Overdiagnosis and “underdiagnosis” in BreastScreen Norway

Kaitlyn M Tsuruda

PhD Thesis

UiO : Faculty of Medicine
University of Oslo

BreastScreen Norway

Cancer Registry of Norway

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-888-5

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

This PhD project was funded by the Dam Foundation, through the Norwegian Breast Cancer Society (grant number 2018/FO201362)

Acknowledgements

This work would have been much more challenging without institutional support from the Cancer Registry of Norway; the University of Oslo; and the national research school in population-based epidemiology, EPINOR. I am also grateful to have received funding for this project from the Dam Foundation via the Norwegian Breast Cancer Society.

I would like to give my heartfelt thanks to Solveig Hofvind, my enthusiastic and positive supervisor. Thank you for giving me the opportunity to become a PhD student, Solveig. And thank you for sharing your extensive knowledge about breast cancer screening with me. It has been a pleasure to learn from you. You helped me push myself when I was ready, and pulled the reins when necessary. Thank you for being a constant source of encouragement and always keeping my personal and professional interests in mind.

I would also like to give special thanks to Marit Veierød, my enthusiastic and positive co-supervisor. Marit, you have been a reliable sounding board for my more creative statistical ideas (and my not-so-creative ones too). Thank you for always pointing me in the right direction when you were unsure yourself. I am grateful for having had the opportunity to learn more about statistics under your guidance.

Solveig and Marit, you have made an excellent team. I have felt well taken care of throughout my PhD and am so fortunate to have supervisors as kind and knowledgeable as you both. Thank you for your insights into research, life, and knitting. It has been a pleasure to have you both on my side this whole time.

To my co-authors, Lars Akslen, Sameer Bhargava, Solveig Roth Hoff, Nehmat Houssami, Tone Hovda, Gunhild Mangerud, and Gunvor Waade: thank you for being a part of this adventure and sharing your experience and knowledge with me. I have learned from all of you and hope we can work together again.

Thank you also to everyone who spent time reading my writing and gave feedback on any text included in – or deleted from – this thesis and its accompanying articles. I appreciate that you took the time to help. My thanks also go to Annette and Iselin for keeping me company (and accountable) during hours and hours and hours of “shut up and write”.

Thank you to all my colleagues at the Registry for your support and insights, and for helping me think through challenging questions. In this regard, I’d like specifically thank the GyM coders and my coworkers in the breast screening section for being especially great to work

with. My officemates Anders, Camilla, Jonas, Marthe, Silje, and Åsne deserve special thanks for supporting me on this journey and making “Teamkontoret” the best place to work.

To Mo Abdolell and Jennifer Payne, thank you for giving me the opportunities, experience, and encouragement earlier in my career to help me get to, and through, this PhD. Your mentorship has meant, and continues to mean, so much.

Thank you to my family and friends near and far for supporting me and cheering me on through another degree’s worth of courses, exams, and deadlines. I’m grateful for the extra encouragement you’ve given me these past few months and hope I see you soon.

Last, but by no means least, thank you Torgeir. It’s been a time, but you have been unconditionally supportive and maintained a positive attitude while enduring the vicissitudes of PhD-supporter life and a shared work from home situation. Thank you for always believing in me. And editing my writing. And making more than your fair share of dinners. What could be more agreeable than that?

Kaitlyn Tsuruda

Oslo, February 2021

Contents

Project motivation

Mammographic screening for breast cancer was established in many European, North American, and Oceanian countries over 20 years ago. This PhD has allowed me to contemplate the impact of breast cancer screening as public health initiative and the social and political history that has shaped it into what it is today. I am grateful for having had the opportunity to explore a field that truly interests me.

Organized mammographic screening was borne out of a desire to help women survive a deadly disease and has been shown to reduce deaths from breast cancer. Most scientists and policymakers familiar with the large body of evidence about mammographic screening agree that its potential benefits outweigh its potential risks. However, polarized views on this topic contribute to it being a recurring point in a highly charged debate. In my opinion, part of the reason this debate is so heated is because there are many who have been impacted by breast cancer, whether directly or indirectly, and truly *care* about women's health.

My thesis investigates some potential risks associated with organized mammographic screening, including overdiagnosis and a contrasting circumstance we have called "underdiagnosis". The intangible nature of these risks can make them contentious topics. This is particularly true of overdiagnosis. In light of this, before I present my work from the last three years, I would like to note that I am employed at the Cancer Registry of Norway in the breast screening section, which administers BreastScreen Norway. I worked in this section for 13 months prior to starting my PhD. My salary comes from a PhD stipend granted by The Dam Foundation via the Norwegian Breast Cancer Society.

Overdiagnosis and "underdiagnosis" are issues that affect how women think about and value screening. There is a need for additional knowledge about these topics internationally, and nationally using Norwegian data. This project investigated concrete aspects of overdiagnosis and "underdiagnosis" that are not well-described in the peer-reviewed literature and largely did so from a Norwegian perspective. The aim of this project was to generate useful information about the potential risks associated with screening. I hope that the results of this research can help us move toward our shared goal of continually improving mammographic screening services for women and reducing the burden of breast cancer.

Publications included in this thesis

Paper 1

Tsuruda KM, Hofvind S, Akslen LA, Hoff SR, Veierød MB. Terminal digit preference: a source of measurement error in breast cancer diameter reporting. Acta Oncol. 2020;59(3):260-7.

Paper 2

Tsuruda KM, Hovda T, Bhargava S, Veierød MB, Hofvind S. Survival among women diagnosed with screen-detected or interval breast cancer classified as true, minimal signs, or missed through an informed radiological review. Eur Radiol. 12 November 2020. https://doi.org/10.1007/s00330-020-07340-4. Online ahead of print.

Paper 3

Tsuruda KM, Veierød MB, Houssami M, Waade GG, Mangerud G, Hofvind S. Women's conceptual knowledge about breast cancer screening and overdiagnosis in Norway: a cross-sectional study. *Submitted.*

Terms and Abbreviations

AJCC	American Joint Committee on Cancer
CI	Confidence interval
ER	Estrogen receptor
HER2	Human epidermal growth factor receptor 2
ICH	Immunohistochemistry
ISH	In situ hybridization
IP	Internet protocol
MRI	Magnetic resonance imaging
PR	Progesterone receptor
SD	Standard deviation
TNM	Tumour-node-metastasis
cTNM	Clinical tumour-node-metastasis stage
cT	Clinical tumour stage
pTNM	Pathological tumour-node-metastasis stage
pT	Pathological tumour stage
ypTNM	Pathological tumour-node-metastasis stage for women who receive neoadjuvant treatment
ypT	Pathological tumour stage for women who receive neoadjuvant treatment

Summary

Attending mammographic screening may help detect breast cancer in an early stage. It may also reassure women that they do not have mammographic signs of the disease. However, participating in screening can also have negative consequences for some, such as overdiagnosis or “underdiagnosis”. Accurate diagnosis and staging is crucial for ensuring that women receive personalized and effective treatment for their disease and can help mitigate some negative consequences associated with over- or “underdiagnosis”. This thesis addresses lesser-studied aspects of these topics with the aim of generating knowledge about the potential risks associated with organized mammographic screening.

In this thesis, overdiagnosis was defined as the diagnosis of a slow growing breast cancer that *would never* present symptomatically during a woman’s lifetime, no matter how long she lived. This term and definition are commonly used in screening. “Underdiagnosis”, however, is not a common term. In this thesis, I defined “underdiagnosis” as the so-called inverse of overdiagnosis: “failing” to diagnose a breast cancer in a woman whose cancer *would* present symptomatically during her lifetime. The definition of overdiagnosis is counterfactual and cannot directly be observed in individuals. The definition of underdiagnosis can also be counterfactual in certain situations. As a result, observational studies investigating these topics are typically based on a set of assumptions.

In the first paper, my co-authors and I described the distribution of tumour diameters reported to the Cancer Registry of Norway. International guidelines specify that tumour diameters should be reported to the nearest millimetre and it is generally assumed that these measurements are accurate. The results of this study showed that radiologists and pathologists have a tendency to round tumour diameter measurements to the nearest whole or half-centimetre value and that this can lead to reporting a cancer as having a lower (but not higher) T stage than its size or spread would dictate. This understaging has the potential to lead to undertreatment. Additional analyses indicated that preferential rounding disproportionally affected women with cancers diagnosed outside of organized screening compared to women diagnosed through BreastScreen Norway. The prevalence of potential understaging due to this type of rounding is difficult to estimate, but I hope that shining a light on this source of measurement error will make breast radiologists and pathologists more cognizant of the impact it can have in their daily practice.

In the second paper, we described how a diagnosis of breast cancer that was or was not retrospectively visible on a woman’s previous screening examination affected her survival.

We compared “missed” cancers (visible in retrospect but not diagnosed at a woman’s previous screening examination) to “true” cancers (that did not have retrospectively visible mammographic signs on a woman’s previous screening examination). Panels of five radiologists made these classifications through a consensus-based review of screening and diagnostic mammograms with access to pathological reports. We considered missed cancers that were diagnosed between screening examinations due to clinical symptoms as underdiagnosed. Further, we considered whether missed cancers that remained asymptomatic and were diagnosed at a woman’s subsequent screening examination could be underdiagnosed if the woman would have benefited from earlier detection. We also considered whether these missed screen-detected cancers could be overdiagnosed if they represented indolent disease.

Missed cancers (including underdiagnosed cancers) are hypothesized to generally have a more favourable prognosis than true cancers because they are thought to be slower-growing. The results of this study did not show a difference in overall survival between missed and true cancers. Effective treatment options may explain this finding, but this result may also be due to low statistical power. The discussion of this paper highlights some challenges associated with secondary use of data from review studies.

In the third paper, my co-authors and I used an online survey to explore women’s conceptual knowledge about breast cancer screening and overdiagnosis. The survey targeted women aged 45–75 and asked them about the breast cancer mortality benefit, false positive screening examinations, and overdiagnosis associated with screening. Most participants in this cross-sectional study chose the correct answers to questions about the first two topics. The proportion of correct responses to questions about overdiagnosis was lower. Responses to individual questions about false positive screening examinations and overdiagnosis suggested that women may confuse these two topics. This is one of the first studies to document Norwegian women’s knowledge about overdiagnosis. Future research using qualitative methods may be warranted to better understand women’s knowledge about false positives versus overdiagnosis. Results from this type of study could be used to improve the quality of Norwegian-language information available to women.

The results of the three studies in this thesis could have implications for clinicians and administrators working in organized screening programs and will be important to consider going forward to develop a multifaceted understanding of the potential harms associated with organized mammographic screening.

Norwegian summary

Deltagelse i Mammografiprogrammet kan bidra til å oppdage og diagnostisere kreft i et tidlig stadium av sykdomsutviklingen og kan forsikre kvinner om at de ikke har mammografiske tegn til brystkreft. Samtidig utgjør overdiagnostikk og «underdiagnostikk» to potensielle risikoer. Denne avhandlingen omhandler aspekter ved disse risikoene, som det foreløpig er forsket relativt lite på. Målet er å få større innsikt i potensielle ulemper knyttet til organisert mammografiscreening.

I dette gradsarbeidet ble overdiagnostikk definert som diagnostisering av sakte-voksende brystkreft, som aldri ville gitt symptomer i kvinnens levetid. «Underdiagnostikk» ble definert som en forsinket brystkreft diagnose, der tidligere diagnostikk ville ført til en mer gunstig prognose.

I den første studien beskriver mine medforfattere og jeg fordelingen av svulststørrelser som blir rapportert til Kreftregisteret. Ifølge internasjonale retningslinjer skal svulstdiameter skal måles og rapporteres til nærmeste millimeter. Det er generelt stor tillit til disse målingene og rapporteringene. Studien viser likevel at radiologer og patologer har en tendens til å rapportere avrundede mål for svulstdiameter, der målet avrundes til nærmeste hele eller halve centimeter. Vår studie viste at denne avrundingen kan føre til at svulster blir klassifisert i et lavere stadium enn de virkelig er, noe som kan få følger for behandlingen kvinner blir tilbudt og gjennomgår.

Den andre studien er knyttet til over- og underdiagnostikk. Her beskriver vi hvordan kvinners overlevelse etter en brystkreftdiagnose påvirkes av hvorvidt svulsten enten var eller ikke var synlig på foregående screeningundersøkelse. Vi sammenlignet overlevelse blant kvinner som fikk diagnostisert brystkreft som hadde mammografiske tegn på forrige screeningundersøkelse, men som ikke ble diagnostisert («oversett» kreft) med de som fikk diagnostisert brystkreft som ikke viste tegn på forrige screeningundersøkelse, men som ga symptomer og ble diagnostisert i perioden mellom to screeningundersøkelser («sann» kreft). Fem brystradiologer klassifiserte tilfellene i en konsensusbasert regranskning av screening og diagnostiske mammogrammer. Informasjon knyttet til diagnostikk, utredning og behandling var tilgjengelig for regranskerne. Brystkreft som i regranskingen ble klassifisert som oversett kan være underdiagnostisert dersom den ble diagnostisert på grunn av symptomer, mens de som forble symptomfri og diagnostisert i neste screeningrunde, kan være overdiagnostisert.

Oversette kreftsvulster (inkludert underdiagnostisert kreft) antas å ha en mer gunstig prognose enn sanne kreftsvulster. Grunnen er at de antas å vokse mer langsomt. Vi fant ingen forskjeller i generell overlevelse for kvinner diagnostisert med svulster som

retrospektivt var klassifisert som oversette eller sanne. Effektiv behandling kan forklare dette funnet, men det er også mulig at resultatet skyldes manglende statistisk styrke. I studien diskuterer vi utfordringer knyttet til gjennomføring og tolkning av resultater fra regranskningsstudier.

I den tredje studien undersøker vi kvinners kunnskap om mammografiscreening og overdiagnostikk. Vi gjennomførte en web-basert spørreskjemaundersøkelse blant kvinner i alderen 45–75 år. Spørsmålene var knyttet til brystkreftdødelighet, falske positive screeningundersøkelser og overdiagnostikk. De fleste kvinnene valgte riktige svar på spørsmålene om brystkreftdødelighet og falske positive, mens andelen riktige svar på spørsmål om overdiagnostikk var noe lavere. Responsen kan tyde på at kvinnene blander begrepene falske positive screeningundersøkelser og overdiagnostikk. Studien er blant de første som dokumenterer kvinners kunnskap om overdiagnostikk i Norge.

Background

Development of breast cancer

Whether invasive breast cancer is limited to the breast or is a systemic disease throughout the body has been questioned for centuries (1). Despite this long history of research, we do not fully understand the natural history of breast cancer, that is, how the disease progresses. This remains an active field of research (2).

As recently as 2006, researchers questioned whether cancerous cells limited to milk ducts in the breast, so-called ductal carcinoma *in situ*, always develop into invasive breast cancer – in other words, whether this is an obligate precursor of invasive disease (3). However, it is now accepted that ductal carcinoma *in situ* is a heterogeneous disease and that not all ductal carcinoma *in situ* will progress to invasive breast cancer if a woman lived indefinitely (3-5). The research presented in this thesis is epidemiological – not biological – in nature and is limited to invasive breast cancer, hereafter referred to as breast cancer.

Historically, breast cancer was considered a progressive and fatal disease (1). The Halstedian model of breast cancer from the late 1800s proposed that breast cancer grew from a single cell that spread to regional lymph tissues before spreading to other body parts (1, 6). The Fisherian model of breast cancer proposed in the 1980s posited that breast cancer was a systemic disease that could spread through the blood and that its aggressiveness was predetermined at its onset based on innate tumour characteristics (2). Our current understanding of breast cancer suggests that it is not a single disease, but many different diseases (so-called “subtypes”) that can range from indolent and slow growing to aggressive with high potential to spread to other organs (Figure 1) (4, 7).

Breast cancers generally have a long natural history and risk for late recurrence (4). Tumour biology, including characteristics such as subtype and size at diagnosis, is an important indicator of a woman’s prognosis and the probability that her cancer will respond to certain treatment. Some tumour characteristics are thought to influence or reflect the potential for tumour growth.

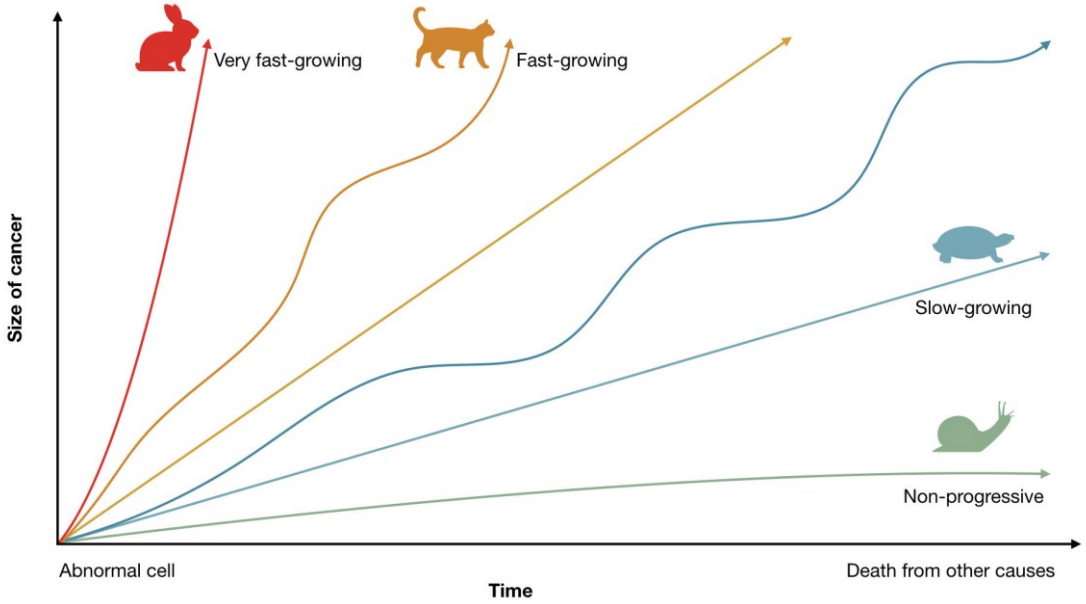


***Figure 1**: Possible growth trajectories for breast cancer, adapted from an infographic courtesy of The National Cancer Institute.*[1]

Tumour characteristics and staging

The Union for International Cancer Control and American Joint Committee on Cancer (AJCC), provide advice about breast cancer staging using the tumour-node-metastasis (TNM) system for malignant tumours (8, 9). The TNM staging system was first published in 1959 and provides information about a patient's prognosis, including whether they would benefit from systemic therapy (9). This staging system defines breast cancer stages based on anatomic features of the cancer, namely a description of the primary tumour (T), regional lymph node involvement (N), and the absence or presence of distant metastases (M) (9). Clinical information obtained from physical examination, imaging, and/or pathology prior to initiating treatment is used to determine the clinical stage (cTNM) and informs decisions regarding treatment before surgery (neoadjuvant treatment) and surgical treatment. Pathological information gained from surgical samples and the aforementioned clinical information is used to determine the pathological stage (pTNM) and informs post-surgery (adjuvant) treatment and follow-up. If a woman receives neoadjuvant treatment, a post-treatment pathological stage (ypTNM) is assigned instead of pTNM. This is to indicate that the residual disease after treatment was assessed, not the untreated disease.

[1]Original infographic available at https://prevention.cancer.gov/news-and-events/infographics/what-cancer-overdiagnosis

Breast cancer tumour categories in the TNM system (so-called T stage) are denoted T1 to T4. Categories T1–T3 are based on maximum (clinical or histopathological) tumour diameter, where T1: ≤20 mm, T2: >20–50 mm, and T3: >50 mm. T4 indicates a tumour with direct extension to the chest wall or skin (8-10). Lymph node categories (N stage) indicate whether regional lymph node metastases are present, and to what extent. N0 indicates no nodal involvement, and N1–N3 indicate increasing nodal involvement at increasingly distant sites. The categories for distant metastases indicate whether they are absent (M0) or present (M1). Breast cancer stages are determined based on the overall TNM classification, as shown in Table 1.

***Table 1**: Breast cancer TNM classifications and corresponding stage classifications. See reference (8) or (9) or for further details.*

TNM classification	Breast cancer stage
T1N0M0	Stage I
T1-T2N1M0 T2-T3N0M0	Stage II
T1-T2N2M0 T3N1-N2M0 T4N0-N2M0 Any T, N3M0	Stage III
Any T, any N, M1	Stage IV

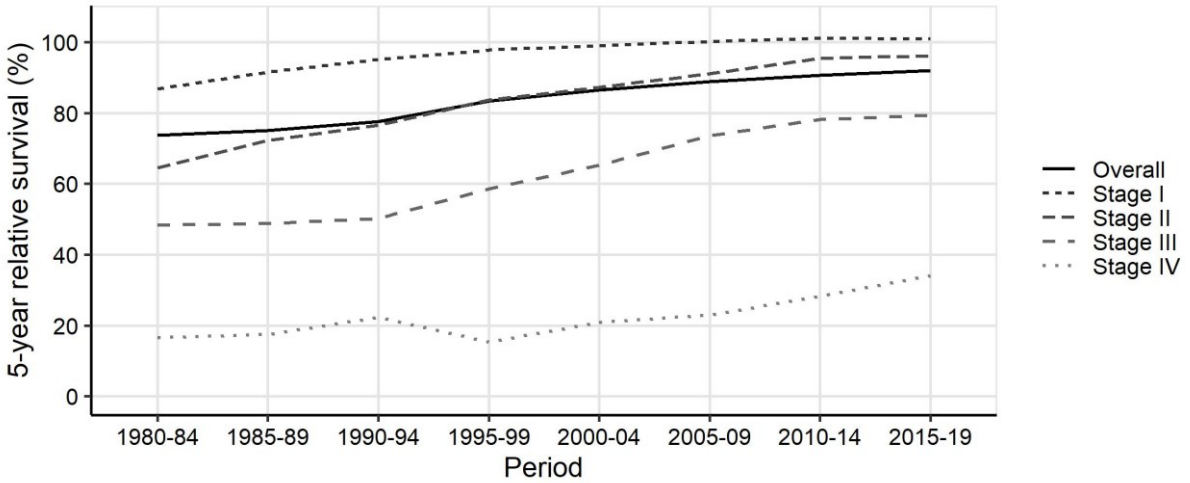


***Figure 2**: Five-year relative survival for women diagnosed with breast cancer in Norway by stage and diagnosis period. See reference (11) for further details.*

Among all women diagnosed with breast cancer during 2015–2019, the probability of surviving five years relative to a similar group of women without breast cancer (relative survival) was 92.0% (11). The probability of survival is higher for earlier stage cancers than later stage cancers: during 2015–2019, the five-year relative survival was 100.9% for women with stage I breast cancer, but 34.0% for women with stage IV breast cancer (Figure 2) (11).

Although TNM classifications and staging provide useful clinical information, knowledge that breast cancer is a heterogeneous disease makes it evident that tumours with a similar TNM classification or stage may not behave similarly over time. Certain tumour characteristics can provide information about a woman's prognosis, so-called prognostic characteristics, and others can provide information about how a woman will respond to certain treatment, so-called predictive characteristics. Commonly reported tumour characteristics include histological grade, estrogen receptor (ER) and progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) status, and Ki67 proliferation. Table 2 highlights selected prognostic and predictive tumour characteristics.

Table 2: *Prognostic and predictive tumour characteristics.*

Tumour characteristic	Prognostic?	Predictive?	Further reading (reference)
Maximum tumour diameter	Yes	-	(4, 9)
Lymph node involvement	Yes	-	(4, 9)
Histologic Grade	Yes	-	(4, 9)
ER status	Yes	Yes	(4, 9)
PR status	-	Yes	(4)
HER2 status	Yes	Yes	(4, 9)
Ki67 proliferation	Yes	-	(4, 12)
Subtype	Yes	Yes	(9, 12)

Tumour characteristics are also used to define molecular subtypes. Breast cancer subtyping provides important information about a patient's prognosis and potential response to treatment (9, 12, 13). Clinical breast cancer subtypes include: Luminal A-like, Luminal B-like, HER2 positive and triple negative. These are defined in Table 3. The majority of breast cancers are ER positive (i.e. luminal) cancers. These are associated with a more favourable prognosis than ER negative cancers (i.e. HER2 positive and triple negative cancers) (4) Triple negative breast cancers in particular are often associated with a poor prognosis (14).

