absence of screening (in grey).

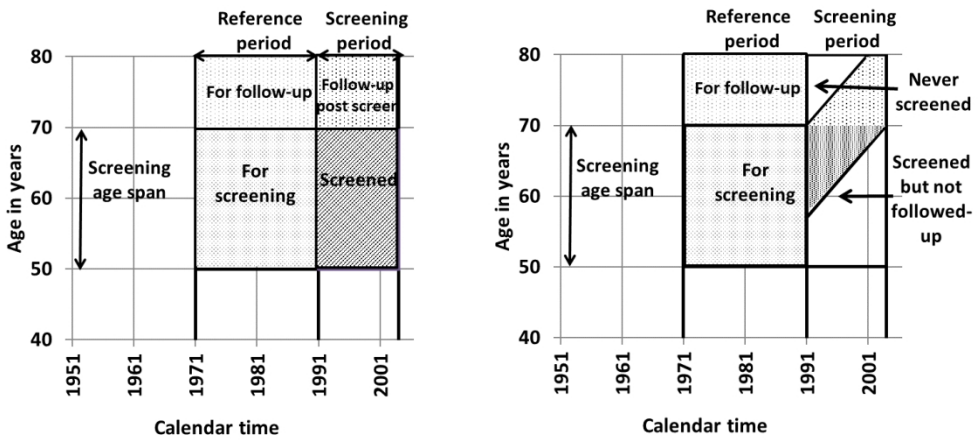

Figure 6: Data analysed by Jørgensen (2009) to estimate overdiagnosis in Denmark [9]. Under the aggregated data hypothesis, some women who were over 70 years of age at the beginning of screening, and therefore have never been screened, are included in the post-screening follow-up. These women are older 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.

Supplementary data

Comprehensive list of studies used to generate Figure 1:

1. Beckmann K, Duffy SW, Lynch J, Hiller J, Farshid G, Roder D. Estimates of over-diagnosis of breast cancer due to population-based mammography screening in South Australia after adjustment for lead time effects. J Med Screen. 2015;22(3):127–135.
2. Paci E, Miccinesi G, Puliti D, et al. Estimate of overdiagnosis of breast cancer due to mammography after adjustment for lead time. A service screening study in Italy. Breast Cancer Res. 2006;8(6):R68.
3. Paci E, Warwick J, Falini P, Duffy SW. Overdiagnosis in screening: Is the increase in breast cancer incidence rates a cause for concern? J Med Screen. 2004;11(1):23–27.
4. Morrell S, Barratt A, Irwig L, Howard K, Biesheuvel C, Armstrong B. Estimates of overdiagnosis of invasive breast cancer associated with screening mammography. Cancer Causes Control. 2010;21(2):275–282.
5. Hellquist BN, Duffy SW, Nystrom L, Jonsson H. Overdiagnosis in the population-based service screening programme with mammography for women aged 40 to 49 years in Sweden. J Med Screen. 2012;19(1):14–19.
6. Jonsson H, Johansson R, Lenner P. Increased incidence of invasive breast cancer after the introduction of service screening with mammography in Sweden. Int J Cancer. 2005;117(5):842–847.
7. Kalager M, Adami H-O, Bretthauer M, Tamimi RM. Overdiagnosis of invasive breast cancer due to mammography screening: Results from the Norwegian Screening Program. Ann Intern Med. 2012;156(7):491–499.
8. Martinez-Alonso M, Vilaprinyo E, Marcos-Gragera R, Rue M. Breast cancer incidence and overdiagnosis in Catalonia (Spain). Breast Cancer Res. 2010;12(4): R58.
9. Svendsen AL, Olsen AH, Von Euler-Chelpin M, Lynge E. Breast cancer incidence after the introduction of mammography screening: What should be expected? Cancer. 2006;106(9):1883–1890.
10. Duffy SW, Agbaje O, Tabar L, et al. Overdiagnosis and overtreatment of breast cancer: Estimates of overdiagnosis from two trials of mammographic screening for breast cancer. Breast Cancer Res. 2005;7(6):258–265.
11. Olsen AH, Agbaje OF, Myles JP, Lynge E, Duffy SW. Overdiagnosis, sojourn time, and sensitivity in the Copenhagen Mammography Screening Program. Breast J. 2006;12(4):338–342.
12. Arrospide A, Rue M, van Ravesteyn NT, et al. Evaluation of health benefits and harms of the breast cancer screening programme in the Basque Country using discrete event simulation. BMC Cancer. 2015;15(1):671.
13. De Gelder R, Fracheboud J, Heijnsdijk EAM, et al. Digital mammography screening: Weighing reduced mortality against increased overdiagnosis. Prev Med. 2011;53(3):134–140.
14. De Gelder R, Heijnsdijk EAM, Van Ravesteyn NT, Fracheboud J, Draisma G, De Koning HJ. Interpreting overdiagnosis estimates in population-based mammography screening. Epidemiol Rev. 2011;33(1):111–121.