***Table 3**: Clinico-pathologic surrogate definitions of intrinsic subtype based on immunohistochemical analyses (12).*

Subtype	Definition
Luminal A	ER and PR positive, HER2 negative, and Ki67 proliferation < 20%[a]
Luminal B[b]	ER positive, PR negative and HER2 negative; or ER positive, HER2 negative and Ki67 proliferation ≥ 20%[a]; or ER positive and HER2 positive
HER2 positive	ER and PR negative, and HER2 positive
Triple negative	ER, PR, and HER2 negative

[a] The exact threshold for Ki67 proliferation varies between laboratories, but 20% has been put forward as a suitable threshold. See reference (12) for more details.
[b] The Luminal B subtype is often subdivided as Luminal B (HER2 negative) and Luminal B (HER2 positive) depending on a woman's HER2 status.

Incidence and risk factors

Incidence describes the number of new cases in a certain population during a given period. It can be measured as a count, proportion, or rate. In Norway, 3726 women were diagnosed with breast cancer during 2019 (15). The median age at diagnosis is 62 years and, on average, one in eleven women will be diagnosed with breast cancer before she turns 75 (15).

Breast cancer is a multifactorial disease with many non-modifiable and modifiable risk factors and causal agents; age and sex are the most influential (4, 15). Other non-modifiable factors that increase breast cancer risk include certain hereditary factors, being younger at menarche, and being older at menopause, among others (4). Modifiable risk factors include being older at first full-term pregnancy, having increased mammographic breast density, engaging in low levels of physical activity, being overweight or obese as an adult (postmenopausal breast cancer only), and using hormone replacement therapy, among others (4, 16, 17).

On a population level, roughly 20% of breast cancers worldwide are attributable to physical inactivity and overweight/obesity (4). Trends in hormone replacement therapy use also affect breast cancer incidence on a population level. In Norway, an estimated 28% of women aged 45–64 were using these drugs during 1996–1997 and 57% of women aged 55–59 had used them at least once (so-called ever users) (18). This is thought to have caused 27% of breast cancers among women in this age group during 1996–1998, and 16.6% of ductal carcinoma in situ and invasive breast cancers among women aged 47–63 during 2003–2008 (19, 20). Hormone replacement therapy use in Norway decreased in the mid-2000s after it was established as a risk factor for breast cancer (21-24). During 2019, 16% of women aged 45 and older used hormone supplements associated with menopause (25).

On an individual level, only 5–10% of women diagnosed with breast cancer have heritable genetic mutations that put them at a high lifetime risk (4). Thus, most women diagnosed with the disease have a low or average lifetime risk of developing breast cancer. However, most risk factors confer only modest change in women's risk of breast cancer and individualized risk models based on established risk factors have only demonstrated moderate discriminatory and calibration accuracy (26).

Screening for breast cancer

Early detection and the principles of screening

When organized breast cancer screening with mammography (mammographic screening) started in the 1960s and 70s, modifiable risk factors for breast cancer were not well understood. Decades later, we still do not know how to entirely prevent the disease among average-risk women. However, after reaching a certain (small) size, many asymptomatic breast tumours can be identified on a mammogram by skilled radiologists. The time between when an asymptomatic breast cancer can first be detected with mammography and when it would cause clinical symptoms is called the *sojourn time* (27). Mammographic screening helps radiologists to find and diagnose small, asymptomatic breast cancers during their sojourn time. Detecting early stage breast cancer in women of a certain age through periodic, population-based mammographic screening is considered the best way to reduce breast cancer mortality. It is thought that early stage breast cancer will respond more readily to treatment and lead to fewer deaths than late stage breast cancer (4).

Most women do not have breast cancer. Therefore, many need to be screened to detect one woman with the disease. Women without breast cancer cannot benefit from early detection and screening program administrators must balance the population-based effects of screening against the individual-level effects. This is challenging because programs are typically monitored using epidemiological methods that measure average (i.e. population-level) effects (28). Although epidemiological studies have generated a wealth of knowledge about screening, this knowledge is not always directly applicable to individuals.

In 1968, before any results were published regarding the population-level effects of organized mammographic screening for breast cancer, the World Health Organization released "Principles and practice of screening for disease" (29). These ten criteria served as guidelines for the establishment of organized screening programs around the world, including mammographic screening programs (Table 4). Additional criteria have been proposed and introduced in the years since, including those put forward by the Norwegian Directorate of Health in 2014 (Table 5) (30-32).

An important aspect of screening is that the screening examination itself does not provide a diagnosis. In the case of mammographic screening, mammograms are used to identify women with abnormal findings that could benefit from being recalled for diagnostic testing that often involves ultrasound and other supplemental imaging. Needle biopsy is necessary to make a definitive diagnosis.

***Table 4**: Ten principles for early detection from the World Health Organization (29).*

1	The condition sought should be an important health problem
2	There should be an accepted treatment for patients with recognised disease
3	Facilities for diagnosis and treatment should be available
4	There should be a recognisable latent or early symptomatic stage
5	There should be a suitable test or examination
6	The test should be acceptable to the population
7	The natural history of the condition, including development from latent to declared disease, should be adequately understood
8	There should be an agreed policy on whom to treat as patients
9	The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole
10	Case-finding should be a continuing process and not a “once and for all” project

***Table 5**: Six additional principles for organized screening programs put forward by the Norwegian Directorate of Health (32).*

1	The health benefits should outweigh the harms
2	The protection of personal privacy and adherence to the law be ensured
3	The program should be ethically acceptable
4	Information about participation should be evidence-based and empower making an informed choice about participation
5	The program should be cost-effective
6	There should be a plan for program administration, quality assurance and evaluation

Effect of screening on breast cancer incidence

Breast cancer incidence has been increasing in Norway for decades (Figure 3) (15). The increase in incidence from 1996–2005 can be partly attributed to changes in background risk, improvements in screening and diagnostic tools, and increased breast cancer awareness, but also to the introduction of organized screening.

When a cohort of women enters a screening program, many of their breast cancers are detected earlier, thereby inflating the incidence in that group. This abrupt increase is called a prevalence peak and is shown around age 50 in Figure 4A. In an ideal program, subsequent rounds of screening will help detect existing (prevalent) and newly occurring (incident) breast cancers earlier than they would have been without screening. The incidence rate is therefore

expected to be higher among screened women than similarly aged non-screened women. When a cohort of women leaves a screening program, their incidence is expected to be lower than a comparable non-screened cohort. This is because many of the prevalent breast cancers in the screened cohort have already been detected through screening and incident cancers can take some time to develop before they become symptomatic. The drop in incidence among women who leave screening is called a compensatory drop. In Figure 4A, women leave screening at age 70 and this drop is visible thereafter.

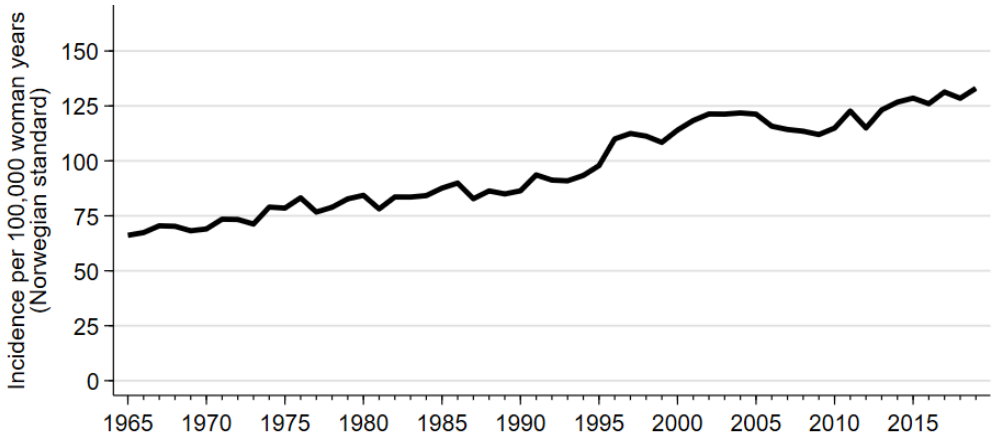


***Figure 3**: Observed breast cancer incidence among women of all ages in Norway from 1965–2019.[2]*

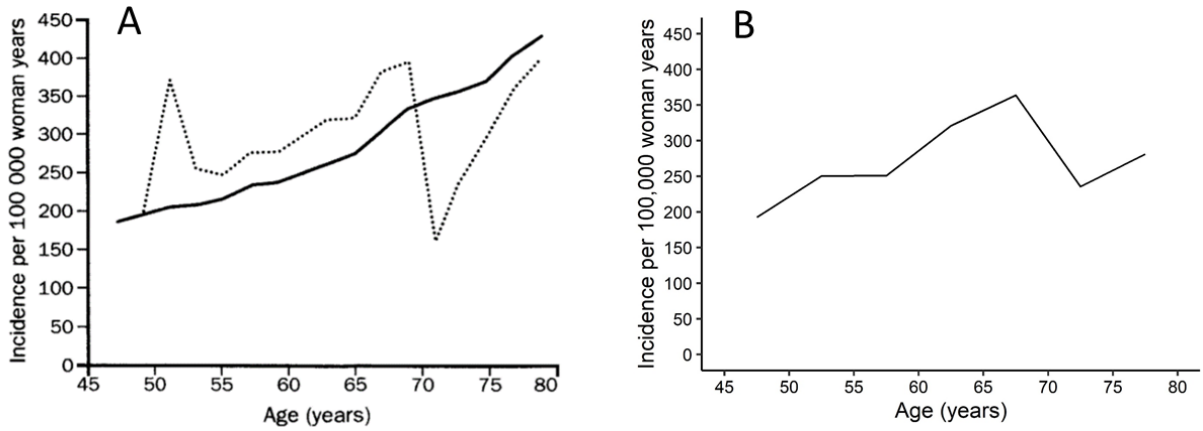


***Figure 4: (A)** Expected breast cancer incidence in two-year age groups for screened (dotted line) and non-screened (solid line) women[3], and **(B)** observed breast cancer incidence among all women aged 45–79 in Norway during 2016.[4]*

---

[2] Figure adapted with permission from Cancer Registry of Norway. Cancer in Norway 2019 - cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2020.
[3] Reprinted from The Lancet, Vol. 343, Boer R, Warmerdam R, de Koning H, van Oortmarssen G, Extra incidence caused by mammographic screening, p. 979, Copyright (1994), with permission from Elsevier.
[4] Figure data from Danckert B, Ferlay J, Engholm G, et al. (2019). NORDCAN: Cancer incidence, mortality, prevalence, and survival in the Nordic countries, Version 8.2 (26.03.2019). Association of the Nordic Cancer Registries. Danish Cancer Society. Available from: http://www.ancr.nu, accessed on 30 December 2020.

In countries with population-based screening programs, factors such as screening at private clinics (particularly among women outside the target age for screening) and irregular attendance or non-participation in organized programs can make it difficult to observe the prevalence peak or compensatory drop. Figure 4B shows the age specific incidence rates for women aged 45–79 in Norway during 2016. The prevalence peak that should be visible around age 50 is not as obvious as the compensatory drop around age 70.

Mammographic screening can also affect breast cancer incidence by leading to the detection of slow growing cancers that would not cause any symptoms during a woman's lifetime (33). These cancers are often called "overdiagnosed" and increase the number of breast cancer diagnoses made compared to a situation with no screening. See page 31 for more details.

The origins of modern-day screening

Screening for breast cancer with mammography gained popularity during the early 1960s and was borne out of a desire to help women survive a deadly disease (recall that the Halstedian model of disease progression prevailed at that time) (34). Reports of diagnostic mammography go back as far as the 1930s, but routine mammographic screening was not reported on before the 1950s (34, 35). General analogue (film) x-ray equipment was used for mammography until the first dedicated mammography device was developed in the 1960s (34). Contemporary mammographic screening uses digital x-ray equipment designed specifically for breast imaging.

In 1963, a randomized controlled trial was started in New York to determine whether mammographic screening together with clinical breast examination could reduce breast cancer mortality among women aged 40–64 (36). Several additional randomized controlled trials followed (4). Together these trials provided the evidence base for implementing organized mammographic screening in countries around the world.

The utility of organized mammographic screening has been debated publicly since at least the late 1970s. This debate started because the Breast Cancer Detection Demonstration Project in the United States was offering mammographic screening to women age 35–74, based on evidence created from women aged 40–64 who participated in the New York trial (37). Later discussions in the 1990s questioned the quality of the Canadian National Breast Screening Studies and the mortality results from Swedish trials (38-41). The randomized trials generally provided a high level of scientific evidence for the questions they were designed to answer, but these debates illustrate how these trials cannot provide definitive answers for all the questions we might have about screening. Further questions arise when considering improvements in mammographic imaging since the 1960s.

Epidemiological considerations

Much of the new knowledge developed about screening comes from observational epidemiological studies. The precision and validity of results, particularly from observational studies, is often discussed or debated. Study precision refers to a "*relative lack of random error*" or "*the quality of being sharply defined or stated"* (27). Epidemiologists divide the concept of validity into internal and external validity. Internal validity relates to the amount of systematic error or bias in a study and refers to how well a study's results relate to the population of interest (27, 42). External validity refers to how well a study's results can be generalized to other populations or settings (27, 42).

Epidemiologists often appraise internal validity by considering the risk and potential impact of selection bias, information bias, and confounding. Selection bias refers to "*bias in the estimated association or effect of an exposure on an outcome that arises from the procedures used to select individuals into the study or the analysis*" (27). Information bias refers to a distortion of study results due to measurement error (27, 42). Confounding refers to a systematic distortion of an effect estimate between exposure and outcome due to a third variable (confounder). A confounder must be a risk factor for the outcome, associated with the exposure in the population of interest, and not affected by the exposure or outcome (42).

To obtain valid results from breast cancer screening studies, careful consideration must go into selecting sufficiently similar control groups and information sources for screening, diagnosis, and cause of death. Sufficiently long follow-up time – preferably 10 years or more – is also important because 5-year breast cancer survival is high (Figure 2), which can affect study precision. Further, when comparing breast cancer mortality among screened versus non-screened women, for example, it is important to ensure that breast cancer deaths among screened women resulted from incident cancers diagnosed *after* a woman was first invited to or attended screening so that the temporal sequence of the exposure and outcome is logical (internal validity) (4). Moreover, researchers must be careful to avoid drawing incorrect conclusions about the effects of screening on individual women based on population-level data (ecological fallacy; external generalizability).

Breast cancer screening in Norway

BreastScreen Norway is the population-based screening program in Norway. The program offers women in the target group a screening appointment every other year for 20 years (10 appointments total). The target group is women aged 50–69, but offers to attend screening are sent based on birth cohorts and residential region, and a woman's age at screening can range from 48–72 (43). Offers to attend screening are sent electronically or by letter-mail and women can call to change their appointment if the suggested time does not suit them. All

offers to attend screening are sent with an information leaflet about different aspects of screening to help inform women about the practical aspects of screening, as well as the major potential benefits and risks. This is discussed in the section *Information about mammographic screening* on page 34.

The Cancer Registry of Norway administers BreastScreen Norway based on national and European guidelines (28, 44). Two radiologists independently read all screening mammograms. A consensus or arbitration meeting determines whether a woman should be recalled for further assessment if both radiologists do not interpret a case as “normal”.

BreastScreen Norway started in one county in late 1995 and three others in early 1996. The program was nationwide by the end of 2005 (45). The transition from screen film mammography to digital mammography was gradual, beginning in a research setting in 2000 (45). Routine screening with digital mammography began at two centres in 2004 and the program was fully equipped with digital mammography in the fall of 2011.

Screen-detected and interval breast cancers

Breast cancers diagnosed because of a recall for further assessment after routine mammographic screening are called screen-detected breast cancers. Because we do not (and cannot) screen continuously for breast cancer, some breast cancers will inevitably be diagnosed between two routine screening examinations. These so-called interval cancers are often – but not always – diagnosed as a result of clinical symptoms. Cancers detected among women who have not attended a planned screening examination in an organized program are referred to as detected “outside screening”.

On a population level, screen-detected cancers are detected earlier in their disease course than interval cancers; the time between the diagnosis of a screen-detected breast cancer and when that cancer would have been detected in the absence of screening is called *lead time* (Figure 5). Slow growing breast cancers with long sojourn times are most susceptible to being screen-detected because they are less likely to become symptomatic between screening examinations. On the other hand, fast growing cancers are more likely to be diagnosed as interval cancers. This phenomenon is known as *length bias* (Figure 6).

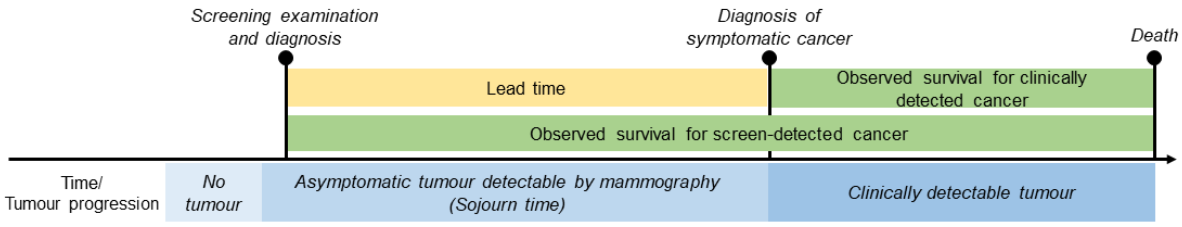


Figure 5: Illustration of breast cancer progression in the context of organized screening.

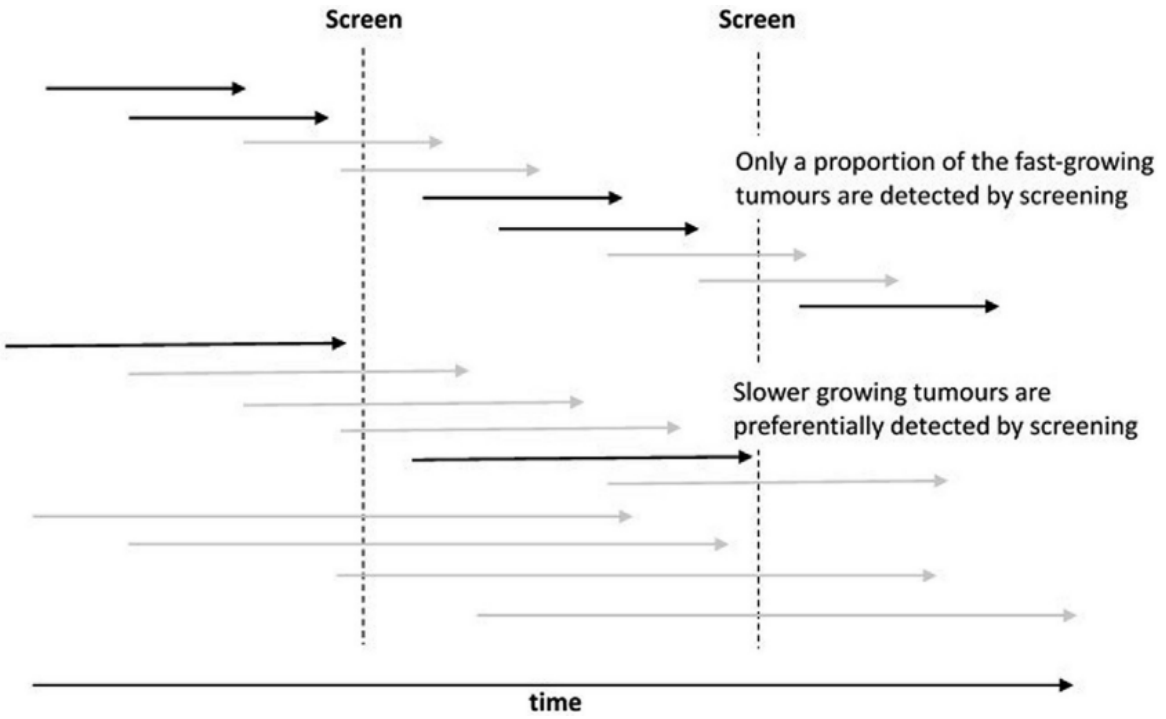


Figure 6*: The arrows represent a cancer's sojourn time. Screen-detected cancers are represented by grey arrows, while interval cancers are represented by black arrows. Periodic screening is more likely to detect slow growing tumours with a longer sojourn time, this is known as length bias.*[5]

Interval cancers represent roughly 25% of breast cancers detected among women attending BreastScreen Norway. On a population level, interval cancers have less favourable tumour characteristics than screen-detected cancers (45). Specifically, interval cancers are more likely to have a larger tumour diameter, be lymph node positive, and have a higher histologic grade than screen-detected breast cancers (46, 47). This is partially due to lead time and length bias. These biases complicate comparing survival outcomes for screen-detected and interval cancers. However, even after adjustment for patient and tumour characteristics, women diagnosed with interval cancer have lower survival than those diagnosed with screen-detected cancer (46, 47). It is desirable to avoid a diagnosis of an interval cancer if the abnormality could have been detected and diagnosed as a screen-detected cancer at the previous screening examination.

Radiographic assessment of "true" and "missed" cancers

Through a process called retrospective radiological review, radiologists can evaluate prior screening mammograms to determine whether a screen-detected or interval cancer was

[5] Reprinted from Clinical Radiology, Vol 73, 1. Wallis MG. How do we manage overdiagnosis/ overtreatment in breast screening? Page 374, Copyright (2018), with permission from Elsevier

visible at that time. Common classifications used in radiological reviews are “true”, “minimal signs”, “missed” and “occult” cancers (28). Definitions used in this thesis for true, minimal signs, and missed cancers are provided in Table 6. These are illustrated for screen-detected and interval cancers in Figure 7. Briefly, true cancers were defined as those diagnosed in women with no visible abnormalities on the prior screening mammograms. Minimal signs cancers were defined as those diagnosed in women with minor abnormalities on the prior screening mammograms at the cancer site, but where these did not necessarily warrant a recall for additional workup. Missed cancers were defined as those diagnosed in women with obvious visible findings at the cancer site on the prior screening mammograms, but that were not diagnosed as a result of attending the screening examination. Occult cancers were defined as those with no mammographically visible malignancy at the time of diagnosis.

The underlying rate of true cancers is positively associated with the amount of time between two regularly scheduled screening examinations (i.e. the duration of screening intervals). Shorter screening intervals will increase the number of true screen-detected cancers and longer screening intervals will increase the rate of true interval cancers (Figure 6). Again, because most women do not have breast cancer, screening programs must strike a balance between screen-detected cancer rates and the frequency that women are offered screening.

***Table 6**: Definitions of true, minimal signs, and missed breast cancers used in this thesis.*

Radiological classification	**Definition**
True	No abnormalities visible on prior screening mammograms at the cancer site (true negative prior screen), followed by a diagnosis of interval breast cancer, or screen-detected breast cancer during the subsequent screening round.
Minimal signs, actionable	Minor abnormalities visible on prior screening mammograms at the cancer site. Recall would have warranted, but was not expected within the screening program.
Minimal signs, non-actionable	Non-specific findings visible on prior screening mammograms at the cancer site. Recall not possible or expected within the screening program.
Missed	Obvious abnormalities visible on prior screening mammograms at the cancer site (false negative prior screen) that resulted in interval breast cancer or screen-detected breast cancer during the subsequent screening round.

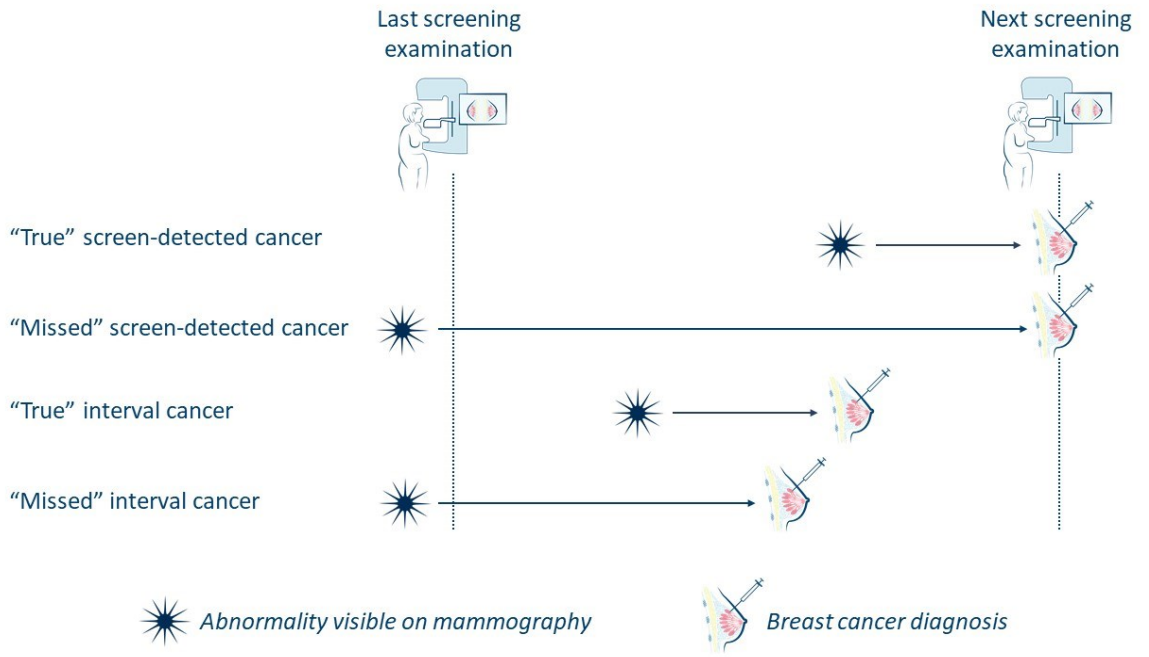


***Figure 7**: Missed cancers are those for which mammographic abnormalities were visible on the prior mammogram, but ultimately not diagnosed as breast cancer. True cancers are those that develop between two screening examinations.*

In addition to being affected by the underlying rates, the *observed* proportions of missed and true cancers are affected by the radiological review methodology used to assign these classifications. This is the case for both screen-detected and interval cancers. Consensus-based reviews can lead to a lower proportion of cancers classified as missed than reviews based on the assessment of a single radiologist (48). However, the proportion of cancers classified as missed is more affected by whether screening and diagnostic imaging is available (so-called informed reviews result in higher rates of missed cancers) or whether only the prior screening images are available (so-called blinded reviews have lower rates of missed cancers) (48-51). Further, radiologists also classify a higher proportion of cancers as missed when reviewing only cancer cases (non-mixed reviews) than when reviewing cancer cases together with normal screening examinations (mixed reviews) (52). It is important to note that nearly all radiologic reviews, regardless of their design, are performed in a study or training setting. This limits the extent that their results can be generalized to an ordinary screening setting.

Among interval cancers, a literature review of radiological review studies concluded that roughly 20–25% may be missed (46). To the best of my knowledge, a comparable literature review about screen-detected cancers has not been published in the peer-reviewed literature. Individual studies report 10–53% of screen-detected cancers as potentially missed

depending on the radiological review process (53-56). Among five studies presenting the proportions of missed screen-detected and missed interval cancers, one study found no difference, one found that the proportion of missed screen-detected cancers was higher than that of missed interval cancers, and three more recent studies found the opposite (54-57). Among the latter, the difference in proportions ranged from 3.4%–13% (54, 57). Women with missed interval cancers could potentially have benefitted from earlier detection at their previous screening examination, but it is less apparent whether women with missed screen-detected cancers could also have benefitted from earlier detection. This is further discussed on the section *"Underdiagnosis"* on page 33.

Benefits associated with screening

The primary aim of mammographic screening is to reduce deaths from breast cancer through early detection and this is the focus of this section. There are other benefits associated with mammographic screening, including that women with screen-detected cancers may receive gentler treatment than if their cancer had been detected due to clinical symptoms.

Reduced breast cancer mortality

Mortality can be calculated as a count or a proportion, but is often calculated as a rate. To evaluate screening, it is generally calculated with respect to death from breast cancer (breast cancer mortality). In 2018, 650 women residing in Norway died of breast cancer and the age-standardized breast cancer mortality rate was 22 per 100,000 woman years (58).

In Norway, breast cancer mortality started decreasing around 1996 (Figure 8). This coincides with an increase in 5-year relative survival that is attributable to improvements in breast cancer treatment and the introduction of organized screening (15). This mortality reduction could reflect the increase in survival, and it is important to consider how changes in incidence and survival may affect mortality. It can be difficult to disentangle these effects but one study has estimated that, for women over 50, organized screening in Norway was associated with a 20% decrease in breast cancer mortality and improvements in treatment were associated with a 23% decrease (59). This estimate is in keeping with those from the initial randomized trials about mammographic screening and later observational studies, which demonstrated that periodic mammographic screening reduces breast cancer mortality among average-risk women aged 50–69 by roughly 22% among those *invited* to screening (4, 59, 60), and approximately 33% among those *attending* screening (60).

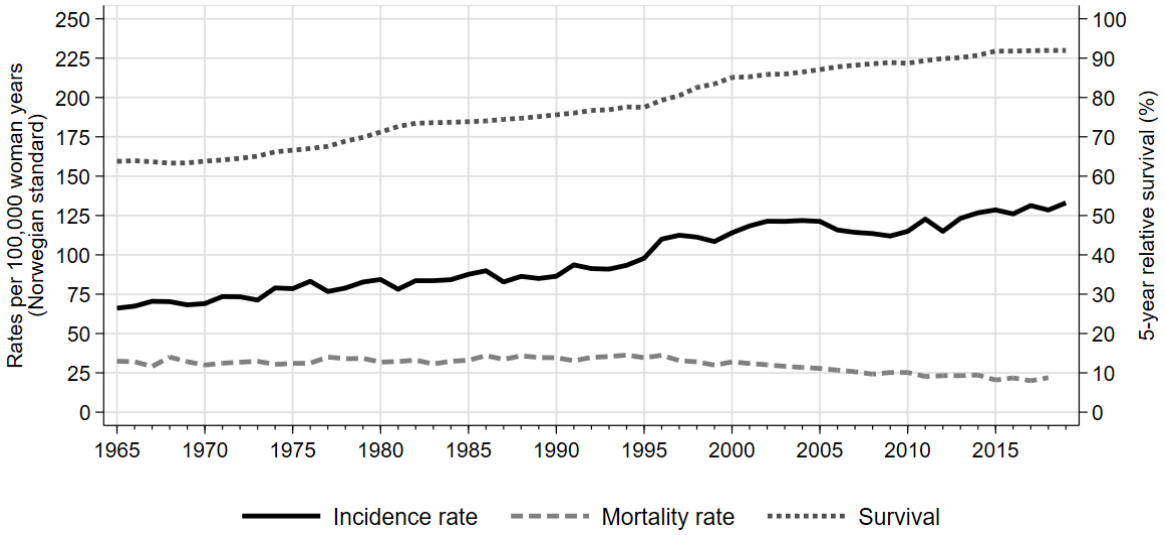


***Figure 8**: Age standardized breast cancer incidence, mortality, and survival in Norway from 1965–2019, among women of all ages.*[6]

Risks associated with screening

Early detection through screening deceases breast cancer mortality, but participating in organized screening can also have negative consequences for women. These include false positive screening examinations, overdiagnosis, and "underdiagnosis". Other risks associated with participating in organized screening include false negative screening examinations and exposure to ionizing radiation (4).

False positive screening examinations

A commonly recognized risk associated with mammographic screening is false positive screening examinations. This refers to "*having a screening mammogram that caused a recall for further assessment and therefore led [a woman] to believe [she] might have breast cancer when [she] do[es] not*" (item 9 on page 13) (61). Women can experience uncertainty, stress, anxiety, and fear after being recalled for further assessment (4, 62). This recall examination may help reassure a woman that she does not have the disease, but will also expose her to further imaging and potentially a breast biopsy. Women who undergo biopsy may wait longer to obtain their test results than women who do not. False positive screening examinations have been studied extensively. One systematic review reported that 46–98% of women were aware of this risk (63). Two other such reviews reported that women appear to be willing to be exposed to this risk for the potential benefit of early detection (62, 64).

[6] Figure adapted with permission from Cancer Registry of Norway. Cancer in Norway 2019 - Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2020.

Overdiagnosis

Overdiagnosis generally refers to “too much medicine”, often in the case where a disease diagnosis doesn’t benefit a patient (65). In mammographic screening, overdiagnosis (sometimes called “overdetection”) typically refers to the detection of a breast cancer that never would have presented symptomatically during a woman’s lifetime in the absence of screening (33). Overdiagnosis may refer to diagnosing a woman with a slow-growing or indolent breast cancer that would never have caused symptoms or death, no matter how long she lived. It may also refer to diagnosing a woman with a progressive breast cancer if she dies of another cause before her breast cancer would have otherwise become symptomatic (66). The consequences of having an overdiagnosed breast cancer, including potentially unbeneficial treatment, are similar for all women with both “types” of overdiagnosed cancers. However, a modelling study estimated that competing risk of death caused less than 4% of overdiagnosed cases (67). This thesis thus limits its scope to overdiagnosis caused by slow-growing or indolent cancers.

Overdiagnosis in mammographic screening was recognized during the 1960s, but was not considered a major risk associated with screening until the 2000s (33, 68). Today, it is considered one of the most serious potential risks associated with screening. Although the definition for overdiagnosis applies to individuals, our current knowledge about the progression of breast cancers in individuals is insufficient to differentiate women whose cancers will develop into harmful disease from those whose will not. As a result, treatment is offered to all women with breast cancer. Women with overdiagnosed breast cancers may undergo treatment that confers little or no medical benefit (overtreatment). Further, these women are unnecessarily exposed to the psychological and physical harms associated with breast cancer diagnosis, treatment, and survivorship (69). However, because most overdiagnosed women are assumed to be diagnosed with cancers that have a favourable prognosis (e.g. small tumours size, low grade), they are likely to be offered less aggressive treatment than women whose cancers are associated with a poor prognosis.

As long as we cannot identify slow-growing or indolent breast cancers that will never cause clinical symptoms in individuals, overdiagnosis will remain an inevitable aspect of breast cancer screening. However, the definition of overdiagnosis can be operationalized, and thereby quantified, on a population-level using epidemiological methods. One such approach compares the breast cancer incidence in a group of screened women to a similar group of non-screened women. Women are followed-up for breast cancer from the start of screening until such a time when the compensatory drop associated with stopping screening has passed. In this way, an excess of breast cancer cases in the screened group indicates overdiagnosis. Ideally, this would be performed in the context of a randomized controlled

trial, however, many studies use retrospective observational data to estimate overdiagnosis. These observational studies are highly sensitive to the study design, data, and analytical approach used (69-72). Although there is no agreement on the optimal methodology to estimate overdiagnosis, studies using individual-level data are generally more valid than studies using group-level information (ecological studies) (43). Aggregated data sources do not always have accurate information about which women attended screening or were diagnosed with breast cancer, or the dates of these events. The risk of misclassification (information bias) is therefore higher in ecological studies than those using individual-level data. Further, researchers may incorrectly draw conclusions about individual-level effects from group-level information (ecological fallacy).

The proportion of overdiagnosis in BreastScreen Norway is debated. Estimates range from 0% to over 75% (73, 74), but most are under 30% (73, 75-79). Less than half of these studies exclusively used individual-level data (73, 76, 77, 80). Studies using aggregate data have produced the highest estimates of overdiagnosis; more recent studies using individual-level data have reported lower rates (Figure 9).

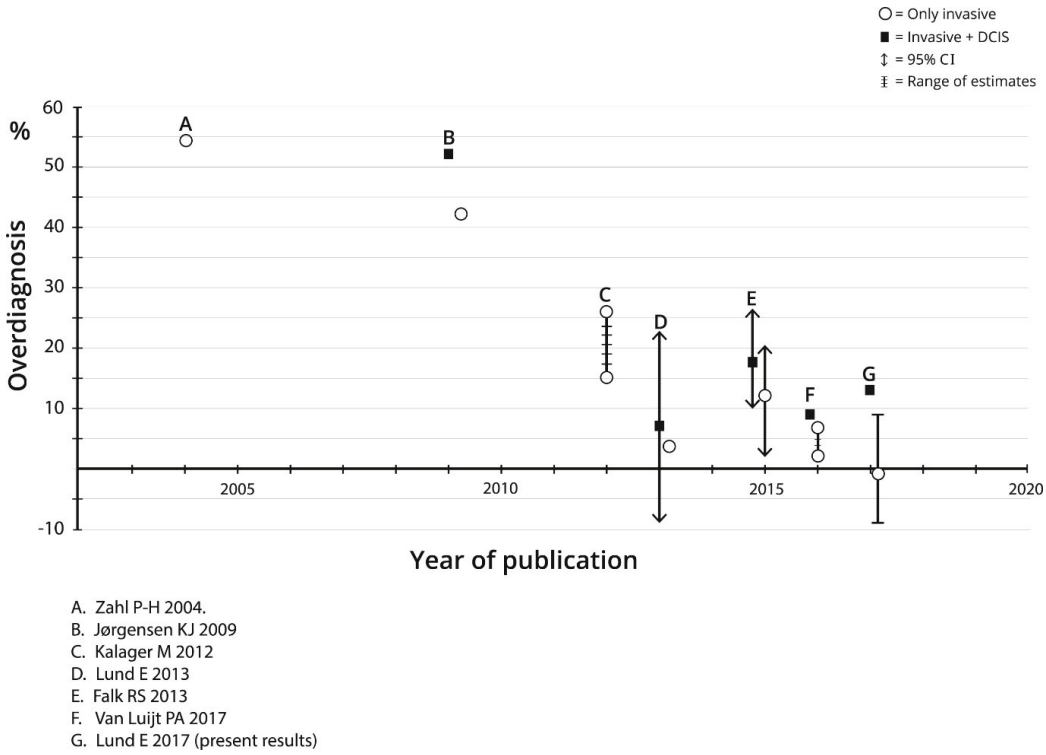


***Figure 9**: Estimates of overdiagnosis in BreastScreen Norway published during 2005–2017.[7] Studies A, B, and C used group-level data.*

[7] Reprinted from European Journal of Cancer, Vol. 89, Lund E, Nakamura A, Thalabard J, No overdiagnosis in the Norwegian Breast Cancer Screening Program estimated by combining record linkage and questionnaire information in the Norwegian Women and Cancer study (DOI: 10.1016/j.ejca.2017.11.003), p. 109, Copyright (2017), this work is licensed under CC BY-NC-ND 4.0.

"Underdiagnosis"

"Underdiagnosis" is not a common term in breast cancer screening and does not have an agreed-upon definition in this field. However, it can be understood in the context of overdiagnosis. Thus, where overdiagnosis can refer to "too much medicine" (65), "underdiagnosis" can refer to "too little medicine". For the purposes of this thesis, I defined "underdiagnosis" in mammographic screening as the inverse of overdiagnosis. Where overdiagnosis was defined as "diagnosing a woman with a slow-growing or indolent breast cancer that would never have caused symptoms or death, no matter how long she lived" (page 31), underdiagnosis was defined as *not* diagnosing a woman with a progressive breast cancer that *would* have caused symptoms or death during her lifetime.

This thesis focused on the potential for underdiagnosis to occur when a cancer retrospectively visible at routine screening was missed and subsequently diagnosed as an interval cancer or screen-detected cancer in the subsequent round (i.e. diagnosed following a false *negative* screening examination). For example, if a "missed" interval cancer was diagnosed after a woman developed clinical symptoms of breast cancer, I assumed this was underdiagnosed based on the definition above. However, I could not determine whether "missed" asymptomatic screen-detected cancers would have eventually caused symptoms during a woman's lifetime (in the absence of her diagnosis). I could only speculate whether these cancers were potentially underdiagnosed based on their tumour characteristics.

The prognostic and predictive tumour characteristics and survival profile of missed (potentially underdiagnosed) screen-detected breast cancers are not well established and I am not aware of any studies evaluating women's survival after a diagnosis of true or missed screen-detected cancer. Among interval breast cancers, it has been hypothesized that missed cancers represent less aggressive disease with favourable prognostic features compared to true cancers that have a shorter sojourn time (56, 81). Indeed, a number of studies have observed that missed interval cancers are less frequently histopathologic grade 3 than true interval breast cancers, but they also often have a larger tumour diameter (82-87). Results have been less consistent for other aspects of histopathology. Three out of four published studies have failed to detect a difference in the overall survival associated with true and missed interval cancers (81, 83, 84, 86). More studies are needed to better describe the impact of missed versus true cancers in organized screening.

Diagnostic accuracy

Because breast cancer staging and personalized treatment are closely tied to histopathologic tumour characteristics, diagnostic accuracy is important to help ensure that women are offered appropriate treatment for their disease. This is particularly the case for women with

overdiagnosed breast cancers, whose treatment confers little or no medical benefit to them. Diagnostic accuracy is therefore crucial to minimize the negative consequences of overtreatment by helping to ensure these women are offered the most targeted and least invasive treatment required to treat their disease. With respect to underdiagnosed cancers, diagnostic accuracy is important to ensure that women are offered appropriate and personalized treatment that can mitigate some of the potentially unfavourable consequences associated with a delayed diagnosis of breast cancer.

Consider the preferential overrepresentation of certain digits (so-called *terminal digit preference*) in reporting numerical tumour descriptors. This is a potential source of measurement error when radiologists and pathologists report the maximum diameter of a tumour (referred to as “tumour diameter” hereafter). The T1–T3 classifications for breast cancer are defined by tumour diameter and inaccurate measurements could affect T staging. Although it has rarely been evaluated as a primary outcome, some studies report that breast pathologists over-report tumour diameters that are a multiple of five millimetres (88-90). Terminal digit preference is not well-described among breast radiologists, but is important to describe because radiological tumour diameter measurements inform cT staging and can therefore affect neoadjuvant and surgical treatment decisions (9). In this thesis, I explored terminal digit preference among breast radiologists and pathologists and examined whether this could cause under- or overstaging and thereby affect treatment decisions.

Information about mammographic screening

Screening is associated with potential benefits and risks, as previously described. Although these risks may be outweighed by the benefits on a population-level, individual women who experience some of the potentially negative aspects of screening may not experience the benefit of early breast cancer detection. For example, a woman with a false positive screening result may never develop breast cancer and never benefit from early detection of the disease. It is therefore important that women have access to accurate and balanced information about the potential consequences of participating in organized screening. This can help them make a decision whether to attend.

In Norway, the national screening program is responsible for developing information about screening for the women it serves and has always included information about screening alongside the invitations it sends to women. Surveys at selected screening units in 2015 and 2018 found that 84%–96% of women attending screening read at least some of the information they received with their invitation (91, 92). The 2015 survey indicated that less than 20% of women had searched for additional information about screening (92). This

underlines the important role that BreastScreen Norway has in communicating accurate and complete information about screening to women.

Informed choice

Initially, much of the information material created for women by screening programs focused on the benefits of screening (93). This aimed to increase the participation rate, which is a quality indicator used to monitor the performance of organized screening programs. In BreastScreen Norway, the participation rate is roughly 75% (45, 94). The 2006 European Guidelines indicated that >70% participation is acceptable and >75% is desirable (28). Participation rates are still under consideration as potential performance indicator by the European Commission Initiative on Breast cancer (28, 95).

Women's autonomy has been increasingly valued by the medical establishment, and informed choice is becoming a key tenet in guidelines for organized screening. For example, the Norwegian Directorate of Health has suggested that information about participation in organized screening programs should be evidence-based and empower making an informed choice about participation (Table 5) (32). Additionally, the 2006 European Guidelines noted that it is *"vital that … women know the pros and cons of breast screening to help them make an informed decision about whether or not to attend*" (Page 382, §12.1) (28). The most recent European Guidelines conditionally recommend providing women with a decision aid to help them make an informed choice about whether they want to attend screening (96).

Informed choice in screening is often defined as a woman making a decision whether to attend screening based on relevant information or knowledge (97). This definition is difficult to operationalize but generally involves assessing women's knowledge and values toward screening, as well as their screening behaviour. A woman makes an informed choice if her decision to attend screening is consistent with her values *and* she demonstrates "adequate" knowledge about screening (97). Measuring informed choice requires making judgements about how to define, measure, and interpret women's knowledge and values and there is no consensus on how to do this (97). Further, the classification "informed"/"not informed" can oversimplify the complex individual components that make up this concept and the interactions that occur between them (97). Reporting detailed information about each individual element of informed choice may provide more nuanced, and perhaps more actionable, information about women's decision making processes.

It is unclear whether screening programs can achieve high participation rates while simultaneously enabling informed choice. Moreover, some women prefer not to make an active decision whether to participate in organized screening (98). However, this does not absolve screening programs from their obligation to provide accurate and balanced

information about screening. Most contemporary European screening programs do provide women with some information about the potential benefits and risks of mammographic screening. Many such programs, including BreastScreen Norway, also have policies to support women in making an informed choice when considering mammographic screening (99). Nonetheless, a number of screening programs have been criticized for providing women with unbalanced or incomplete information about screening, particularly with respect to overdiagnosis (100-105). It is important for women to be aware of, and have access to, information about overdiagnosis and overtreatment because their knowledge about this topic can affect how they think about and value mammographic screening (106-108).

BreastScreen Norway's information leaflet has described overdiagnosis since 2009 (45). This information was updated in 2017 and 2020 (45, 109). The program aimed to encourage women to make an informed choice while making these updates. More detailed information about overdiagnosis and other aspects of screening is available on the program's website.[8] A routine questionnaire sent with screening invitations in 2015 indicated that 96% of 154,884 participants felt that the quality of information offered by BreastScreen Norway was "good" or "very good" (results unpublished). Focus groups performed in the same year indicated that participating women did not feel the need for information about the benefits and harms of mammographic screening, but they felt it might be important for others (92).

Women's knowledge about overdiagnosis in Norway

A systematic review of literature from countries in the International Cancer Screening Network's breast cancer division indicated that women have low levels of awareness and knowledge about overdiagnosis (63). Studies have shown that overdiagnosis is difficult to define and understand, and that it is not uncommon for women to confuse false positive screening examinations and overdiagnosis (106, 107, 110).

In Norway, a 2018 survey (n = 204) conducted as part of a Bachelor's degree thesis reported that 65% of women attending screening reported that they had heard of overdiagnosis and knew what it was and 16% reported having heard of the word, but were unsure of its meaning (91). Women were not asked to provide a definition for overdiagnosis in this study, however focus group discussions in 2015–2016 among screening-aged women indicated that overdiagnosis was difficult for them to understand and was sometimes confused with false positive screening examinations (92). Additional information about the extent to which women in Norway are knowledgeable about mammographic screening, overdiagnosis in particular, is scarce.

[8] https://www.kreftregisteret.no/screening/Mammografiprogrammet/

Research objectives

The objective of this PhD project was to investigate over and underdiagnosis in BreastScreen Norway. This project consisted of three papers based on separate studies that explored concepts related to overdiagnosis, underdiagnosis, and women's knowledge about breast cancer in general and overdiagnosis in particular.

The research objectives this PhD project sought to address were:

1. To determine whether terminal digit preference is present among tumour diameter data registered at the Cancer Registry of Norway and evaluate whether this had the potential to lead to under- or overstaging.
2. To determine whether tumour histopathology and survival are differentially associated with true, minimal signs, and missed screen-detected and interval cancers, and explore whether this could indicate if missed cancers were over- or underdiagnosed.
3. To describe Norwegian women's conceptual knowledge about mammographic screening in general, and overdiagnosis in particular.

Paper 1 addressed the first research objective, Paper 2 the second, and Paper 3 the third.

Materials and methods

The three studies included in this thesis are summarized in Table 7.