15. de Koning HJ, Draisma G, Fracheboud J, de Bruijn A. Overdiagnosis and overtreatment of breast cancer: Microsimulation modelling estimates based on observed screen and clinical data. Breast Cancer Res. 2006;8(1):202.
16. Gunsoy NB, Garcia-Closas M, Moss SM. Modelling the overdiagnosis of breast cancer due to mammography screening in women aged 40 to 49 in the United Kingdom. Breast Cancer Res. 2012;14(6):R152.
17. Gunsoy NB, Garcia-Closas M, Moss SM. Estimating breast cancer mortality reduction and overdiagnosis due to screening for different strategies in the United Kingdom. Br J Cancer. 2014;110(10):2412–2419.
18. Seigneurin A, Francois O, Labarere J, Oudeville P, Monlong J, Colonna M. Overdiagnosis from non-progressive cancer detected by screening mammography: Stochastic simulation study with calibration to population based registry data. BMJ. 2011;343:d7017.
19. van Ravesteyn NT, Stout NK, Schechter CB, et al. Benefits and harms of mammography screening after age 74 years: Model estimates of overdiagnosis. J Natl Cancer Inst. 2015;107(7): djv103.
20. Beckmann KR, Lynch JW, Hiller JE, et al. A novel case-control design to estimate the extent of over-diagnosis of breast cancer due to organised population-based mammography screening. Int J Cancer. 2015;136(6): 1411–1421.
21. Coldman A, Phillips N. Incidence of breast cancer and estimates of overdiagnosis after the initiation of a population-based mammography screening program. CMAJ. 2013;185(10):E492–E498.
22. Duffy SWT, Olsen L, Vitak AH, et al. Absolute numbers of lives saved and overdiagnosis in breast cancer screening, from a randomized trial and from the Breast Screening Programme in England. J Med Screen. 2010;17(1):25–30.
23. Falk RS, Hofvind S, Skaane P, Haldorsen T. Overdiagnosis among women attending a population-based mammography screening program. Int J Cancer. 2013;133(3):705–712.
24. Heinavaara S, Sarkeala T, Anttila A. Overdiagnosis due to breast cancer screening: Updated estimates of the Helsinki service study in Finland. Br J Cancer. 2014;14(10):413.
25. Jørgensen KJ, Gøtzsche PC. Overdiagnosis in publicly organised mammography screening programmes: Systematic review of incidence trends. BMJ. 2009;339(7714):206–209.
26. Jorgensen KJ, Zahl PH, Gotzsche PC. Overdiagnosis in organised mammography screening in Denmark. A comparative study. BMC Womens Health. 2009; 9:36.
27. Lund E, Mode N, Waaseth M, Thalabard J-C. Overdiagnosis of breast cancer in the Norwegian Breast Cancer Screening Program estimated by the Norwegian Women and Cancer cohort study. BMC Cancer. 2013;13(1):614.
28. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: An independent review. Br J Cancer. 2013;108(11):2205–2240.
29. Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year followup for breast cancer incidence and mortality of the Canadian National Breast Screening Study: Randomised screening trial. BMJ. 2014;348:g366.
30. Moss S. Overdiagnosis and overtreatment of breast cancer: Overdiagnosis in randomised controlled trials of breast cancer screening. Breast Cancer Res. 2005;7(5):230–234.
31. Njor SH, Olsen AH, Blichert-Toft M, Schwartz W, Vejborg I, Lynge E. Overdiagnosis in screening mammography in Denmark: Population based cohort study. BMJ. 2013;346:f1064.