***Table 7**: Study characteristics of the three papers included in this thesis.*

	Paper 1 **Terminal digit preference**	**Paper 2** **Survival associated with true and missed cancers**	**Paper 3** **Knowledge about screening and overdiagnosis**
Ethical approval	PVO 19/02585	PVO 2016/4696	Not applicable[a]
Data sources	Cancer Registry of Norway	Cancer Registry of Norway Radiologic review	Anonymous online survey
Study design	Cross-sectional	Cohort	Cross-sectional
Study sample after exclusions	14,168 women of all ages diagnosed with T1–T3 breast cancer during 2012–2016 (all modes of detection)	1022 women diagnosed with screen-detected breast cancer and 788 women diagnosed with interval breast cancer	1892 women aged 45–75 in 2020
Primary outcome	Terminal digit preference	Overall survival	Knowledge about breast cancer screening and overdiagnosis
Main statistical methods	Descriptive statistics	Cox regression	Descriptive statistics

[a] *This study was exempt from ethical approval because it was based on anonymous information.*

Ethical approvals

The study for Paper 1 fell under the umbrella of quality assurance and improvement since it could highlight measurement error in data registered at the Cancer Registry of Norway. This study was approved by the privacy ombudsman at the Oslo University Hospital (PVO 19/02585). Access to the study data was sought from the Registry's data delivery unit (ref. 19/39, DU-3292) after determining that a data privacy impact assessment was not required.

The study for Paper 2 was a part of a quality assurance and improvement project approved by the privacy ombudsman at the Oslo University Hospital in 2016 (PVO 2016/4696). To limit the risk of re-identifying individuals, all dates were set to the 15th of the month and no unique personal identifiers were included in the analytical data set.

The study for Paper 3 was exempt from review by the local research ethics board and institutional privacy ombudsman. This was because the study data came from a questionnaire that only collected categorical data that were considered anonymous. Submitting a completed questionnaire was deemed as consent to participate. Non-consenting women could exit the questionnaire at any time with no consequence.

Data sources

Cancer Registry of Norway

The Cancer Registry of Norway provided data used in Paper 1 and Paper 2. The Registry was established in 1952 and registers information about all cancer cases diagnosed among residents in Norway. Cancer reporting is mandated by law, and the Registry obtains information about cancer patients from clinical sources, pathology reports, and death certificates (Figure 10) (15). Nearly 100% of breast cancer cases in Norway are reported to the Registry and over 99% of these are morphologically verified (111).

As the administrator of BreastScreen Norway, the Registry also maintains individual-level records of screening invitations, attendance, screening results, and any follow-up or diagnosis of screen-detected or interval breast cancer. Information about mode of detection (screen-detected, interval, or outside screening) is therefore available for all women diagnosed with breast cancer. Women have the right to opt out of having their normal screening results saved at the Registry and used for research. As of December 31, 2016, less than 2% of women had ever exercised this right (45).

The Registry is also responsible for the National breast cancer quality registry (Norwegian: *Nasjonalt kvalitetsregister for brystkreft*). Using an electronic portal, this quality registry collects information about diagnosis, treatment, and follow-up for all female breast cancer patients (11). Surgical and pathological information is highly complete, but radiological and oncological information is less so. Oncological information was not used in this thesis.

Radiological information is submitted by radiologists via an electronic “radiology form”. Radiologists working at breast diagnostic centres in BreastScreen Norway have had access to and submitted a version of this form since the inception of the program. These forms have historically been submitted for nearly all women diagnosed within BreastScreen Norway (112, 113). In 2015, radiology forms were submitted for 47% of cancers diagnosed outside the program (113). At this time, one large referral hospital did not have access to the electronic system required to access and submit these forms to the Registry. This access was established in 2016, and forms were submitted for 72% of all women diagnosed with

breast cancer that year (114). To date, access has not been established at private centres and some smaller public centres – both of which diagnose cancers outside of BreastScreen Norway – and the overall submission rate therefore cannot reach 100% (11). Moreover, although processes for reminding radiologists about missing forms are well established for women diagnosed within BreastScreen Norway, these were not established for women diagnosed outside the program until 2017–2018 (115). This has also contributed to the excess of missing radiology forms for cancers diagnosed outside the screening program.

Paper 1 used information about breast cancer diagnoses and associated tumour characteristics (including tumour diameter) sent to the Registry by radiologists and pathologists. Paper 2 used information about breast cancer diagnoses and associated tumour characteristics sent to the Registry by pathologists. This paper also used information on emigration and cause of death from the Registry. Information about missing data for individual variables is provided in the section *Missing data* on page 48.

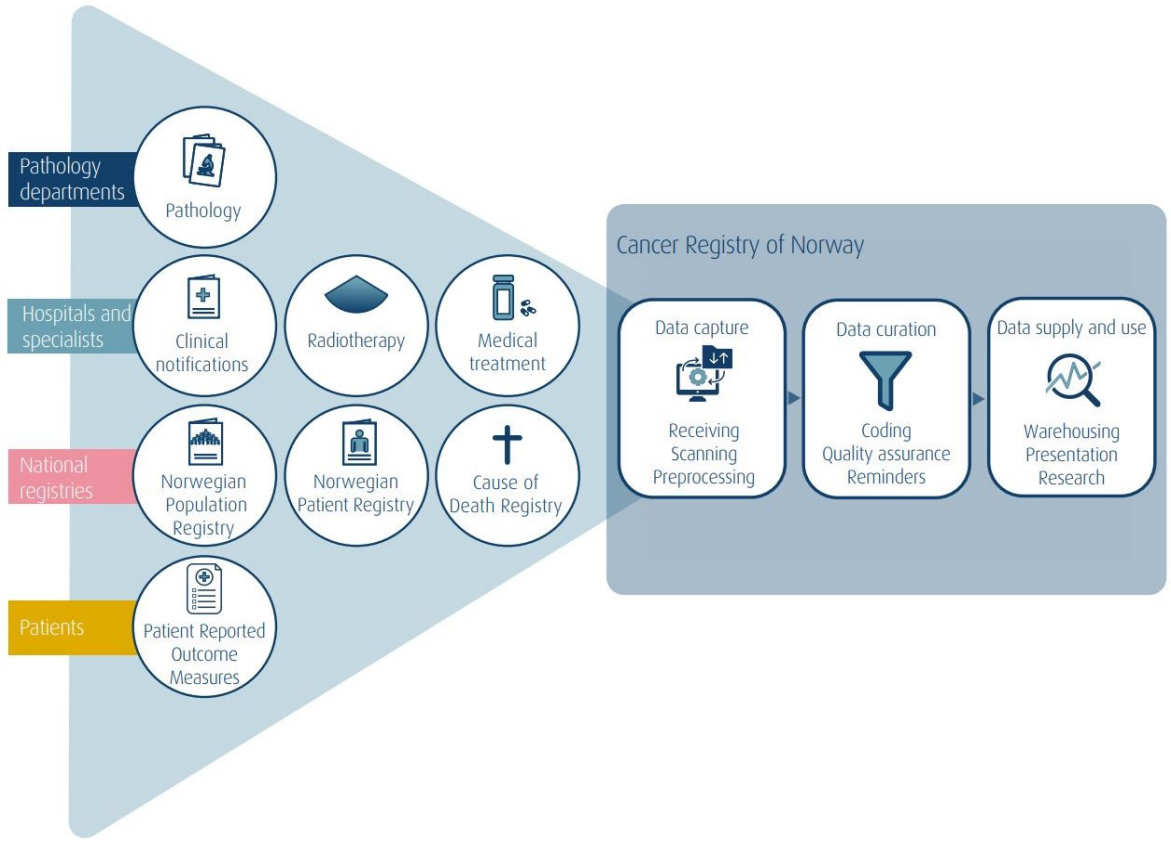


***Figure 10**: Data sources and registration processes at the Cancer Registry of Norway.*[9]

[9] Figure reprinted with permission from Cancer Registry of Norway. Cancer in Norway 2019 - Cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2020.

Retrospective radiological review

Paper 2 linked data from the Cancer Registry of Norway to radiologic classifications obtained during a nationwide radiological review of ductal carcinoma *in situ* and breast cancer diagnosed in women attending BreastScreen Norway. The purpose of the review was "quality improvement for radiologists' performance and the program" (116). The review itself is described in detail elsewhere (116). It was performed between September 2016 and April 2017 by 37 radiologists who had read at least 5000 mammograms during the past two years. It involved all 16 breast centres across the country. Centres were randomly paired and a panel of five radiologists reviewed cases from paired centres. Each panel consisted of two radiologists from the reviewing centre, two from the paired centre, and Dr. Tone Hovda. Dr. Hovda and a representative from the Registry were present at every session to ensure consistency in the review process, data collection and coding.

The panels performed an informed review using screening and diagnostic images. Information on surgical treatment and tumour histopathology was provided after each case was reviewed. Based on whether a tumour was retrospectively visible on the prior screening examination, a panel of radiologists classified all cancer cases as "true", "minimal signs", "missed", or "occult" (defined on page 27). This classification was determined through consensus, or a majority vote if consensus could not be reached.

Each centre aimed to identify a sample of 75 screen-detected and 75 interval cancers diagnosed with digital mammography for review. However, some low-volume centres oversampled screen-detected cancers to obtain 150 cases for review.

Online questionnaire

Paper 3 used data collected through an online questionnaire between April 8 and June 8, 2020. The questionnaire targeted women aged 45–69 living in Norway and was written in Norwegian (bokmål). It was shared through a Facebook post made by the Cancer Registry of Norway's Kreftsjekken page[10], which has roughly 8200 followers (mostly women).

The questionnaire was administered through a secure platform for survey management developed by the University of Oslo (Nettskjema) (117). Internet protocol (IP) addresses and other such personal identifiers were not saved. Further, Nettskjema did not save information about the time a survey was submitted, and cookies were not used while the survey was in progress (118). In other words, responses were not saved temporarily while the survey was being completed, they were saved only when submitted by a participant.

[10] https://www.facebook.com/kreftsjekken/

Our questionnaire included four sociodemographic questions and five background questions; the latter included two questions about breast cancer screening. Further, the questionnaire included nine graded conceptual questions covering three themes associated with screening: breast cancer mortality benefit (n = 2), false positive screening examinations (n = 1), and overdiagnosis (n = 6). Both graded and non-graded questions about breast cancer screening were based on the work of Hersch and colleagues (119). Key questions and response alternatives used in our study are described in the section *Key variables* on page 44; these have been back translated to English.

Questions and potential responses from the work of Hersch and colleagues were translated from English to Norwegian (bokmål) by four of the authors of Paper 3, who included a native English speaker and native Norwegian speakers (119). The Norwegian questions are presented in the Appendix. They were tested by five women aged 50–69 before the survey was made publicly available.

Study designs

The studies for Papers 1 and 3 were both cross-sectional. Cross-sectional studies describe the characteristics of a group at a particular point in time (27). In this thesis, Paper 1 used retrospectively available data from 2012–2016 and Paper 3 used data collected in 2020.

The study for Paper 2 used a historical cohort design. A strength of cohort studies is that they can make use of temporal information about an exposure that precedes the outcome of interest (120). In our study, women were selected based on a diagnosis of screen-detected or interval cancer, which was then classified as true, minimal signs, or missed. These women were followed-up from their diagnosis of breast cancer until December 31, 2018 and we used this temporal information to investigate overall survival. This is described in more detail in the section *Main statistical analyses* on page 49.

Study samples

The sample described in Paper 1 included 16,767 women of all ages residing in Norway who were first diagnosed with invasive breast cancer during 2012–2016. We excluded women with pT4 lesions, whose mammographic or histopathologic tumour diameter was > 99 mm, and women for whom no tumour diameter information was registered. The final sample consisted of 14,468 women (Figure 11). Mammographic tumour diameter was available for 7792 women, histopathologic tumour diameter for 13,541 women, and both for 6865 women.

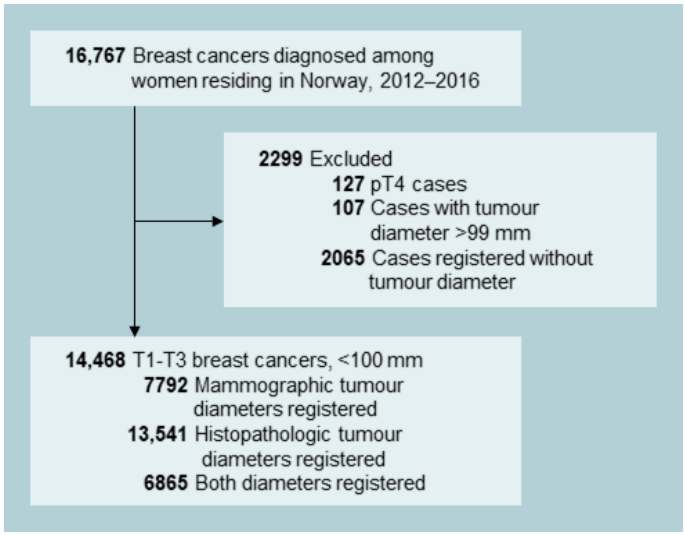


***Figure 11**: Number of individuals included and excluded in Paper 1. Individuals were excluded sequentially using the exclusion criteria.*

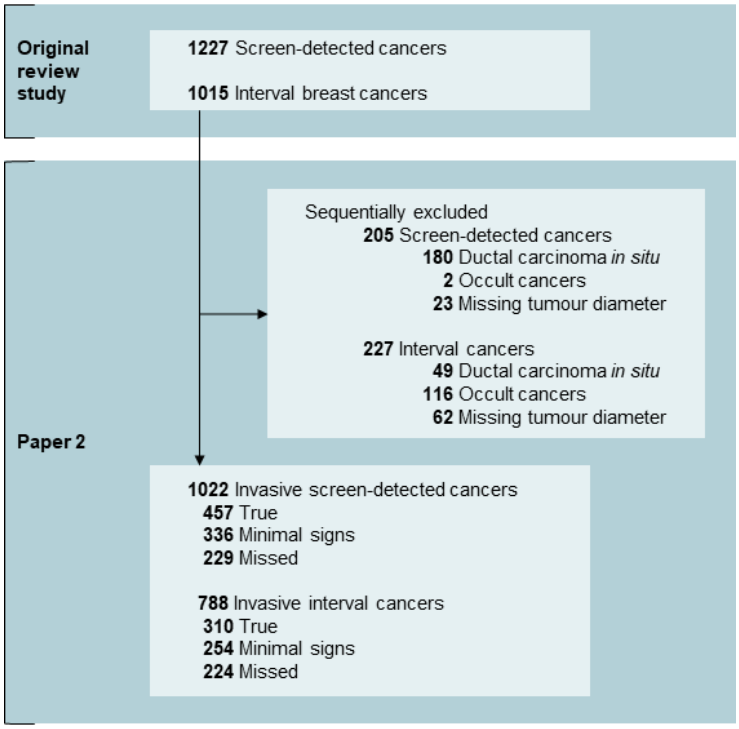


***Figure 12**: Number of individuals included and excluded in Paper 2. Individuals were excluded sequentially using the exclusion criteria.*

The sample used in Paper 2 was based on the 1127 screen-detected and 1015 interval ductal carcinoma *in situ* and breast cancers included in the aforementioned radiological review. These cancers were diagnosed during 2005–2016. We excluded women diagnosed with ductal carcinoma *in situ* (n = 180 screen-detected and 49 interval cancers), women with

mammographically occult tumours (n = 2 screen-detected and 116 interval cancers), and women with no recorded tumour diameter information (n = 23 screen-detected and 62 interval cancers) (Figure 12). The final sample consisted of 1022 women with screen-detected cancer and 788 women with interval cancer classified as true, minimal signs, or missed. We followed these women from the date of their breast cancer diagnosis until death from any cause, emigration, or December 31, 2018.

In Paper 2, we also conducted a sensitivity analysis in which we did not exclude women missing tumour diameter information. This sample consisted of 1045 women with screen-detected cancer and 850 women with interval breast cancer.

Paper 3 was based on a convenience sample of 2033 women who responded to our questionnaire. We excluded women outside the ages of 45–74 (n = 13) and limited our analyses to women without missing data (n = 128). This left 1892 women in the final sample.

Key variables

The tumour diameter was the primary outcome in Paper 1. This refers to the longest distance across the outermost boundaries of an invasive breast cancer and was measured in millimetres (mm). The examining radiologists measured mammographic tumour diameter from screening or diagnostic mammograms. These could be acquired from standard 2D digital mammography, tomosynthesis, or synthetic 2D images derived from tomosynthesis, and could include spot compression views (with or without magnification). Pathologists measured histopathologic tumour diameter using a transparent ruler. Measurements were taken from the macroscopic examination (formalin fixed specimen) if they could not be obtained using the microscopic slide. Measurements were taken from a single tissue slice or estimated across all slices containing microscopically verified invasive tumour tissue (9).

In Paper 2, the primary exposure was whether women were classified as having a true, minimal signs, or missed cancer based on the classification assigned during the radiological review. The review classified minimal signs cancers as either actionable or non-actionable, but we grouped these categories together.

The primary outcome in Paper 2 was overall survival. We calculated survival time from the date of a histologically verified diagnosis of screen-detected or interval cancer until death from any cause. Women's survival time was censored if they emigrated before or were still alive on December 31, 2018. Breast cancer specific survival was measured from a woman's diagnosis until death from breast cancer. Women's survival time was censored if they died from any other cause, emigrated before, or were still alive on December 31, 2018.

The covariates in Paper 2 included age at diagnosis (five-year age groups), tumour diameter (continuous variable, measured as described above), histologic grade, and subtype. We derived information about subtype using the surrogate definition of intrinsic subtype from the St Gallen consensus based on individual-level information about ER and PR status, as well as HER2 positivity and Ki67 expression (12). These variables are defined in Table 8; the subtype definitions are presented in Table 3.

***Table 8**: Definitions of variables used to determine subtype in Paper 2.*

Variable	Description
Estrogen receptor (ER) status	Positive if the sample displayed ≥10% reactivity, and negative otherwise. We prioritized histopathologic samples from the diagnostic biopsy and used histopathologic results from the surgical specimen if the former were not available.
Progesterone receptor (PR) status	Positive if the sample displayed ≥10% reactivity, and negative otherwise. We prioritized histopathologic samples from the diagnostic biopsy and used histopathologic results from the surgical specimen if the former were not available.
HER2[a] positivity	If in situ hybridization (ISH) was performed, a borderline or amplified result was considered positive. If ISH was not performed, positivity was determined using immunohistochemistry (IHC). HER2 positivity was assessed using the surgical specimen.
Ki67 expression	Percentage of Ki67 positive cells from a sample of 500 tumour cells from the surgical specimen.

[a] *Human epidermal growth factor receptor 2*

Variables used in Paper 3 described women's sociodemographic characteristics and conceptual knowledge about mammographic screening. They were derived from the online questionnaire. Briefly, the questionnaire asked women about their age (44 or younger, 45–49, 50–54, 55–59, 60–64, 70–74, 75 or older), highest completed formal education, region of residence, and birth country. These variables were designed to match categorizations used in population statistics by Statistics Norway (121-124).

As described earlier, questions about breast cancer screening were based on the work of Hersch and colleagues (119). All questions were closed-form. A back translation of the background and graded questions and potential answers women could select are outlined in Table 9. The graded questions were assigned marks based on the rubric published by Hersch and colleagues (119).

***Table 9**: Translated survey questions and potential responses; where applicable, correct answers are marked in italics. Graded conceptual knowledge was assessed based on a published questionnaire and rubric.[a]*

Thematic category	Survey question	Response format	Marks if correct
Background questions			
Not assigned	Have you ever searched for information about mammographic screening?	- Yes, a lot - Yes, some - No - Can't remember/unsure	NA
Not assigned	Where have you searched for information about mammographic screening? (Multiple options can be selected)	- Family doctor/other health personnel - Friends and family - Cancer Registry of Norway/BreastScreen Norway's website - Other websites - Scientific literature - Books - Newspapers and periodicals - Other	NA
Not assigned	Choose the sentence you think is correct	- *Mammographic screening is having a mammogram when you haven't noticed a lump or other breast cancer symptoms* - Mammographic screening is having a mammogram when you have noticed a lump or other breast cancer symptoms - Unsure	NA
Not assigned	Have you ever heard of these three terms?[b] - False positive screening examinations - Overdiagnosis - Overdetection	- Yes - No - Unsure	NA
Not assigned	Choose the sentence you think best describes "false positive screening examinations"	- *Abnormal findings on a screening mammogram, but where additional examination doesn't find breast cancer* - Breast cancer that would have never been detected if a woman hadn't attended screening - Unsure	NA

Graded questions			
Breast cancer mortality benefit	Who do you think has the highest chance of dying from breast cancer?	- Women who attend screening for breast cancer/mammographic screening - *Women who do NOT attend screening for breast cancer/mammographic screening* - Unsure	2
	Can mammographic screening detect all breast cancers?	- Yes - *No* - Unsure	1
False positive screening examinations	Will all women who have abnormal findings on a screening mammogram be diagnosed with breast cancer?	- Yes - *No* - Unsure	1
Overdiagnosis	Who do you think has the highest chance of being diagnosed with breast cancer?	- *Women who attend screening for breast cancer/mammographic screening* - Women who do NOT attend screening for breast cancer/mammographic screening - Unsure	1
	Cross off the sentences you think are true	- [Correct if not crossed off] All breast cancers will eventually lead to sickness and death if they are not diagnosed and treated[c] - [Correct if not crossed off] Doctors can distinguish harmful breast cancer that needs treatment from “nice” breast cancer that doesn’t need treatment with certainty[c] - *[Correct if crossed off] Slow growing breast cancers that are treated even though they would not have caused sickness exist[c]* - *[Correct if crossed off] Mammographic screening leads to the diagnosis of slow growing tumours and unnecessary treatment[c]*	1 1 1 1
	Choose the definition you think best describes “overdiagnosis”	- Abnormal findings on a screening mammogram, but where additional examination doesn’t find breast cancer - *Breast cancer that would have never been detected if a woman had not attended screening* - Unsure	1

[a] *Adapted from Hersch J, Barratt A, Jansen J, et al. Use of a decision aid including information on overdetection to support informed choice about breast cancer screening: a randomised controlled trial. The Lancet. 2015;385(9978):1642-52.*

[b] *These terms were shown together in a grid, and participants could select one response for each term.*

[c] *These sentences were shown together in a grid and participants had the option to cross off any item.*

Missing data

Missing data are common in medical research and refer to incomplete information about study subjects. Information may be missing because was it was not acquired (i.e. does not exist), or because it was not registered with a particular data source (i.e. is not available). In the case of questionnaire data, a participant may not respond to certain questions.

Pathological tumour diameter is a key variable in Papers 1 and 2. It is generally highly complete at the Cancer Registry of Norway (11). However, pT information is missing for women who receive neoadjuvant therapy and ypT information is available instead. This is because histopathologic tumour diameter measurements made after neoadjuvant treatment may not reflect the diameter of the untreated tumour. Neoadjuvant treatment is often offered to women with larger and/or more aggressive tumours and is therefore disproportionally offered to women with interval cancers or cancers detected outside of screening. Thus, larger (untreated) tumours are more likely to have missing pT information and the size of the missing (untreated) histopathologic tumour diameters is missing not at random.

Mammographic tumour diameter is a key variable in Paper 1. It is reported to the Cancer Registry of Norway via the “radiology form” submitted by radiologists. As described previously, radiologic information is less complete than histopathologic information. Further, mammographic tumour diameter information is not always reported on the radiology form and is therefore missing at a higher rate than the form itself. This missingness is likely attributable to mammography occult tumours (which cannot be measured), or to the field simply not being filled in. Overall, there is a high proportion of missing mammographic tumour diameter information. Mammographic information is more complete for women with screen-detected cancers than women with larger tumours detected outside screening (11).

Papers 1 and 3 were based on a complete case analysis whereby women were excluded from the final analytical sample if they had missing data for any variable of interest. In Paper 2, we used multiple imputation with chained equations to handle missing data (125). We assumed that the probability of a missing value was random given a set of measured auxiliary variables (this type of missing data is called “missing at random”) (126). Given data about detection mode and year of diagnosis, we believed this assumption was justified for missing information about histopathologic grade; lymph node status; ER, PR, and HER2 positivity; and Ki67 expression, but not tumour diameter. Without more information about tumour diameter or receipt of neoadjuvant therapy, we decided to exclude women missing pT tumour diameter information from our primary analysis.

Survey weights

Poststratification is an adjustment technique used in survey research to rebalance a sample to better reflect the population with respect to a set of variables (127). In Paper 3, we calibrated survey weights for age, education, and region based on population statistics from Statistics Norway (121-123). Control statistics from Statistics Norway represented women aged ≥ 40 for education, and women aged 45–74 for region. Because our study variables for age, education, and region used the same categorization as those from Statistics Norway, no variable transformations were required to calibrate the survey weights.

Survey weights were calibrated using an iterative proportional fitting (raking) procedure (128). This procedure produces weights that preserve the marginal distributions of the auxiliary (weighting) variables, but does not guarantee that the joint distributions are preserved.

Main statistical analyses

We used various descriptive techniques across the three papers. Generally, we used means and standard deviations (SD) to describe normally distributed variables such as age, and medians to describe skewed variables such as tumour diameter. We constructed 95% confidence intervals (CIs) for medians using quantile regression with standard errors based on 100 bootstrap replications. Frequencies and proportions were used to describe the distribution of categorical variables observed in the different study samples; 95% CIs for proportions were constructed using the Wilson score interval (129). Further, we used histograms and bar charts to illustrate the distributions of pertinent variables.

The primary analysis in Paper 1 focused on describing the frequencies of the last (terminal) digit of tumour diameters recorded at the Cancer Registry of Norway. We created histograms that plotted the frequencies of tumour diameters in 1 mm increments to visualize the distribution of tumour diameters. Further, we created histograms to visualize the frequencies of the terminal digits of tumour diameters and reported the proportion of terminal digits that corresponded to half- or whole-centimetre values (i.e. had 0 or 5 as a terminal digit). This was to evaluate whether there was an excess of certain digits in the reported measurements. For this thesis, I also performed these analyses stratified by mode of detection: screen-detected or “clinically detected”. Clinically detected cancers represented both interval cancers and those detected outside of the screening program among women aged 50–69.

The analysis in Paper 2 used frequencies and proportions to describe the tumour histopathology of true, minimal signs, and missed cancers, both screen-detected and

interval. The main analysis in this paper described the survival associated with a diagnosis of true, minimal signs, or missed cancer stratified by mode of detection (screen-detected or interval). We used Kaplan-Meier curves to describe the overall survival associated with these classifications. We used the Nelson-Aalen estimator to describe the cumulative risk of death from breast cancer for women with interval cancers (130). There were too few deaths among women with screen-detected cancer to perform a similar analysis. We used the log-rank test to test for differences in overall survival and breast cancer specific survival between the classification groups (130).

Using 40 sets of imputed data, we used Cox regression to estimate the relative hazard (risk) of death from any cause for minimal signs or missed cancers compared to true cancers. Due to the length and lead time associated with screen-detected cancers, this analysis was performed separately for screen-detected and interval cancers. Multivariable models were adjusted for age at diagnosis, tumour diameter, histopathologic grade, and subtype. These variables were selected as confounders based on a priori knowledge generated from studies of missed and true interval cancers (82-87). Hazard ratios and their corresponding 95% CIs were derived using Rubin's rules (131).

Cox regression assumes that the relative risk of death is proportional for different levels of a variable throughout the entire follow-up period (130). We evaluated the validity of this assumption using Schoenfeld residuals and graphical methods (130). There are multiple ways to handle non-proportionality, should it arise. A simple method is to create a stratified regression model using a categorical variable with non-proportional hazards (130, 132). A second method involves splitting the follow-up period to obtain proportionality in each individual period (132). We used both of these methods in Paper 2.

Lastly, we performed a sensitivity analysis to evaluate the effect of including women with missing tumour diameter information on the observed hazard ratios. To this end, we repeated the Cox regression methodology described above using the larger sample, but excluded tumour diameter as a covariate in the regression models.

The primary analysis in Paper 3 described women's knowledge about breast cancer screening and overdiagnosis. We used frequencies and proportions to describe how women responded to different questions in our questionnaire. Further, we calculated the mean number of marks awarded in each of the three thematic categories covered by the questionnaire, and overall. In this paper, we tested the null hypothesis that certain demographic variables were not independent of women's answers to questions about screening. Because this analysis was based on a survey sample, we used the second-order correction for Pearson's chi-square test statistic described by Rao and Scott (133).

Main findings

Paper 1 – Terminal digit preference

The age of study subjects ranged from 23–103 years (mean 62). Fifty seven percent of women were aged 50–69 and 35% of all cancers were screen-detected.

Breast radiologists and pathologists tended to record rounded tumour diameter measurements from mammograms and histopathologic specimens, respectively. National and international guidelines indicate that measurements should be made to the nearest millimetre (8, 9, 13, 134), but many tumour diameters registered at the Cancer Registry of Norway appeared to be rounded to the nearest whole or half-centimetre (Figure 13).

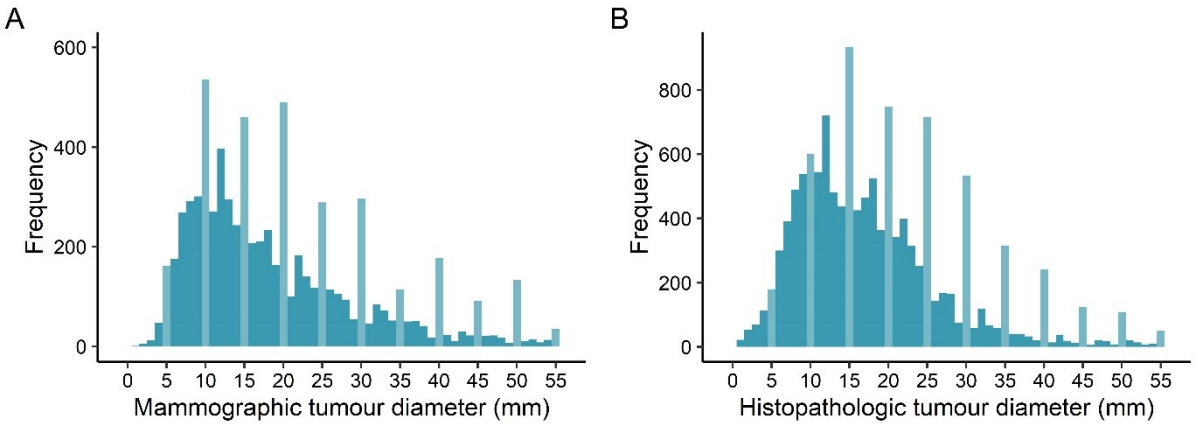


***Figure 13**: Longest recorded tumour diameter for T1–T3 breast cancers ≤ 55 mm diagnosed during 2012–2016, as measured by **(A)** radiologists from mammography (7792) and **(B)** pathologists from surgical specimens (cases reported at whole numbers only, n = 13,167)*

Overall, 38.7% of mammographic tumour diameters and 35.8% of histopathologic measurements were measured as a whole- or half-centimetre value. Radiologists demonstrated a preference for whole-centimetre values over half- centimetre values, but this was not observed among pathologists (Table 10).The proportion of rounded cases generally increased with increasing tumour diameter.

The distributions of tumour diameters observed in the full sample (Figure 13) were similar to those observed for screen-detected cancers (Figure 14). The distributions of tumour diameters for clinically detected cancers among women aged 50–69 also showed signs of terminal digit preference (Figure 15). Further, a high frequency of cases were reported as having a 20 mm diameter on mammography. A higher proportion of cases were reported as having whole-centimetre values by radiologists than by pathologists for screen-detected cancers (21.0% vs 16.5%) and clinically detected cancers (26.1% vs 18.7%) (Table 10).

***Table 10**: Proportion of tumours registered as a whole- or half-centimetre value (millimetre measurement with a terminal digit of 0 or 5) by radiologists from mammography or pathologists from surgical specimens for T1–T3 breast cancers diagnosed during 2012–2016*

Terminal digit	Overall[a]	Screen-detected[b]	Clinically detected[c]
Radiologists			
0	23.4%	21.0%	26.1%
5	15.3%	15.1%	14.4%
Total	38.7%	36.1%	40.5%
Pathologists			
0	17.9%	16.5%	18.7%
5	17.9%	16.5%	18.3%
Total	35.8%	33.0%	37.0%

[a] *Among women of all ages and through any mode of detection*
[b] *Among women attending BreastScreen Norway*
[c] *Interval cancers among women attending BreastScreen Norway and clinically detected cancers diagnosed among women aged 50–69 who did not attend BreastScreen Norway*

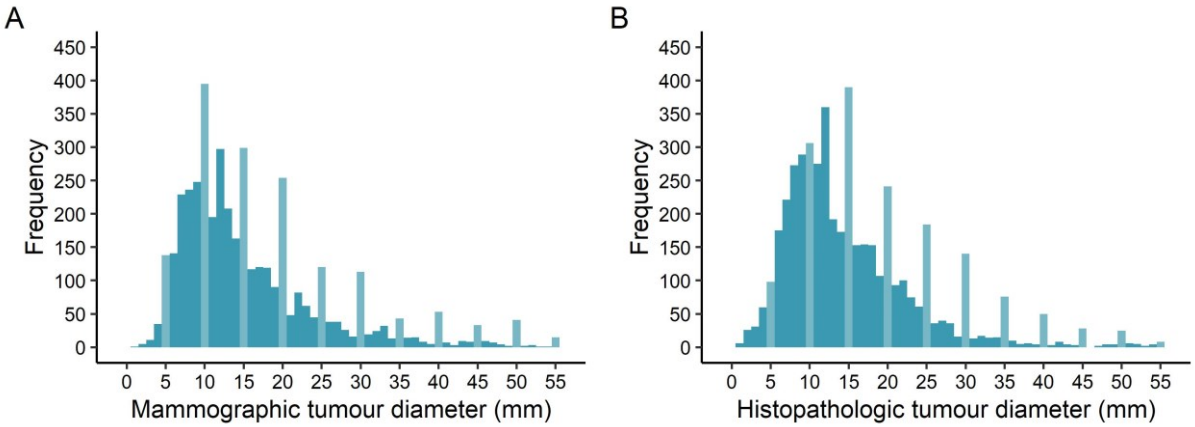


***Figure 14**: Longest recorded tumour diameter for* screen-detected *T1–T3 breast cancers ≤ 55 mm diagnosed during 2012–2016, as measured by **(A)** radiologists from mammography (n = 4357) and **(B)** pathologists from surgical specimens (cases reported at whole numbers only, n = 4839)*

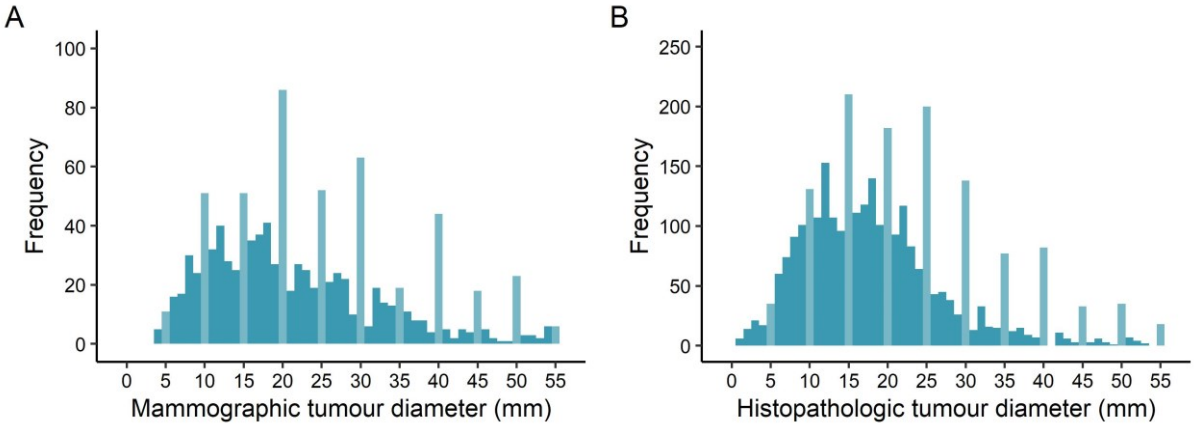


***Figure 15:** Longest recorded tumour diameter for* clinically detected *T1–T3 breast cancers ≤ 55 mm diagnosed during 2012–2016 among women aged 50–69, as measured by **(A)** radiologists from mammography (n = 1138) and **(B)** pathologists from surgical specimens (cases reported at whole numbers only, n = 3183)*

Paper 2 – Survival associated with true, minimal signs, and missed cancers

The mean age at diagnosis was similar for women with true, minimal signs, and missed screen-detected cancers (62, 62, and 63, respectively). It was also similar for women with true, minimal signs, and missed interval cancers (59, 60, and 61, respectively)

True cancers were associated with a number of less favourable tumour characteristics than minimal signs and missed cancers when stratified by mode of detection. Within each mode of detection, minimal signs and missed cancers had similar tumour characteristics. There was a higher proportion of histopathologic grade 3 tumours among true screen-detected cancers (30.0%) than minimal signs (14.9%), or missed screen-detected cancers (13.7%). Further, true screen-detected cancers were more likely to be triple negative than minimal signs or missed screen-detected cancers (9.8% versus 2.3% and 2.9%). Similarly, a higher proportion of true interval cancers were histopathologic grade 3 (46.7%) than minimal signs (36.1%) or missed interval cancers (35.9%). The proportion of triple negative true interval cancers was 18.1%, compared to 14.5% for minimal signs and 9.6% for missed interval cancers.

The median follow-up was 5.4 years (range 0.2–12.8) for women with screen-detected cancer and 5.6 years (range 0.3–14.8) for women with interval cancer. During the follow up period, there were 43 (4.2%) deaths from any cause among women with screen-detected cancer, and 81 (10.3%) among women with interval cancer. The Kaplan-Meier estimates of overall survival did not differ between true, minimal signs, and missed screen-detected (p = 0.82, Figure 16A) or interval cancers (p = 0.43, Figure 16B). Additionally, the Nelson-Aalen cumulative hazard estimates did not differ for interval breast cancer deaths from true (16 deaths), minimal signs (11 deaths), or missed cancers (12 deaths; p = 0.80, Figure 17).

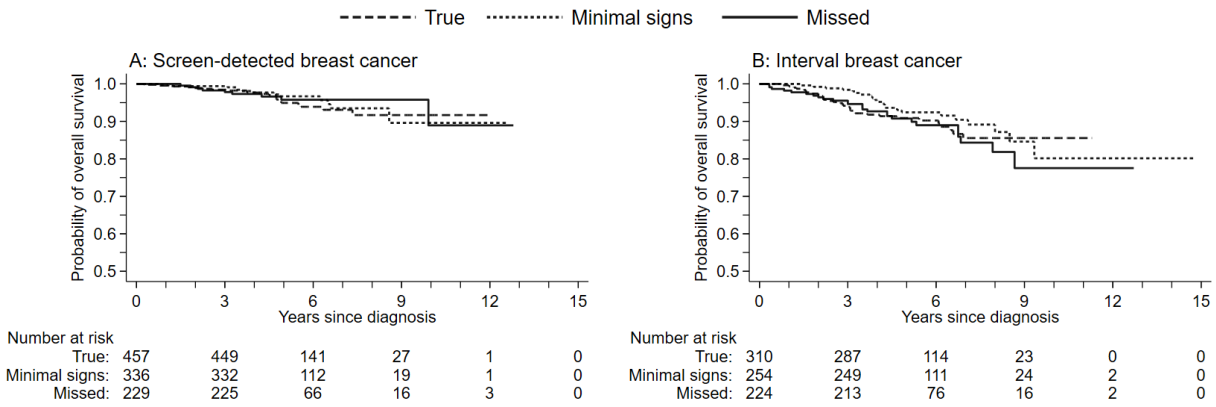


***Figure 16**: Kaplan-Meier estimates of overall survival for true, minimal signs, and missed **(A)** screen-detected, and **(B)** interval breast cancers.*

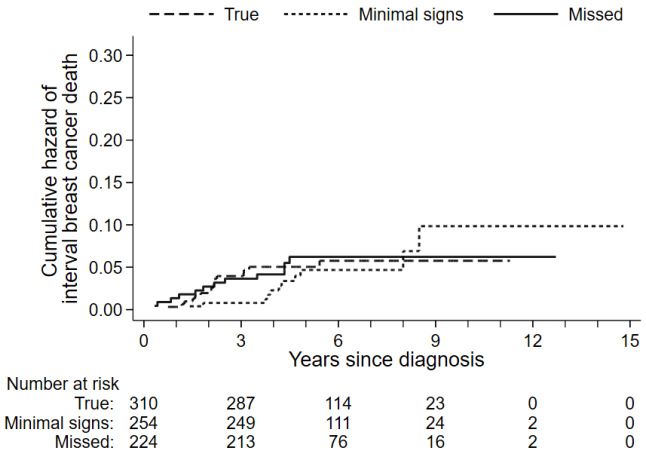


***Figure 17**: Nelson-Aalen estimate of the cumulative hazard of dying from interval breast cancer.*

Using the imputed data sets in the multivariable Cox regression, we found that the risk of death from any cause did not differ between women with minimal signs screen-detected cancers and women with true screen-detected cancers [HR = 1.04, 95% CI (0.51, 2.13)]. We obtained similar results for women with missed versus true screen-detected cancers [HR = 1.10, 95% CI (0.49, 2.46)]. Due to lack of proportional hazards, we split the follow-up time at three years after diagnosis for women with interval cancers. Women with minimal signs interval cancers had a lower risk of death from any cause than those with true interval cancers during the first three years after diagnosis [HR = 0.29, 95% CI (0.10, 0.86)]. This finding did not persist after the first three years [HR = 1.40, 95% CI (0.70, 2.80)]. Women with missed and true interval cancers had a similar risk of death from any cause both before and after three years of follow up. Our main conclusions did not change based on results from the sensitivity analysis.

Paper 3 – Women's conceptual knowledge about screening and overdiagnosis

Compared to statistics from Statistics Norway, women who responded to our survey were, on average, younger and had longer formal post-secondary education than the female population aged 45–74. After weighting, the marginal distributions of age, education, and region were similar for participants and the population. Over half of women in the weighted sample (52.4%) reported that they had previously looked up at least some information about mammographic screening.

Roughly 40% of women indicated that they had heard of a "false positive mammography examination" (Figure 18A), and roughly half (51.3%) indicated that they had heard of "overdiagnosis" (Figure 18B). Similar proportions of women indicated that they were unsure of having heard either term (8.7% for false positive screening examinations and 10.7% for overdiagnosis).

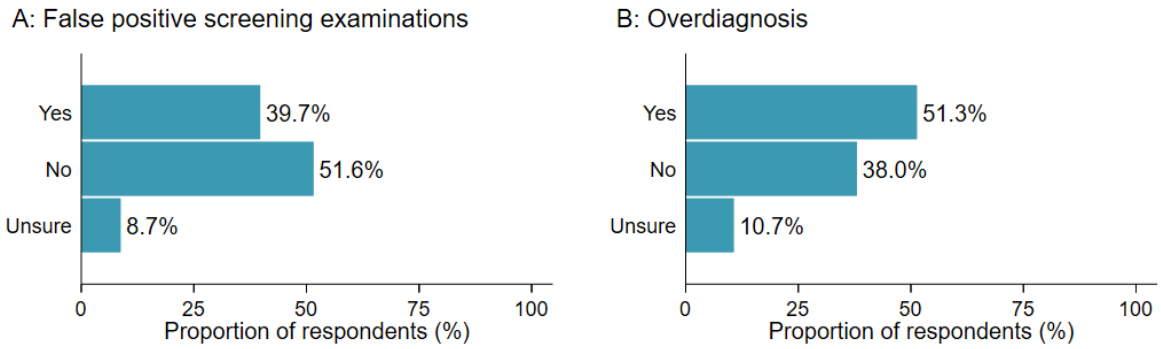


***Figure 18**: Proportion of women reporting whether they had heard of the terms "false positive screening examination" (panel A) or "overdiagnosis" (panel B), n = 1892.*

One question asked women to select the correct definition for "false positive screening examinations" from a definition for a false positive screening examination, a definition for overdiagnosis, and "unsure". A similar question asked women to select the correct definition of overdiagnosis from the same alternatives. The majority of women (86.2%) chose the correct definition for the question about false positive screening examinations and a minority (14.8%) chose the correct definition for the question about overdiagnosis. Most women (63%) chose the definition for false positive screening examinations to describe both terms.

Questions with the highest proportion of "unsure" responses were related to overdiagnosis: 22.1% of women were unsure if they had ever heard of the term "overdetection" and 17.1%

were unsure which of the two aforementioned definitions best represented the term “overdiagnosis”. The proportion of “unsure” responses for all other questions was ≤ 12%.

With respect to the graded questions, the mean score for the two questions about the breast cancer mortality benefit was 2.59 of a possible 3 (Figure 19A), and it was 0.93 of a possible 1 for the question about false positive screening examinations (Figure 19B). The mean score was 2.23 of a possible 6 for questions about overdiagnosis (Figure 19C). The mean total score was 5.75 of a possible 10.

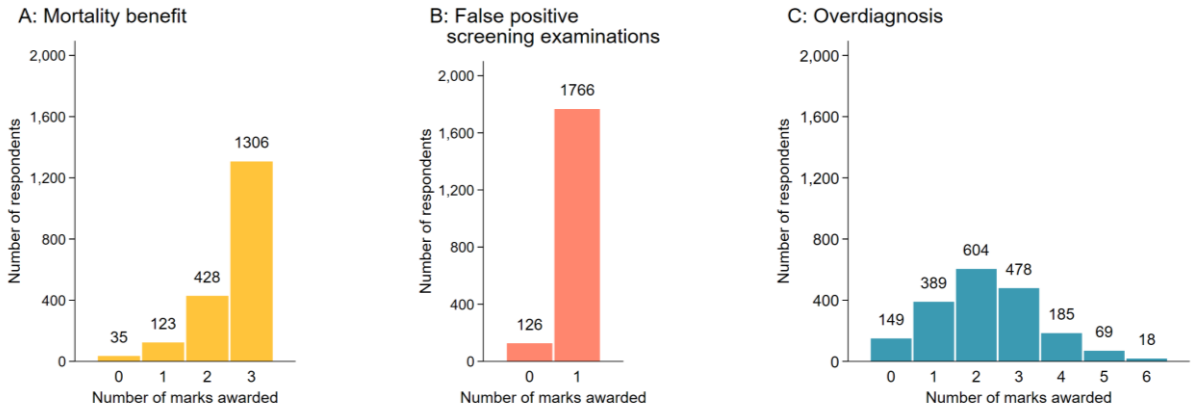


Figure 19: *Distribution of marks assigned for correct responses in the weighted study sample (n = 1892) within the thematic categories of* ***(A)*** *breast cancer mortality benefit,* ***(B)*** *false positive screening examinations, and* ***(C)*** *overdiagnosis. Questions and answers were adapted from Hersch J, Barratt A, Jansen J, et al. Use of a decision aid including information on overdetection to support informed choice about breast cancer screening: a randomised controlled trial. The Lancet. 2015;385(9978):1642-52.*

Discussion

Interpretation of main findings

Under- and overstaging

Paper 1 investigated whether terminal digit preference was present among tumour diameters reported to the Cancer Registry of Norway. We analysed tumour diameters for breast cancers diagnosed during 2012–2016 from any mode of detection among women of all ages and investigated whether this could lead to over- or understaging. We did not include information about terminal digit bias stratified by mode of detection in Paper 1, but included supplemental analyses in this thesis.

Our findings confirmed previous reports of terminal digit preference among breast pathologists (88-90, 135), and demonstrated that this preference also exists among breast radiologists. Overall, we found that terminal digit preference among radiologists and pathologists appeared to be more prevalent among larger tumours where the relative effect of this measurement error would have been lesser than that for smaller tumours. We suggested that understaging could occur if tumour measurements were rounded down to the nearest 5 mm increment (i.e. whole or half centimetre value) and then fell on boundary values that define a T stage. We posited that overstaging could not occur if tumour measurements were rounded up to the nearest 5 mm increment because the lower bounds of T stages are exclusive. Examples of rounding and its potential effect are given in Table 11.