32. Peeters PHM, Verbeek ALM, Straatman H, et al. Evaluation of overdiagnosis of breast cancer in screening with mammography: Results of the Nijmegen Programme. Int J Epidemiol. 1989;18(2):295–299.
33. Puliti D, Miccinesi G, Zappa M, Manneschi G, Crocetti E, Paci E. Balancing harms and benefits of service mammography screening programs: A cohort study. Breast Cancer Res. 2012;14(1):R9.
34. Puliti D, Zappa M, Miccinesi G, Falini P, Crocetti E, Paci E. An estimate of overdiagnosis 15 years after the start of mammographic screening in Florence. Eur J Cancer. 2009;45(18):3166–3171.
35. Ripping TM, Verbeek AL, Fracheboud J, de Koning HJ, van Ravesteyn NT, Broeders MJ. Overdiagnosis by mammographic screening for breast cancer studied in birth cohorts in The Netherlands. Int J Cancer. 2015;137(4):921–929.
36. Waller M, Moss S, Watson J, Møller H. The effect of mammographic screening and hormone replacement therapy use on breast cancer incidence in England and Wales. Cancer Epidemiol Biomarkers Prev. 2007;16(11):2257–2261.
37. Zackrisson S, Andersson I, Janzon L, Manjer J, Garne JP. Rate of overdiagnosis of breast cancer 15 years after end of Malmo¨ mammographic screening trial: Follow-up study. BMJ. 2006;332(7543):689–692.
38. Zahl PH, Strand BH, Maehlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: Prospective cohort study. BMJ. 2004;328(7445):921–924.
39. Yen AM, Duffy SW, Chen TH, et al. Long-term incidence of breast cancer by trial arm in one county of the Swedish Two-County Trial of mammographic screening. Cancer. 2012;118(23):5728–5732.
40. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. N Engl J Med. 2012;367(21):1998–2005.
41. Michalopoulos D, Duffy SW. Estimation of overdiagnosis using short-term trends and lead time estimates uncontaminated by overdiagnosed cases: Results from the Norwegian Breast Screening Programme. J Med Screen 2016, Vol. 23(4) 192–202
42. Jørgensen KJ, Gøtzsche PC, Kalager M and Zahl PH. Breast Cancer Screening in Denmark A Cohort Study of Tumor Size and Overdiagnosis Ann Intern Med. 2017;166:313-323. doi:10.7326/M16-0270
43. Lund E, Nakamura A, Thalabard JC. No overdiagnosis in the Norwegian Breast Cancer Screening Program estimated by combining record linkage and questionnaire information in the Norwegian Women and Cancer study. European Journal of Cancer 2018; 89: 102e112
44. Njor SH, Paci E and 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 : 1287–1294
45. Glasziou PP, Jones MA, Pathirana T, et al. Estimating the magnitude of cancer overdiagnosis in Australia. Med J Aust 2020; doi: 10.5694/mja2.50455.
46. Lynge E, Beau AB, von Euler-Chelpin M, Napolitano G, Njor S, Olsen AH, Schwartz W, Vejborg I. Breast cancer mortality and overdiagnosis after implementation of population-based screening in Denmark. Breast Cancer Research and Treatment (2020) 184:891–899. https://doi.org/10.1007/s10549-020-05896-9.
47. Yang L, Wang S, Huang Y. An exploration for quantification of overdiagnosis and its effect for breast cancer screening. Chin J Cancer Res 2020;32(1):26-35
48. Katalinic, A., Eisemann, N., Kraywinkel, K., Noftz, M.R. and Hübner, J. (2020), Breast cancer incidence and mortality before and after implementation of the German mammography screening program. Int. J. Cancer, 147: 709-718. https://doi.org/10.1002/ijc.32767

49. Fann JC, Chang KJ, Hsu CY, Yen AM, Yu CP, Chen SL, Kuo WH, Tabár L, Chen HH. Impact of Overdiagnosis on Long-Term Breast Cancer Survival. Cancers (Basel). 2019 Mar 7;11(3):325. doi: 10.3390/cancers11030325. PMID: 30866499; PMCID: PMC6468420.