Table 11: *Fictional examples of how preferential rounding due to terminal digit preference could lead to understaging for breast cancers with a tumour diameter close to boundary values*

Before rounding		After rounding		Potential effect
Tumour diameter	**T stage**	**Tumour diameter**	**T stage**	**on T stage**
18 mm	T1	20 mm	T1	None
21 mm	T2	20 mm	T1	Understaged
49 mm	T2	50 mm	T2	None
52 mm	T3	50 mm	T2	Understaged

In our sample, roughly 36% of histopathologic measurements were measured as a whole- or half-centimetre value, compared to 23% in a registry-based study performed by den Bakker and Damhuis in the Netherlands (88). In the latter, breast pathologists appeared reluctant to record tumour diameters that fell on T stage category boundary values. Although the TNM

guidelines clearly specify categorization rules for tumours that are on boundary values (9), den Bakker and Damhuis speculated that clinicians may hesitate to set tumour sizes on thresholds that may require additional discussion at multidisciplinary team meetings (88). We did not observe a reluctance among Norwegian radiologists or pathologists to record tumour diameters that fell on a T stage category boundary, which likely explains why a higher portion of cases were measured as a whole- or half-centimetre value in our study. In fact, we observed that many clinically detected tumours had a mammographic tumour diameter of 20 mm (T1 upper bound). We hypothesize that the risk of understaging due to terminal digit preference is mediated by tumour characteristics (including tumour size) and regional differences in clinical practice.

When stratified by mode of detection, we observed that a lower proportion of screen-detected cancers had a tumour diameter with a terminal digit of 0 or 5 than clinically detected cancers among similarly-aged women. This finding held for mammographic and pathologic assessments. It is difficult to estimate the true proportion of rounded measurements without reviewing the original samples, but this finding may suggest that terminal digit preference is somewhat less prevalent among screen-detected cancers. This can potentially be explained by the fact that, overall, terminal digit preference was less prevalent among smaller tumours, which are disproportionately screen-detected. If this understaging resulted in undertreatment, it would have disproportionally affected women diagnosed with interval cancers or diagnosed outside BreastScreen Norway. We are not aware of any other studies that have stratified analyses of terminal digit preference by mode of detection.

Overdiagnosis

Paper 2 considered whether women with missed cancers that were subsequently screen-detected benefited from their diagnosis, or whether they could have been overdiagnosed. A woman could have been overdiagnosed with a missed screen-detected cancer if her cancer represented slow-growing or indolent disease that would never have caused symptoms or death, no matter how long she lived. Assuming that all missed screen-detected cancers were asymptomatic when diagnosed, it is possible that some women diagnosed with missed screen-detected cancers have been overdiagnosed.

We hypothesized that if women with missed screen-detected cancers were overdiagnosed, they would have a longer (infinite) lead time and therefore better survival than women with true screen-detected cancers. However, we did not observe a difference in the overall survival for women with missed and true screen-detected cancers in our study sample, and posited that overdiagnosis was not highly prevalent among women with missed screen-

detected cancers. Some individuals with missed screen-detected cancers may have been overdiagnosed, but it is not possible to identify who based on the scientific knowledge available about overdiagnosis. Breast cancer specific outcomes may provide more insight into this epidemiological perspective on possible overdiagnosis. However, the 10-year relative survival of women diagnosed with breast cancer is 85.8% in Norway (15), and a larger sample and/or longer follow-up would be required to obtain over 80% statistical power to investigate such outcomes. Additionally, advances in treatment over the course of a decade can be significant and may make the individual effects in such studies difficult to disentangle.

Women's knowledge about overdiagnosis is important in the context of making an informed choice whether to attend screening. Paper 3 described the knowledge women in Norway have about breast cancer screening with a focus on overdiagnosis. Although half (51.3%) of participants reported having heard of the term "overdiagnosis", a minority (14.8%) correctly identified its definition. Our results are similar to those reported a British study, in which 64% of women eligible for organized mammographic screening reported being aware of overdiagnosis (110). Our results are also corroborated by an Australian study evaluating a decision aid, which reported that only 12% of women who had not been provided with information about overdiagnosis (i.e. in the control group) correctly distinguished overdiagnosis from false positive screening examinations (136). Overall, the results reported in Paper 3 indicated that women's conceptual knowledge about overdiagnosis was relatively lower than their knowledge about the breast cancer mortality benefit and false positive examinations associated with screening. However, the varying number of questions on each topic may have also affected these results.

Since a high proportion of women have reported reading at least some of the information developed and provided by BreastScreen Norway (91, 92), the results from our study may suggest that women find it difficult to fully understand overdiagnosis as described in the information material revised in 2017. However, overdiagnosis in mammographic screening has been intensely discussed in newspapers and on high profile radio and television shows in the Norwegian mainstream media in recent years, including *Dagsnytt 18*, *Folkeopplysningen,* and *Debatten* (137-140). This could also affect women's understanding about overdiagnosis. These programs have highlighted disagreements between experts about the extent of overdiagnosis and may not provide clear information about this complicated topic for the public.

Underdiagnosis

Paper 2 touched on the topic of underdiagnosis due to a false negative screening examination and delayed breast cancer diagnosis that lead to a diagnosis of screen-detected breast cancer in the subsequent screening round or interval cancer (i.e. “missed” screen-detected or interval cancers). Women with missed screen-detected cancers could be underdiagnosed if their cancers have more aggressive tumour characteristics and women experience poorer survival outcomes than women with true screen-detected cancers. In this study, we observed that missed screen-detected cancers had similar or more favourable histopathology than true screen-detected cancers. Further, women with missed screen-detected cancers did not have worse overall survival than women with true screen-detected cancers. Since these women were diagnosed with breast cancer as a result of their routine screening examination, we assumed they did not experience any clinical symptoms at the time of screening. It follows that, as a group, women with missed screen-detected cancers were not underdiagnosed and likely would not have benefited from earlier detection. As discussed above, it also seems unlikely that overdiagnosis is a major concern among women with missed screen-detected cancers. Thus, in terms of tumour histopathology and survival outcomes, these women appear to have received a timely diagnosis of breast cancer.

Missed interval cancers can be considered underdiagnosed in the sense that they could have been diagnosed earlier as asymptomatic screen-detected cancers, which, on a population level, have more favourable tumour histopathology and better survival than symptomatic interval cancers (45-47). Paper 2 did not provide information about whether missed interval cancers would have had a favourable prognosis had they been detected earlier as screen-detected cancers. Instead, we compared true and missed interval cancers and evaluated whether true cancers were associated with a less favourable prognosis than missed cancers.

We observed that true interval cancers were more likely to be grade 3 than missed cancers, which is supported by other peer-review studies (82-84, 86, 87). We also observed that true and missed interval cancers had a similar tumour diameter. This diverges from previously published results indicating that true interval cancers are smaller (82, 83, 85-87) and may partially be because we reported the median diameter, which is less sensitive to outliers than the mean. Another possible explanation is that a higher proportion of women with larger cancers were offered neoadjuvant therapy in our contemporary sample and excluded from our analysis. We also observed that true interval cancers were more likely to be triple negative than missed interval cancers. Overall, although true interval cancers were associated with some less favourable tumour characteristics, we did not observe any differences in the risk of death from breast cancer for women with true or missed interval

cancers. This may be due to the availability of effective treatment options for more aggressive cancers, but is likely also be related to the precision of our results.

Precision of results

The results from Paper 1 are based on a nation-wide registry with highly complete pathological data. However, 45% of women in our study sample were missing information about mammographic tumour diameter. The proportion of missing data was 14.2% among screen-detected cancers and 63.4% among clinically detected cancers. The reasons for this are discussed in the section *Missing data* (page 48). This missing data would have decreased the precision of the proportions estimated in Paper 1. However, mammographic tumour diameter information was available for nearly 8000 women overall and the effect of random error was therefore likely minimal. With respect to the stratified analyses, over 4000 women were included in the screen-detected group. The clinically detected group was limited to women aged 50–69 for comparability, and mammographic tumour diameter information was available for 1138 women. The histograms plotted these data in 1 mm increments over a range of 55 mm and the number of observations in some “bins” was scarce. This would have decreased the precision of the observed results, particularly where smaller differences were observed between groups.

The primary results from Paper 2 were derived from Cox regression models generated using data from a sample of women whose mammograms were included in a retrospective radiologic review. The purpose of this review was “quality improvement for radiologists’ performance and the program” (116). Radiological reviews are resource intensive and the sample size in Paper 2 was limited by a number of practical factors including the number of cases that could be reviewed by radiologists and the availability of digital screening and diagnostic mammograms from lower volume centres. Further, the median follow-up time in this study was roughly 5.5 years and there were a modest number of deaths during the follow-up period; the five-year relative survival from breast cancer has been over 88% since 2005–2009 (15). The sample size, modest number of deaths, study specific exclusion criteria, missing histopathology data, and adjustment for covariates limited the statistical power of the Cox regression models and decreased the precision of the resulting estimates (141). It is worth noting, however, that the adjustments improved the internal validity of our estimates. Although the adjusted complete case analyses left out over 25% of the available observations due to missing histopathology information, we used multiple imputation to overcome this limitation in our primary analysis. The final sample size was modest, but our

study is one of the largest to evaluate the overall survival associated with missed and true cancers in population-based screening to date. This is an important strength of our work.

In Paper 3, we reported point estimates for proportions and means, but no corresponding confidence intervals or standard deviations. The sample size in this study was large enough to produce stable estimates of proportions and means (n = 1892) and small changes to the study weights would not have substantially affected our results (this is can be observed by comparing the unweighted and weighted results). With respect to the means, these corresponded to the mean number of marks awarded to women for correct answers to graded questions and was interpreted with respect to the total number of marks available. Although we did not present standard deviations for these estimates, we provided histograms that showed the distribution of the underlying data for each of these means. From these, we can observe that there is less variability in the mean marks for questions about the breast cancer mortality benefit and false positive screening examinations associated with screening than for questions about overdiagnosis or overall.

Internal validity

In Paper 1, mammographic tumour diameter information was missing for nearly half of the sample. As noted earlier, women without a record of this information are more likely to have clinically diagnosed cancers. Indeed, screen-detected cancers were overrepresented in the final sample. Because our results indicated that terminal digit preference was more frequent among larger (i.e. clinically detected) tumours, the direction of this selection bias was likely was toward the null (less/no preferential rounding). The proportion of mammographic tumour diameters reported with a terminal digit of 0 or 5 is therefore likely to be higher in the population than observed in our study. Moreover, the overall distribution of mammographic tumour diameters in the population may be less right skewed than our figures suggest.

The analysis in Paper 2 used multiple imputation with chained equations to overcome the limitations associated with incomplete histopathology data (125). This technique can produce unbiased results if the data are missing at random, but results may be biased if the data are missing not at random (142). We assumed that missing histopathologic tumour diameter data was missing not at random because the probability of missingness was associated with neoadjuvant therapy and, therefore, tumour diameter. However, tumour diameter is an important confounder in the relationship between being diagnosed with a true, missed, and minimal signs cancer, and survival (81, 85-87, 143). In Paper 2, we excluded women who were missing tumour diameter information in order to adjust for this confounder in the Cox regression models, but this would have introduced a selection bias into this analyses. For

this reason, we performed a sensitivity analysis in which we included women with missing information about tumour diameter in the Cox regression models (but did not include it as a covariate in the model to prevent women with missing data from being excluded by the software). The results of the sensitivity analysis did not change the conclusions we drew from the original analysis.

Information bias is a relevant concern in Paper 3. We based the conceptual questions and rubric used in our study on those developed by Hersch and colleagues (119). These questions and rubric have been used in randomized controlled trials to evaluate the effect of a decision aid on informed choice in the context of deciding whether to participate in organized screening (119, 144). The multiple choice and true/false nature of the questions we used simplified complex and debated topics in breast cancer screening, particularly estimates of overdiagnosis (69). However, we felt these conceptual questions would be more valid in Norway than the corresponding set of numeric questions. In fact, our decision to exclude numeric questions from our study was heavily influenced by challenges in establishing a conclusive Norwegian reference standard to questions like, “If these 1000 [ordinary women who are 50 years old] have screening every 2 years for 20 years, in that time about how many will be diagnosed and treated for a breast cancer that is not harmful?” (119). Estimates of overdiagnosis are highly sensitive to the assumptions and methods used to determine them (69). It is difficult to estimate the prevalence overdiagnosis in most organized screening programs, but the non-random, stepwise implementation of BreastScreen Norway further complicates such estimations. Published rates range from 0 to over 75%, making it difficult to determine a “correct” reference standard (43, 73, 74). Consequently, we excluded numeric questions from our questionnaire.

Feedback in the pilot testing indicated that women felt the questions were easy to understand. However, it is possible that women did not understand or misinterpreted the conceptual questions used in our questionnaire and this could have affected the internal validity of our conclusions. We often provided an “unsure” option that women could select if they were not certain of the correct answer and only 6% (n = 128) of participants were excluded due to missing data caused by not responding to a question at all. This could indicate that women felt sure enough about their understanding of the questions to provide an answer. The highest proportions of “unsure” responses were observed for questions related to overdiagnosis: 22.1% were unsure of having heard of the word “overdetection”, and 17.1% were unsure when asked to choose the correct definition of overdiagnosis. This finding is in keeping with published literature indicating that overdiagnosis is difficult to define and understand (106, 107). The proportion of “unsure” responses for all other questions was less than 13%. Guessing could have also affected the internal validity of the data we

collected in this study, particularly since only one question about false positive screening examinations was included in the set of graded questions. However, most questions had a relatively high ratio of correct to incorrect responses (or vice versa), and this is unlikely to have been a major limitation. Similarly, although women could have looked up correct answers to questions they were unsure about online, the overall score for graded questions was 5.75 of a possible 10 and we do not think this was a major threat to the internal validity of this study.

Selection bias is also a relevant concern with respect to Paper 3. The population of interest was all women aged 45–74 living in Norway, but the study was based on a convenience sample recruited primarily through a Facebook page administered by the Cancer Registry of Norway. The resulting selection bias was mainly due to internet and social media access precluding some women from having access to the survey, and a volunteer bias related to women deciding to complete the survey. Regarding the former, 92% of women in Norway aged 45–54 and 60% of women aged 65–74 used social media in late 2018 (145). Further, 87% of Norwegian residents aged 16–74 communicate with public authorities online (146). Regarding the latter, participants were younger and had longer formal education compared to population statistics from Statistics Norway. Even after weighting the sample, a higher proportion of women in our study indicated that they had looked up information about screening (52.4%) than was reported among a sample of women attending screening in 2015 (< 20%) (92). Thus, this volunteer bias likely persisted in the weighted sample. We hypothesize that participants were more interested in, or knowledgeable about, breast cancer screening than women of a similar age in the general population. As such, we posit that the estimates derived from our survey overestimate those that would be observed from a more representative sample of general population.

External validity

Breast cancer screening programs have been established worldwide and under- and overdiagnosis are inevitable consequences of any such program based on the knowledge that exists today about the natural history of breast cancer. This makes the results presented in this thesis of interest internationally. However, screening programs in countries outside of Europe, the United States in particular, are administered very differently than programs that follow the European Guidelines (147). Moreover, international treatment guidelines and clinical practice may also differ. This limits the degree to which the results included in this thesis can be generalized to other countries.

With respect to Paper 1, the preferential rounding appears to be common internationally and across cancer types. It has been reported regarding the measurement of breast cancers in Sweden, the Netherlands, and the United States, (88-90), as well as the measurement of colorectal cancers (148), kidney cancers (149), lung cancers (88, 150), and melanomas (151) in the United States, the United Kingdom, and Norway. It therefore seems reasonable to hypothesize that terminal digit preference is present in tumour diameter data for other cancer types in Norway and other countries. However, the “stage border avoidance” observed by den Bakker and Damhuis in the Netherlands was not corroborated by our findings, which could indicate that specific findings related to terminal digit preference may not be as generalizable between settings (88).

Another important factor limiting the external generalizability of the results described in Paper 1 and Paper 2 is the exclusion of women with missing tumour diameter data, particularly related to those who received neoadjuvant therapy. Because these tumours are likely inherently different from those diagnosed in women who do not receive neoadjuvant therapy, we caution against generalizing our results to women who have received neoadjuvant therapy. This may not substantially affect the external generalizability of the results of Paper 1 since this exclusion does not have a major impact on our overall conclusions. However, it is an important limitation for the interpretation of results from Paper 2 because these women are likely to have a worse prognosis (and therefore poorer survival) than the women who were included in our main analysis.

In Paper 2, the exposure classifications of true, minimal signs, and missed cancers were made through an informed, consensus-based radiological review. The methodology used to classify cancers affects the proportion of cases in each category and the informed nature of the review likely resulted in a higher proportion of missed cancers than would have been observed in a blinded review setting (48-51). Additionally, the proportion of cases in each category was also affected by the classification definitions used in this study. A previous informed, consensus-based review performed in Norway using different classification definitions observed a lower proportion of missed cancers and a higher proportion of minimal signs cancers (54). This illustrates the challenge in generalizing results from review studies, even those conducted within a screening program. Across screening programs, rates of true and missed cancers are also affected by organizational factors including screening sensitivity and screening intervals. Because of this, the underlying rates of true, minimal signs, and missed cancers – whether screen-detected or interval – is probably unknowable. Overall, our survival results from Paper 2 corroborate the finding of no difference in overall survival between true and missed cancers observed in three studies conducted in other countries

with different review methods, but this may be related to the precision of estimates in all studies (83, 84, 86).

The results in Paper 3 reflect conceptual knowledge among a select group of women residing in Norway, but can be compared to results from Australia (Hersch et al.) and Spain (Pérez-Lacasta et al.) derived from a similar questionnaire (119, 144). These studies both used a randomized controlled trial to investigate the effect of a decision aid on informed choice – women in these studies were aged 48–50, but results from the control groups offer an opportunity to explore the potential external validity of our results. Our study reported a similar mean number of marks assigned within the topic of the breast cancer mortality benefit and false positive screening examinations associated with screening. With respect to overdiagnosis, the mean number of marks awarded was lower in our study than observed in Australia or Spain (2.32, 3.48, and 2.88, respectively). This may be because we asked women one fewer question. Overall, it appeared that our results may be broadly generalizable to other Western countries with organized screening programs, however, there was some variation between studies in women's responses to individual questions. For example, 49% of women in our study incorrectly reported that the statement "all breast cancers will eventually cause illness and death if they are not found and treated" compared to 33% in Hersch et al. (119), and 24.9% in Pérez-Lacasta et al. (144). This could reflect differences in study design and sampling, but likely also reflects local variations in knowledge that limit the external generalizability of results from these types of questionnaires.

Conclusions

In the first study (Paper 1), we aimed to determine whether terminal digit preference was present among tumour diameter data at the Cancer Registry of Norway. We observed that breast radiologists and pathologists exhibited a preference for 0s and 5s as a terminal digit when they reported tumour diameter measurements to the Registry. We also described a mechanism by which this preferential rounding could lead to potentially understaged breast cancers among women of all ages. Additional analyses included in this thesis demonstrated that terminal digit preference disproportionally affected cancers detected outside of organized screening among women aged 50–69.

In the second study (Paper 2), we aimed to determine whether a diagnosis of true, minimal signs, or missed cancer was differentially associated with tumour histopathology and women's survival when stratified by mode of detection. Further, we explored whether this information could be used to indicate whether missed cancers were over- or underdiagnosed. Within each mode of detection, we observed that missed cancers had

some favourable tumour characteristics compared to true cancers, but women in all groups experienced similar overall survival. Based on the results of this study, we posited that women with missed screen-detected cancers received a timely diagnosis and were not likely to be over- or underdiagnosed. Further, we posited that missed interval cancers were not underdiagnosed compared to true interval cancers. Some of these cancers may have been underdiagnosed in the sense that they could have benefited from being diagnosed earlier as screen-detected cancers, but this was not examined in this study.

In the third study (Paper 3), we aimed to address the research gap that exists around Norwegian women's conceptual knowledge about mammographic screening in general, and overdiagnosis in particular. We identified that women in Norway reported less knowledge about overdiagnosis than about the breast cancer mortality benefit and false positive screening examinations associated with organized screening.

Directions for future research

The papers in this thesis investigated overdiagnosis and underdiagnosis within BreastScreen Norway; they also describe a mechanism for potential understaging in breast cancer diagnosis. The results of these studies ought to have implications for radiologists, pathologists, administrators, and other stakeholders in organized screening programs and should be important to consider going forward to develop a multifaceted understanding of the potential harms associated with organized mammographic screening. The results described in these papers also point to a number of areas for future research.

Tumour diameter provides important information about a woman's prognosis and is therefore a key variable for breast cancer researchers. Accurate tumour diameter measurement can be affected by a variety of factors including mammographic density, multifocal or diffuse lesions, and histopathologic grade (152-154). Some rounding may therefore be expected for cases that are challenging to measure accurately and terminal digit preference may play a role in the outcome of such measurements. Indeed, terminal digit preference may be more prevalent among cases with certain tumour characteristics that make them borderline candidates for certain therapies. The relationship between terminal digit preference and tumour histopathology has not been investigated and could provide more information about the potential clinical implications associated with this type of rounding, particularly for women who are diagnosed with interval breast cancers or outside BreastScreen Norway.

Missing information about histopathologic tumour diameter (pT) affected the external generalizability of the results described in Papers 1 and 2. Although it is not often discussed, this is a challenge for many studies that include information about histopathologic tumour diameter because standard regression models omit observations with missing data (155).

Moreover, this disproportionally excludes information from women who have larger or more aggressive cancers (e.g. diagnosed with breast cancer outside of an organized screening program), which may threaten the external generalizability of research results. Histopathologic tumour diameter data is difficult to impute because it is completely missing for all women who receive neoadjuvant therapy. As more women are offered this type of therapy, radiologic measurements of tumour diameter, such as those from mammography, ultrasound, or magnetic resonance imaging (MRI), may provide a surrogate for histopathologic information. However, this information is missing at the Cancer Registry of Norway for a substantial proportion of women diagnosed with cancers outside the screening program (11). Moreover, radiologic measurements cannot always measure the same aspects of a tumour as histopathologic measurements. For example, the longest axis may be different on standard 2D mammography compared to a surgical specimen, and the limits of agreement between radiologic and histopathologic measurements can be wide enough to have clinical implications (156, 157). A discussion of how to address the challenge of missing tumour diameter information in research should be a priority.

As highlighted in Paper 2, review studies can provide important information about the prognosis of women with potentially over- or underdiagnosed cancers. However, review studies are resource intensive and death is relatively rare among women with breast cancer. This limits the statistical power of such studies. Future studies could consider treatment regimens and quality of life measures as alternative endpoints to generate shorter-term information about outcomes for women with missed and true cancers. Longer follow up would be advantageous to improve the precision of published estimates of overall survival and provide more information about breast cancer specific outcomes that could better relate to over- and underdiagnosis, particularly for women with screen-detected cancers. International multi-centre studies and meta-analyses offer another potential solution to improve the precision of published survival estimates, but heterogeneity in the organization of screening programs and review study designs likely precludes this from being a viable solution. Machine-learning algorithms could review prior screening mammograms and triage or identify true negative screens in the future (158). Since true interval cancers are typically the largest group in review studies, this has the potential to reduce the classification workload for radiologists and simultaneously facilitate larger review studies (46). However, this could detract from the educational aspect of such review studies.

As described earlier, the definition of underdiagnosis presented in this thesis suggests that missed interval cancers are underdiagnosed. It follows that decreasing the number of missed interval cancers is desirable. However, increasing a radiologist's sensitivity to reduce the rate of missed interval cancers could result in decreased specificity. Since most women do not

have breast cancer, this could unnecessary increase the rate of false positive screening examinations. Moreover, increased sensitivity could also increase the risk of overdiagnosis since some women may be recalled for findings that, previously, would not have warranted a recall and some of these women could risk being diagnosed with a slow-growing or indolent breast cancer as a result. Studies evaluating interventions to improve screening sensitivity (e.g. screen-detected cancer rates) should also focus on screening specificity (e.g. recall rates), positive and negative predictive values, and interval cancer rates. Additionally, information about tumour characteristics, such as tumour morphology, diameter, and grade, should be reported, as this may provide insights into the potential for overdiagnosis associated with any gains in screening sensitivity.

Many European screening programs have policies in place to enable women to make an informed choice about screening, which includes providing information about different aspects of screening, including overdiagnosis. Providing information about overdiagnosis can be challenging because epidemiological and biological knowledge about this complex topic in mammographic screening is limited. The results from Paper 3 do not provide information about women's values and cannot be used to provide information about whether women in Norway can make an informed choice about screening. Focus group studies centered on the topic of overdiagnosis may provide information about where women have previously heard of the term, what specific aspects of overdiagnosis are unclear to them, and how they would like these to be explained to them. This research may be used to generate women-oriented solutions for more clearly conveying information about this challenging topic, for example with respect to the content or format of this information. BreastScreen Norway continually updates its information materials and knowledge gained from this type of research could help improve the quality of information about overdiagnosis available to women in the future. In turn, this would further the program's goal of enabling women to make an informed choice about screening.

References

1. Sanders-Goebel P. Crisis and controversy: historical patterns in breast cancer surgery. Can Bull Med Hist. 1991;8(1):77-90.
2. Baum M. Modern concepts of the natural history of breast cancer: a guide to the design and publication of trials of the treatment of breast cancer. Eur J Cancer. 2013;49(1):60-4.
3. Erbas B, Provenzano E, Armes J, Gertig D. The natural history of ductal carcinoma in situ of the breast: a review. Breast Cancer Res Treat. 2006;97(2):135-44.
4. International Agency for Research on Cancer. Breast Cancer Screening. IARC Handbook of Cancer Prevention. Volume 15. Lyon, France; 2016.
5. Schnitt SJ. Diagnosis of ductal carcinoma in situ in an era of de-escalation of therapy. Mod Pathol. 2021;34(Suppl 1):1-7.
6. Jatoi I. Breast cancer: a systemic or local disease? Am J Clin Oncol. 1997;20(5):536-9.
7. Goldhirsch A, Wood WC, Coates AS, et al. Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol. 2011;22(8):1736-47.
8. Union for International Cancer Control. TNM classification of malignant tumours. 8th ed. Oxford Hoboken: John Wiley & Sons; 2017.
9. Badve SS, Beitsch PD, Bose S, et al. Part XI, Breast. In: Amin MB, Edge S, Greene F, Byrd DR, Brookland RK, Washington MK, et al., editors. AJCC Cancer Staging Manual. 8th ed: Springer International Publishing; 2017.
10. Hayes DF, Allred C, Anderson BO, et al. Part VII, Breast. In: Edge S, Byrd D, Compton CC, Fritz AG, Greene F, Trotti III A, editors. AJCC Cancer Staging Manual. 7th ed: Springer International Publishing; 2010.
11. Cancer Registry of Norway. Årsrapport 2019 med resultater og forbedringstiltak fra Nasjonalt kvalitetsregister for brystkreft. [National quality registry for breast cancer annual report, 2019.] Oslo: Cancer Registry of Norway; 2020. Norwegian.
12. Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24(9):2206-23.
13. Norwegian Directorate of Health. Brystkreft – handlingsprogram [National breast cancer guidelines] Oslo: Norwegian Directorate of Health; 2020 [updated August 14, 2020; cited January 1, 2021]. Available from: https://helsedirektoratet.no/retningslinjer/nasjonalt-handlingsprogram-med-retningslinjer-for-diagnostikk-behandling-og-oppfolging-av-pasienter-med-brystkreft.
14. Gluz O, Liedtke C, Gottschalk N, et al. Triple-negative breast cancer--current status and future directions. Ann Oncol. 2009;20(12):1913-27.
15. Cancer Registry of Norway. Cancer in Norway 2019 - cancer incidence, mortality, survival and prevalence in Norway. Oslo Cancer Registry of Norway; 2020.
16. World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and breast cancer. Available at http://dietandcancerreport.org.
17. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Combined estrogen-progestogen menopausal therapy. A review of human carcinogens - Part A: pharmaceuticals. IARC monographs on the evaluation of carcinogenic risks to humans. Volume 100A. Lyon, France: International Agency for Research on Cancer; 2012. p. 249-82.
18. Bakken K, Eggen AE, Lund E. Hormone replacement therapy in Norwegian women, 1996–1997. Maturitas. 2001;40(2):131-41.
19. Saether S, Bakken K, Lund E. The risk of breast cancer linked to menopausal hormone therapy. Tidsskr Nor Laegeforen. 2012;132(11):1330-4.
20. Bakken K, Alsaker E, Eggen AE, Lund E. Hormone replacement therapy and incidence of hormone-dependent cancers in the Norwegian Women and Cancer study. Int J Cancer. 2004;112(1):130-4.
21. Million women study collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study. Lancet. 2003;362(9382):419-27.
22. Collaborative group on hormonal factors in breast cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705

women with breast cancer and 108 411 women without breast cancer. Lancet. 1997;350(9084):1047-59.
23. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA. 2002;288(3):321-33.
24. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Combined estrogen-progestogen contraceptives and combined estrogen-progestogen menopausal therapy. Lyon, France: International Agency for Research on Cancer; 2007.
25. The Norwegian Institute of Public Health. Norwegian Prescription Database 2020 [cited December 30, 2020]. Available from: http://www.norpd.no/.
26. Louro J, Posso M, Hilton Boon M, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. Br J Cancer. 2019;121(1):76-85.
27. Porta M, editor. A Dictionary of Epidemiology. 6th ed. New York, NY: Oxford University Press; 2014.
28. Perry N, Broeders M, de Wolf C, et al., editors. European guidelines for quality assurance in breast cancer screening and diagnosis. 4th ed. Luxembourg: Office for Official Publications of the European Communities; 2006.
29. Wilson JMG, Jungner G, World Health Organization. Principles and practice of screening for disease. Geneva: World Health Organization; 1968.
30. Andermann A, Blancquaert I, Beauchamp S, Déry V. Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. Bulletin of the World Health Organization. 2008;86(4):317-9.
31. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme: Public Health England; [updated October 23, 2015; cited September 23, 2020]. Available from: https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes/criteria-for-appraising-the-viability-effectiveness-and-appropriateness-of-a-screening-programme.
32. Norwegian Directorate of Health. Styringsstruktur og strategi for nasjonale screeningprogrammer [Structure and strategy for national screening programs]. Oslo: Norwegian Directorate of Health; 2014. Report No. IS-2242. Norwegian.
33. Welch HG, Black WC. Overdiagnosis in cancer. J Natl Cancer Inst. 2010;102(9):605-13.
34. Gold RH, Bassett LW, Widoff BE. Highlights from the history of mammography. Radiographics. 1990;10(6):1111-31.
35. Gershon-Cohen J, Hermel MB, Berger SM. Detection of breast cancer by periodic x-ray examinations. A five-year survey. JAMA. 1961;176:1114-6.
36. Shapiro S. Periodic screening for breast cancer: the HIP Randomized Controlled Trial. Journal of the National Cancer Institute Monographs. 1997(22):27-30.
37. Pellerin R. Chatper 3: The mammogram wars begin. Conspiracy of hope. New Brunswick, Canada: Goose Lane Editions; 2018.
38. Rosen M, Rehnqvist N. Breast screening. No need to reconsider breast screening programme on basis of results from defective study. BMJ. 1999;318(7186):809-10.
39. Sjonell G, Stahle L. Scientific foundation of mammographic screening is based on inconclusive research in Sweden. BMJ. 1999;319(7201):55.
40. Kopans DB. The Canadian screening program - a different perspective. AJR Am J Roentgenol. 1990;155(4):748-9.
41. Boyd NF, Jong RA, Yaffe MJ, et al. A critical appraisal of the Canadian National Breast Cancer Screening Study. Radiology. 1993;189(3):661-3.
42. Rothman K, Greenland S, Lash T. Validity in epidemiologic studies. In: Rothman K, Greenland S, Lash T, editors. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
43. Falk RS. Why are results of organised mammography screening so difficult to interpret? Tidsskr Nor Laegeforen. 2014;134(11):1124-6.
44. Cancer Registry of Norway. Kvalitetsmanual [Quality manual] 2019 [updated September 2, 2019. Available from: https://www.kreftregisteret.no/Generelt/Rapporter/Mammografiprogrammet/Kvalitet/
45. Hofvind S, Tsuruda K, Mangerud G, et al. The Norwegian Breast Cancer Screening Program, 1996-2016: celebrating 20 years of organised screening in Norway. Oslo: Cancer Registry of Norway; 2017.

46. Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. NPJ Breast Cancer. 2017;3:12.
47. Hofvind S, Holen Å, Román M, et al. Mode of detection: an independent prognostic factor for women with breast cancer. J Med Screen. 2015;23(2):89-97.
48. Hofvind S, Skaane P, Vitak B, et al. Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology. 2005;237(2):437-43.
49. Lekanidi K, Dilks P, Suaris T, et al. Breast screening: What can the interval cancer review teach us? Are we perhaps being a bit too hard on ourselves? Eur J Radiol. 2017;94:13-5.
50. Ciatto S, Catarzi S, Lamberini MP, et al. Interval breast cancers in screening: the effect of mammography review method on classification. Breast. 2007;16(6):646-52.
51. Gordon PB, Borugian MJ, Warren Burhenne LJ. A true screening environment for review of interval breast cancers: pilot study to reduce bias. Radiology. 2007;245(2):411-5.
52. Moberg K, Grundstrom H, Lundquist H, et al. Radiological review of incidence breast cancers. J Med Screen. 2000;7(4):177-83.
53. Hoff SR, Samset JH, Abrahamsen AL, et al. Missed and true interval and screen-detected breast cancers in a population based screening program. Acad Radiol. 2011;18(4):454-60.
54. Hoff SR, Abrahamsen AL, Samset JH, et al. Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography - results from a retrospective review. Radiology. 2012;264(2):378-86.
55. Broeders MJM, Onland-Moret NC, Rijken HJTM, et al. Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. Eur J Cancer. 2003;39(12):1770-5.
56. Saarenmaa I, Salminen T, Geiger U, et al. The visibility of cancer on earlier mammograms in a population-based screening programme. Eur J Cancer. 1999;35(7):1118-22.
57. van Breest Smallenburg V, Setz-Pels W, Groenewoud JH, et al. Malpractice claims following screening mammography in The Netherlands. Int J Cancer. 2012;131(6):1360-6.
58. Cancer Registry of Norway. Cancer in Norway 2018 - cancer incidence, mortality, survival and prevalence in Norway. Oslo Cancer Registry of Norway; 2019.
59. Sebuødegård S, Botteri E, Hofvind S. Breast cancer mortality after implementation of organized population-based breast cancer screening in Norway. J Natl Cancer Inst. 2020;112(8):839-46.
60. Dibden A, Offman J, Duffy SW, Gabe R. Worldwide review and meta-analysis of cohort studies measuring the effect of mammography screening programmes on incidence-based breast cancer mortality. Cancers (Basel). 2020;12(4).
61. Baldeh T, Saz-Parkinson Z, Muti P, et al. Development and use of health outcome descriptors: a guideline development case study. Health Qual Life Outcomes. 2020;18(1):167.
62. Long H, Brooks JM, Harvie M, et al. How do women experience a false-positive test result from breast screening? A systematic review and thematic synthesis of qualitative studies. Br J Cancer. 2019;121(4):351-8.
63. Seaman K, Dzidic PL, Castell E, et al. A systematic review of women's knowledge of screening mammography. Breast. 2018;42:81-93.
64. Mathioudakis AG, Salakari M, Pylkkanen L, et al. Systematic review on women's values and preferences concerning breast cancer screening and diagnostic services. Psychooncology. 2019;28(5):939-47.
65. Carter SM, Rogers W, Heath I, et al. The challenge of overdiagnosis begins with its definition. BMJ. 2015;350:h869.
66. Marcus PM, Prorok PC, Miller AB, et al. Conceptualizing overdiagnosis in cancer screening. J Natl Cancer Inst. 2015;107(4).
67. Falk RS, Hofvind S. Overdiagnosis in mammographic screening because of competing risk of death. Cancer Epidemiol Biomarkers Prev. 2016;25(5):759-65.
68. Feinleib M, Zelen M. Some pitfalls in the evaluation of screening programs. Arch Environ Health. 1969;19(3):412-5.
69. Marmot MG, Altman DG, Cameron DA, et al. The benefits and harms of breast cancer screening: an independent review. Br J Cancer. 2013;108(11):2205-40.
70. Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. BMJ. 2015;350:g7773.

71. Biesheuvel C, Barratt A, Howard K, et al. Effects of study methods and biases on estimates of invasive breast cancer overdetection with mammography screening: a systematic review. Lancet Oncol. 2007;8(12):1129-38.
72. Puliti D, Duffy SW, Miccinesi G, et al. Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. J Med Screen. 2012;19 Suppl 1:42-56.
73. 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. Eur J Cancer. 2018;89:102-12.
74. Zahl PH, Maehlen J. Overdiagnosis of breast cancer after 14 years of mammography screening. Tidsskr Nor Laegeforen. 2012;132(4):414-7.
75. Kalager M, Adami H, 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):491-9.
76. Lund E, Mode N, Waaseth M, Thalabard JC. Overdiagnosis of breast cancer in the Norwegian Breast Cancer Screening Program estimated by the Norwegian Women and Cancer cohort study. BMC Cancer. 2013;13:614.
77. Falk RS, Hofvind S, Skaane P, Haldorsen T. Overdiagnosis among women attending a population-based mammography screening program. Int J Cancer. 2013;133(3):705-12.
78. van Luijt PA, Heijnsdijk EA, van Ravesteyn NT, et al. Breast cancer incidence trends in Norway and estimates of overdiagnosis. J Med Screen. 2017;24(2):83-91.
79. Lousdal ML, Møller MH, Kristiansen IS, et al. The Screening Illustrator: separating the effects of lead-time and overdiagnosis in mammography screening. Eur J Public Health. 2018;28(6):1138-42.
80. 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;23(4):192-202.
81. Fong Y, Evans J, Brook D, et al. The incidence and 10-year survival of interval breast cancers in Wales. Clin Radiol. 2014;69(4):e168-72.
82. Brekelmans CT, van Gorp JM, Peeters PH, Collette HJ. Histopathology and growth rate of interval breast carcinoma: characterization of different subgroups. Cancer. 1996;78(6):1220-8.
83. Vitak B, Olsen KE, Manson JC, et al. Tumour characteristics and survival in patients with invasive interval breast cancer classified according to mammographic findings at the latest screening: a comparison of true interval and missed interval cancers. Eur Radiol. 1999;9(3):460-9.
84. Porter GJ, Evans AJ, Burrell HC, et al. Interval breast cancers: prognostic features and survival by subtype and time since screening. J Med Screen. 2006;13(3):115-22.
85. Hofvind S, Geller B, Skaane P. Mammographic features and histopathological findings of interval breast cancers. Acta Radiol. 2008;49(9):975-81.
86. Weber RJ, van Bommel RM, Louwman MW, et al. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. Breast Cancer Res Treat. 2016;158(3):471-83.
87. Sala M, Domingo L, Louro J, et al. Survival and disease-free survival by breast density and phenotype in interval breast cancers. Cancer Epidemiol Biomarkers Prev. 2018;27(8):908-16.
88. den Bakker MA, Damhuis RAM. Pentameric last-digit preference and stage border avoidance in pathology measurement. Histopathology. 2018;73(3):510-3.
89. Coburn NG, Clarke-Pearson E, Chung MA, et al. A novel approach to T classification in tumor-node-metastasis staging of breast cancer. Am J Surg. 2006;192(4):434-8.
90. Tabár L, Tucker L, Davenport RR, et al. The use of mammographic tumour feature significantly improves outcome prediction of breast cancers smaller than 15 mm: a reproducibility study from two comprehensive breast centres. memo - Magazine of European Medical Oncology. 2011;4(3):149-57.
91. Candidates 913, 934, 919. Kvinners vurdering av overdiagnose og informasjon om overdiagnose i mammografiscreening [Women's assessments of overdiagnosis and information about overdianosis in mammographic screening] [Bachelor's thesis]. Oslo: Oslo Metropolitan University; 2018.
92. Hofvind S, Mangerud G, Ertzaas AK, et al. Prosjektrapport: Revisjon av informasjonsmateriellet i Mammografiprogrammet [Project report: revision of information material in BreastScreen Norway]. Oslo: Cancer Registry of Norway; 2019. Norwegian.

93. Hersch JK, Nickel BL, Ghanouni A, et al. Improving communication about cancer screening: moving towards informed decision making. Public Health Res Pract. 2017;27(2).
94. Larsen M, Moshina N, Sagstad S, Hofvind S. Factors associated with attendance and attendance patterns in a population-based mammographic screening program. J Med Screen. June 17, 2020. doi: 10.1177/0969141320932945. Online ahead of print.
95. Muratov S, Canelo-Aybar C, Tarride JE, et al. Monitoring and evaluation of breast cancer screening programmes: selecting candidate performance indicators. BMC Cancer. 2020;20(1):795.
96. European Commission Initiative on Breast Cancer. Inviting and informing women about organised screening programmes: European Commission; 2020 [updated May 28, 2020; cited July 7, 2020]. Available from: https://healthcare-quality.jrc.ec.europa.eu/european-breast-cancer-guidelines/Invitation-to-screening-and-decision-aid#recs-group-2.
97. Ghanouni A, Renzi C, Meisel SF, Waller J. Common methods of measuring 'informed choice' in screening participation: challenges and future directions. Prev Med Rep. 2016;4:601-7.
98. Østerlie W, Solbjør M, Skolbekken JA, et al. Challenges of informed choice in organised screening. J Med Ethics. 2008;34(9):e5.
99. Ritchie D, Van Hal G, Van den Broucke S. How is informed decision-making about breast cancer screening addressed in Europe? An international survey of 28 countries. Health Policy. 2020.
100. Ballesteros-Peña S, Gavilán-Moral E. Content of official written information for women on breast cancer screening in Spain. Rev Esp Salud Pública. 2018;92:e201810076.
101. Castagno R, Canuti D, Petrella M, et al. Information provided by Italian breast cancer screening programmes: a comparison between 2001 and 2014. Epidemiol Prev. 2015;39(3):48-51.
102. Gummersbach E, Piccoliori G, Zerbe CO, et al. Are women getting relevant information about mammography screening for an informed consent: a critical appraisal of information brochures used for screening invitation in Germany, Italy, Spain and France. Eur J Public Health. 2010;20(4):409-14.
103. Saalasti-Koskinen U, Makela M, Saarenmaa I, Autti-Ramo I. Personal invitations for population-based breast cancer screening. Acad Radiol. 2009;16(5):546-50.
104. Jorgensen KJ, Gotzsche PC. Content of invitations for publicly funded screening mammography. BMJ. 2006;332(7540):538-41.
105. Hofmann B. Informing about mammographic screening: ethical challenges and suggested solutions. Bioethics. 2020;34(5):483-92.
106. Hersch J, Jansen J, Barratt A, et al. Women's views on overdiagnosis in breast cancer screening: a qualitative study. BMJ. 2013;346:f158.
107. Waller J, Douglas E, Whitaker KL, Wardle J. Women's responses to information about overdiagnosis in the UK breast cancer screening programme: a qualitative study. BMJ Open. 2013;3(4):e002703.
108. Hersch J, McGeechan K, Barratt A, et al. How information about overdetection changes breast cancer screening decisions: a mediation analysis within a randomised controlled trial. BMJ Open. 2017;7(10):e016246.
109. Cancer Registry of Norway. Informasjonsmateriell [Information material in BreastScreen Norway]: Cancer Registry of Norway; 2020 [updated 21 December, 2020; cited 29 December, 2020]. Available from: https://www.kreftregisteret.no/screening/mammografiprogrammet/informasjonsmateriell
110. Waller J, Whitaker KL, Winstanley K, et al. A survey study of women's responses to information about overdiagnosis in breast cancer screening in Britain. Br J Cancer. 2014;111(9):1831-5.
111. Larsen IK, Smastuen M, Johannesen TB, et al. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. Eur J Cancer. 2009;45(7):1218-31.
112. Cancer Registry of Norway. Årsrapport 2013-2014. [National quality registry for breast cancer annual report, 2013-2014.] Oslo: Cancer Registry of Norway; 2015. Norwegian.
113. Cancer Registry of Norway. Årsrapport 2015. [National quality registry for breast cancer annual report, 2015.] Oslo: Cancer Registry of Norway; 2016. Norwegian.
114. Cancer Registry of Norway. Årsrapport 2016 med resultater og forbedringstiltak. [National quality registry for breast cancer annual report, 2016.] Oslo: Cancer Registry of Norway; 2017. Norwegian.

115. Cancer Registry of Norway. Årsrapport 2017 med resultater og forbedringstiltak fra Nasjonalt kvalitetsregister for brystkreft. [National quality registry for breast cancer annual report, 2017.] Oslo: Cancer Registry of Norway; 2018. Norwegian.
116. Hovda T, Tsuruda K, Hoff SR, et al. Radiological review of prior screening mammograms of screen-detected breast cancer. Eur Radiol. 1 October 2020. doi: 10.1007/s00330-020-07130-y. Online ahead of print.
117. University of Oslo. Nettskjema: University of Oslo; 2020 [cited December 3, 2020]. Available from: https://www.uio.no/english/services/it/adm-services/nettskjema/.
118. University of Oslo. [Legger Nettskjema igjen skjulte elektroniske spor som kan identifisere personer?] [Internet]. University of Oslo; 2019 [updated February 7, 2020; cited September 23, 2020]. Available from: https://www.uio.no/tjenester/it/applikasjoner/nettskjema/hjelp/tips-triks/elektroniske-spor.html.
119. Hersch J, Barratt A, Jansen J, et al. Use of a decision aid including information on overdetection to support informed choice about breast cancer screening: a randomised controlled trial. Lancet. 2015;385(9978):1642-52.
120. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet. 2002;359(9303):341-5.
121. Statistics Norway. Educational attainment of the population. Table 1: Age groups and educational attainment [updated February 27, 2020; cited March 24, 2020]. Available from: https://www.ssb.no/utniv/.
122. Statistics Norway. Population. 07459: Population, by sex and one-year age groups (M) 1986 - 2020 [updated February 27, 2020; cited March 24, 2020]. Available from: https://www.ssb.no/statbank/table/07459/.
123. Statistics Norway. Population. Table 2: Population by sex, age and marital status (per January 1) [updated February 27, 2020; cited March 24, 2020]. Available from: https://www.ssb.no/befolkning/statistikker/folkemengde/aar-per-1-januar.
124. Statistics Norway. Immigrants and Norwegian-born to immigrant parents. 05182: Persons, by immigration category and sex (C) 1970 - 2020 [updated March 9, 2020; cited March 24, 2020]. Available from: https://www.ssb.no/en/statbank/table/05182/.
125. van Buuren S. Flexible Imputation of Missing Data. Boca Raton, FL: Chapman and Hall/CRC; 2012.
126. Bjørnstad J, Lydersen S. Missing data. In: Veierød MB, Lydersen S, Laake P, editors. Medical statistics in clinical and epidemiological research. 1 ed. Oslo, Norway: Gyldendal; 2012. p. 429-61.
127. Lohr S. Nonresponse. Sampling: design and analysis. 2nd ed. Boston, MA: Brooks/Cole, Cengage Learning; 2009. p. 329-57.
128. Kolenikov S. Calibrating survey data using iterative proportional fitting (raking). The Stata Journal. 2014;14(1):22-59.
129. Lydersen S, Fagerland MW, Laake P. Categorical data and contingency tables. In: Veierød MB, Lydersen S, Laake P, editors. Medical statistics in clinical and epidemiological research. 1st ed. Oslo: Gylendal Norsk Forlag; 2012. p. 48-89.
130. Aalen O, Riksheim M. Survival analysis. In: Veierod MB, Lydersen S, Laake P, editors. Medical statistics in clinical and epidemiological research. 1 ed. Oslo, Norway: Gyldendal; 2012. p. 167-98.
131. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. 2009;28(15):1982-98.
132. Bellera CA, MacGrogan G, Debled M, et al. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC Med Res Methodol. 2010;10:20.
133. Lohr S. Categorical data anaylsis in complex surveys. Sampling: design and analysis. 2nd ed. Boston, MA: Brooks/Cole, Cengage Learning; 2009. p. 401-22.
134. Kvalitetsmanual i Mammografiprogrammet - Patologi [Quality manual for BreastScreen Norway - Pathology]. Oslo: BreastScreen Norway, Cancer Registry of Norway; 2018.
135. Verschraegen C, Vinh-Hung V, Cserni G, et al. Modeling the effect of tumor size in early breast cancer. Ann Surg. 2005;241(2):309-18.
136. Hersch J, Barratt A, Jansen J, et al. The effect of information about overdetection of breast cancer on women's decision-making about mammography screening: study protocol for a randomised controlled trial. BMJ Open. 2014;4(5).

137. Folkeopplysningen 6 - Kroppen på service [Television broadcast]. Norway: NRK1; First broadcast: September 7, 2020. Running time: 40:00. Norwegian. Available from: https://tv.nrk.no/serie/folkeopplysningen/2020/KMTE50003120/avspiller.
138. 31. juli 2019 – Dagsnytt 18 [Television broadcast]. Norway: NRK2; First broadcast: July 31, 2019. Running time: 59:00. Norwegian. Available from: https://tv.nrk.no/serie/dagsnytt-atten-tv/201907/NNFA56073119/avspiller (starting at 34:59).
139. 31. juli 2019 – Dagsnytt 18 [Radio broadcast]. Norway: NRK Nyheter; First broadcast: July 31, 2019. Running time: 59:00. Norwegian. Available from: https://radio.nrk.no/serie/dagsnytt-atten/sesong/201907/NMAG03015119 (starting at 34:59).
140. 13. feb. 2014 – Debatten: helsefrykt [Television broadcast]. Norway: NRK1; First broadcast: February 13, 2014. Running time: 59:00. Norwegian. Available from: https://tv.nrk.no/serie/debatten/201402/NNFA51021314/avspiller.
141. Chow S, Shao J, Wang H. Comparing time-to-event data. Sample size calculations in clinical research. Chapman & Hall/CRC Biostatistics Series. 2nd ed: Chapman and Hall/CRC; 2008. p. 163-86.
142. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338.
143. Carter CL, Allen C, Henson DE. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer. 1989;63(1):181-7.
144. Pérez-Lacasta MJ, Martínez-Alonso M, Garcia M, et al. Effect of information about the benefits and harms of mammography on women's decision making: the InforMa randomised controlled trial. PLOS ONE. 2019;14(3):e0214057.
145. Statistics Norway. ICT usage in households; 11437: Use of social media (per cent), by sex, age, contents and year [updated August 29, 2019; cited December 10, 2020]. Available from: https://www.ssb.no/en/statbank/table/11437.
146. Eurostat. Individuals who used the internet for interaction with public authorities (ISOC_R_GOV_I) 2020 [cited December 3, 2020]. Available from: https://ec.europa.eu/eurostat/databrowser/view/isoc_r_gov_i/default/table?lang=en.
147. Elmore JG, Lee CI. What do the European Breast Cancer Screening Guidelines portend for U.S. practice? Ann Intern Med. 2020;172(1):65-6.
148. Hayes SJ. Does terminal digit preference occur in pathology? J Clin Pathol. 2008;61(8):975.
149. Hayes SJ. Terminal digit preference occurs in pathology reporting irrespective of patient management implication. J Clin Pathol. 2008;61(9):1071-2.
150. Anderson KR, Heidinger BH, Chen Y, et al. Measurement bias of gross pathologic compared with radiologic tumor size of resected lung adenocarcinomas. Am J Clin Pathol. 2017;147(6):641-8.
151. Veierød MB, Page CM, Aaserud S, et al. Melanoma staging: Varying precision and terminal digit clustering in Breslow thickness data is evident in a population-based study. J Am Acad Dermatol. 2018;79(1):118-25 e1.
152. Heusinger K, Lohberg C, Lux MP, et al. Assessment of breast cancer tumor size depends on method, histopathology and tumor size itself. Breast Cancer Res Treat. 2005;94(1):17-23.
153. Stein RG, Wollschlager D, Kreienberg R, et al. The impact of breast cancer biological subtyping on tumor size assessment by ultrasound and mammography - a retrospective multicenter cohort study of 6543 primary breast cancer patients. BMC Cancer. 2016;16:459.
154. Gruber IV, Rueckert M, Kagan KO, et al. Measurement of tumour size with mammography, sonography and magnetic resonance imaging as compared to histological tumour size in primary breast cancer. BMC Cancer. 2013;13:328.
155. Harrell FE, Jr. Regression Modeling Strategies. 2 ed: Springer International Publishing; 2015.
156. Marinovich ML, Macaskill P, Irwig L, et al. Agreement between MRI and pathologic breast tumor size after neoadjuvant chemotherapy, and comparison with alternative tests: individual patient data meta-analysis. BMC Cancer. 2015;15:662.
157. Marinovich ML, Macaskill P, Bernardi D, Houssami N. Systematic review of agreement between tomosynthesis and pathologic tumor size for newly diagnosed breast cancer and comparison with other imaging tests. Expert Rev Med Devices. 2018;15(7):489-96.
158. Sechopoulos I, Mann RM. Stand-alone artificial intelligence - The future of breast cancer screening? Breast. 2020;49:254-60.

Appendix

Online questionnaire (Norwegian language)

Table 12: *Original survey questions and potential responses; where applicable, correct answers are marked in italics.*[a]

Survey question	**Response format**
Har du noen gang oppsøkt informasjon om mammografiscreening?	- Ja, mye - Ja, noe - Nei - Husker ikke/usikker
Hvor har du søkt etter informasjon om mammografiscreening? (Flere kryss er mulig)	- Fra fastlege/annet helsepersonell - Venner og familie - Kreftregisterets/Mammografiprogrammets nettsider - Andre nettsider - Fagartikler - Bøker - Aviser og ukeblader - Annet
Velg den setningen du mener er riktig.	- *Mammografiscreening er å ta mammografi når du ikke har merket kul eller andre symptomer på brystkreft* - Mammografiscreening er å ta mammografi når du har oppdaget en kul eller har andre symptomer fra brystet - Vet ikke
Har du noen gang hørt om disse tre begrepene?[b] - Falske positive mammografiundersøkelser - Overdiagnostikk - Overdeteksjon	- Ja - Nei - Vet ikke
Velg den setningen du mener forklarer begrepet «falsk positiv screeningundersøkelse» best.	- *Unormale funn på screeningmammogrammene, men hvor tilleggsundersøkelser ikke viste brystkreft* - Brystkreft som aldri ville blitt oppdaget dersom man ikke hadde deltatt i screening Vet ikke
Hvem tror du har høyest sannsynlighet for å dø av brystkreft?	- Kvinner som deltar i screeningprogram for brystkreft/mammografiscreening - *Kvinner som IKKE deltar i screeningprogram for brystkreft/mammografiscreening* - Vet ikke
Vil screening med mammografi finne all brystkreft?	- Ja - *Nei* - Vet ikke
Vil alle kvinner som har unormale funn på screeningmammogrammet få påvist brystkreft?	- Ja - *Nei* - Vet ikke

Survey question	Response format
Hvem tror du har høyest sannsynlighet for å få påvist brystkreft?	- *Kvinner som deltar i screeningprogram for brystkreft/mammografiscreening* - Kvinner som IKKE deltar i screeningprogram for brystkreft/mammografiscreening - Vet ikke
Sett kryss ved de setningene du mener er sanne: (Flere kryss er mulig)	- [Correct if not crossed off] All brystkreft vil til slutt føre til sykdom og død hvis den ikke blir påvist og behandlet[c] - [Correct if not crossed off] Med stor sikkerhet kan leger skille farlige brystkreftsvulster som må behandles fra «snille» svulster som ikke trenger behandling[c] - *[Correct if crossed off] Det finnes saktevoksende brystkreft som blir behandlet selv om de ikke ville gitt helseproblemer*[c] - *[Correct if crossed off] Mammografiscreening fører til diagnose av saktevoksende svulster og unødvendig behandling*[c]
Velg den setningen du mener forklarer begrepet «overdiagnostikk» best.	- Unormale funn på screeningmammogrammene, men hvor tilleggsundersøkelser ikke viste brystkreft - *Brystkreft som aldri ville blitt oppdaget dersom man ikke hadde deltatt i screening* - Vet ikke

[a] *Questions and responses were adapted from Hersch J, Barratt A, Jansen J, et al. Use of a decision aid including information on overdetection to support informed choice about breast cancer screening: a randomised controlled trial. The Lancet. 2015;385(9978):1642-52.*
[b] *These terms were shown together in a grid, and participants could select one response for each term.*
[c] *These sentences were shown together in a grid and participants had the option to cross off any item.*

Paper 1

Tsuruda KM, Hofvind S, Akslen LA, Hoff SR, Veierød MB. Terminal digit preference: a source of measurement error in breast cancer diameter reporting. Acta Oncol. 2020;59(3):260-7.

I

ACTA ONCOLOGICA
2020, VOL. 59, NO. 3, 260–267
https://doi.org/10.1080/0284186X.2019.1669817

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Terminal digit preference: a source of measurement error in breast cancer diameter reporting

Kaitlyn M. Tsuruda[a,b], Solveig Hofvind[a,c], Lars A. Akslen[d,e], Solveig R. Hoff[f,g] and Marit B. Veierød[h]

[a]Cancer Registry of Norway, Oslo, Norway; [b]Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway; [c]Faculty of Health Sciences, Department of Life Sciences and Health, Oslo Metropolitan University, Oslo, Norway; [d]Centre for Cancer Biomarkers CCBIO, Department of Clinical Medicine, University of Bergen, Bergen, Norway; [e]Department of Pathology, Haukeland University Hospital, Bergen, Norway; [f]Department of Radiology, Aalesund Hospital, Aalesund, Norway; [g]Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway; [h]Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway

ABSTRACT

Objectives: Women diagnosed with breast cancer are offered treatment and therapy based on tumor characteristics, including tumor diameter. There is scarce knowledge whether tumor diameter is accurately reported, or whether it is unconsciously rounded to the nearest half-centimeter (terminal digit preference). This study aimed to assess the precision (number of digits) of breast cancer tumor diameters and whether they are affected by terminal digit preference. Furthermore, we aimed to assess the agreement between mammographic and histopathologic tumor diameter measurements.

Material and Methods: This national registry study included reported mammographic and registered histopathologic tumor diameter information from the Cancer Registry of Norway for invasive breast cancers diagnosed during 2012–2016. Terminal digit preference was assessed using histograms. Agreement between mammographic and histopathologic measurements was assessed using the intraclass correlation coefficient (ICC) and Bland-Altman plots.

Results: Mammographic, histopathologic, or both tumor measurements were available for 7792, 13,541 and 6865 cases, respectively. All mammographic and 97.2% of histopathologic tumor diameters were recorded using whole mm. Terminal digits of zero or five were observed among 38.7% and 34.8% of mammographic and histopathologic measurements, respectively. There was moderate agreement between the two measurement methods (ICC = 0.52, 95% CI: 0.50–0.53). On average, mammographic measurements were 1.26 mm larger (95% limits of agreement: −22.29–24.73) than histopathologic measurements. This difference increased with increasing tumor size.

Conclusion: Terminal digit preference was evident among breast cancer tumor diameters in this nationwide study. Further studies are needed to investigate the potential extent of under-staging and under-treatment resulting from this measurement error.

ARTICLE HISTORY
Received 19 June 2019
Accepted 14 September 2019

Introduction

Preferential overrepresentation of certain terminal digits (terminal digit preference) is a well-known source of measurement error in medicine, including the measurement of malignant tumors [1–3]. Some evidence suggests that pathologists measuring breast tumor diameter favor terminal digits of zero or five, though few studies have evaluated this type of measurement error as a primary objective [3–6]. Moreover, terminal digit preference in breast radiology is not well described.

Pretreatment clinical tumor size (cT) informs neoadjuvant and surgical treatment decisions, while pathological tumor size (pT) informs adjuvant therapy and follow-up. The tumor-node-metastasis (TNM) framework defines breast tumor size categories (T categories) T1–T3 based on maximum diameter, where T1: ≤20 mm, T2: >20–50 mm, and T3: >50 mm [7,8]. T4 refers to any tumor with direct extension to the chest wall and/or skin, and TX refers to cases where the tumor diameter has not or cannot be assessed [7–9].

In Norway, the clinician responsible for securing a patient's diagnosis reports the cT category from either palpation (calipers), ultrasound, mammography, or magnetic resonance imaging. In the absence of information from palpation, radiologic imaging, including mammography, provides influential information when determining cT, but cT can under- or overestimate pT [10,11]. Accurate estimates of cT and pT are important for optimal clinical decision-making and personalized treatment.

CONTACT Marit B. Veierød m.b.veierod@medisin.uio.no Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, P.O. Box 1122 Blindern, 0317 Oslo, Norway

Supplemental data for this article can be accessed here.

This article has been republished with minor changes. These changes do not impact the academic content of the article.

The World Health Organization (WHO) and American Joint Committee on Cancer have advised for many years that breast cancer tumor measurements be reported to the nearest mm – recent editions provide detailed guidance on rounding [7–9]. Norwegian guidelines are based on the WHO guidelines and also adhere to those from the national breast cancer screening program, BreastScreen Norway [12,13].

In Norway, mammographic and histopathological tumor diameters are routinely reported and registered at the Cancer Registry of Norway (CRN). We wanted to determine whether terminal digit preference was present among these population-based data, and describe the precision to which pathologic tumor diameter was reported. As a secondary objective, we aimed to assess the agreement between mammographic and histopathologic tumor diameter measurements.

Methods

Data source

All data were extracted from the CRN. Mandatory reporting of cancer cases from multiple sources ensures complete and high quality data [14,15]. Roughly 30% of all breast cancers nationwide are screen-detected through BreastScreen Norway, which is administered by the CRN and offers biennial mammographic screening to all female residents aged 50–69 [15,16].

Information about mammographic and histopathologic tumor diameter is sent to the CRN via standardized electronic radiology reports from radiologists, and via electronic or paper-based histopathology reports from pathologists. Radiology reports are registered automatically at the CRN, while histopathology reports are registered manually.

Study sample

We included information about mammographic and histopathological tumor diameter from all incident invasive breast cancers diagnosed during 2012–2016 among women residing in Norway.

The first invasive breast cancer diagnosed per woman during the study period was included. Only the largest lesion was included for women diagnosed with multifocal or bilateral breast cancer. We excluded women with pT4 lesions, cases where the longest measured mammographic or histopathologic tumor diameter was $\geq$100 mm, and cases with no recorded tumor diameter (e.g., pTX).

This project was approved by Oslo University Hospital's privacy ombudsman (PVO 19/02585) [17].

Study variables

We extracted information about women's age and year of diagnosis, and mode of detection, which was classified as screen-detected (diagnosed within six months after a positive screening mammogram in BreastScreen Norway), or clinically detected (not screen-detected). Histologic tumor classification (invasive carcinoma of no special type (NST), lobular, or other), the largest mammographic and histopathologic tumor diameter measured (mm; hereafter referred to as mammographic and histopathologic tumor diameter, respectively), and cT and pT classifications were also extracted.

Mammographic tumor diameter

Mammographic tumor diameter was measured on digitally acquired images using electronic calipers in a clinical setting by the examining radiologist. Measurements could be taken from screening or diagnostic mammographic examinations, from either standard 2D images (with or without contrast), spot compression views (with or without magnification), tomosynthesis images, or synthetic 2D images derived from tomosynthesis. Images were displayed on DICOM-compliant workstations after retrieval from a picture archiving and communication system (PACS). The type of mammography equipment, viewing workstations, and PACS varied between centers.

There are no national guidelines describing mammographic tumor measurement methods. For mammographically visible cases, radiologists typically reported the tumor diameter in whole mm in the patient's radiology report on the day of the examination. Thereafter, this measurement was entered into the report sent to the CRN.

Histopathologic tumor diameter

Based on national recommendations, tumor diameter was determined from microscopic slides by measuring the outermost boundaries of the invasive lesion, and measured to the nearest mm using a transparent ruler [12]. If this could not be done using the available slides, the measurement was based on the macroscopic examination (formalin fixed specimen), either from a single tissue slice, or as an estimate across all tissue slices containing microscopically verified invasive tumor tissue [7].

Measurement precision

All mammographic measurements must be sent to the CRN as whole mm.

Histopathologic tumor diameter could be registered with one decimal place if it was reported as such (e.g., for small invasive lesions), but national recommendations advise pathologists to report tumor diameter using whole mm [12,13].

T-category information

The clinician responsible for securing a patient's diagnosis, often the patient's surgeon, reported the cT category directly to the CRN. Due to a high proportion of missing and unavailable data [18], we derived a surrogate assessment, mT, using the mammographic tumor diameter. The mT variable was used in all analyses requiring information about cT.

The pT category was manually set by dedicated coders at the CRN based on the maximum tumor diameter and tumor

invasion information included in the submitted pathology reports.

Statistical analysis

Descriptive results were presented as means (standard deviations, SDs, or ranges), or frequencies (%). Histograms were used to present the frequency distributions of tumor diameter measurements, stratified by measurement type (mammographic vs histopathologic), and the number of decimal places (zero or one). To focus on the cut-points for T1–T3 tumors, some figures were restricted to cases with tumors $\leq$55 mm. Using the same stratification variables, we reported the proportion of different terminal digits among tumors of varying sizes and used histograms to visualize overall terminal digit frequencies.

To study agreement between mammographic and histopathologic tumor diameter for cases where both measurements were available, we estimated means and 95% confidence intervals (CIs); Pearson's correlation coefficient (r) and an asymptotic 95% CI based on Fisher's z transformation; and the intraclass correlation coefficient (*ICC*) and 95% CI based on a single measurement, one-way random effects model for absolute agreement [19]. A scatterplot and Bland-Altman plot were used to visualize the differences between mammographic and histopathologic measurements and to calculate the absolute and relative mean differences and 95% limits of agreement (LOA) using raw and natural log transformed data. We presented the relative mean difference and 95% LOA because the magnitude of the differences appeared to be associated with tumor diameter [20]. Lastly, the mT and pT categories were compared using proportions of agreement/disagreement, and a weighted Kappa with Cicchetti-Allison weights (κ_w). Bowker's test was used to test symmetry.

An ad hoc analysis of cases with a terminal digit of zero or five from mammography or histopathology was conducted using a scatterplot, and Bland-Altman plots for raw and natural log transformed data.

All statistical analyses were performed in R version 3.5.0 for Windows [21]. The *irr* package was used to calculate the *ICC* and corresponding CI, while the *DescTools* package was used to calculate κ_w and its corresponding CI [22,23].

Results

During 2012–2016, 16,767 women were diagnosed with invasive breast cancer in Norway. After applying exclusion criteria, the final study sample consisted of 14,468 invasive breast tumors with mammographic or histopathologic tumor diameter information in the same number of women (Figure 1(A)). Tumor diameter information from mammography, histopathology, or both sources was available for 7792; 13,541; and 6865 cases, respectively (Figure 1(B)).

The mean age at diagnosis in the final study sample was 62 years (range 23–103), and 57% of women were aged 50–69 (Table 1). Overall, roughly 35% of cancers were screen-detected, and 82% of all cancers were invasive carcinoma of no special type (Table 1). The subsample of women with mammographic tumor information was slightly older than that with histopathologic information (means (SDs) 62.2 (11.7), and 61.7 (12.6) years, respectively).

Precision

All mammographic tumor diameters were recorded as whole numbers (median 18 mm, range 1–99). Nearly all (97.2%, $n = 13{,}167$) of the histopathologic diameters were recorded as whole numbers (median 17 mm, range 1–95); the remaining cases (2.8%) were recorded with a single decimal (median 7.4 mm, range 0.1–50.1).

Terminal digit preference

Histograms of the distribution of mammographic and histopathologic tumor diameters recorded as whole numbers showed high frequencies of multiples of five and drops around these peaks that corresponded to numbers ending in

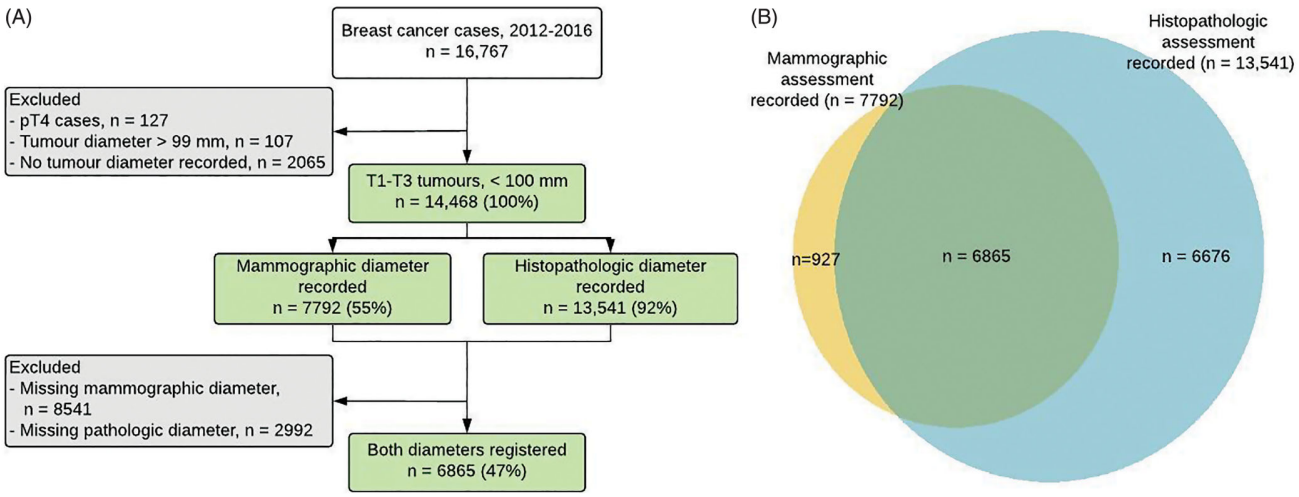


Figure 1. (A) Flow diagram (B) Venn diagram indicating whether tumor diameter information on mammographic (yellow area), histopathologic (blue area), or both (green area) was available for women diagnosed with invasive T1–T3 breast cancer in Norway during 2012–2016, $n = 14{,}468$.

Table 1. Characteristics of women diagnosed with invasive T1–T3 breast cancer in Norway during 2012–2016, $n = 14{,}468$, stratified by whether mammographic, histopathologic , or both assessments were available.

Characteristic	Total, *n* (%) $n = 14{,}468$	Mammographic assessments, *n* (%) $n = 7792$	Histopathologic assessments, *n* (%) $n = 13{,}541$	Both assessments, *n* (%) $n = 6865$
Age at diagnosis				
20–49	2503 (17.3)	893 (11.5)	2249 (16.6)	639 (9.3)
50–69	8251 (57.0)	5313 (68.2)	7985 (59.0)	5047 (73.5)
70+	3714 (25.7)	1586 (20.4)	3307 (24.4)	1179 (17.2)
Year of diagnosis				
2012	2482 (17.2)	849 (10.9)	2434 (18.0)	801 (11.7)
2013	2810 (19.4)	1222 (15.7)	2715 (20.1)	1127 (16.4)
2014	2993 (20.7)	1567 (20.1)	2806 (20.7)	1380 (20.1)
2015	3072 (21.2)	1942 (24.9)	2831 (20.9)	1701 (24.8)
2016	3111 (21.5)	2212 (28.4)	2755 (20.3)	1856 (27.0)
Method of detection				
Screen-detected	5078 (35.1)	4357 (55.9)	4987 (36.8)	4266 (62.1)
Clinically detected	9390 (64.9)	3435 (44.4)	8554 (63.2)	2626 (37.9)
Histological type				
Invasive NST[a]	11883 (82.1)	6579 (84.4)	11122 (82.1)	5818 (84.8)
Lobular	1571 (10.9)	765 (9.8)	1467 (10.9)	661 (9.6)
Other	1014 (7.0)	448 (5.8)	952 (7.0)	386 (5.6)

[a]NST: no special type.

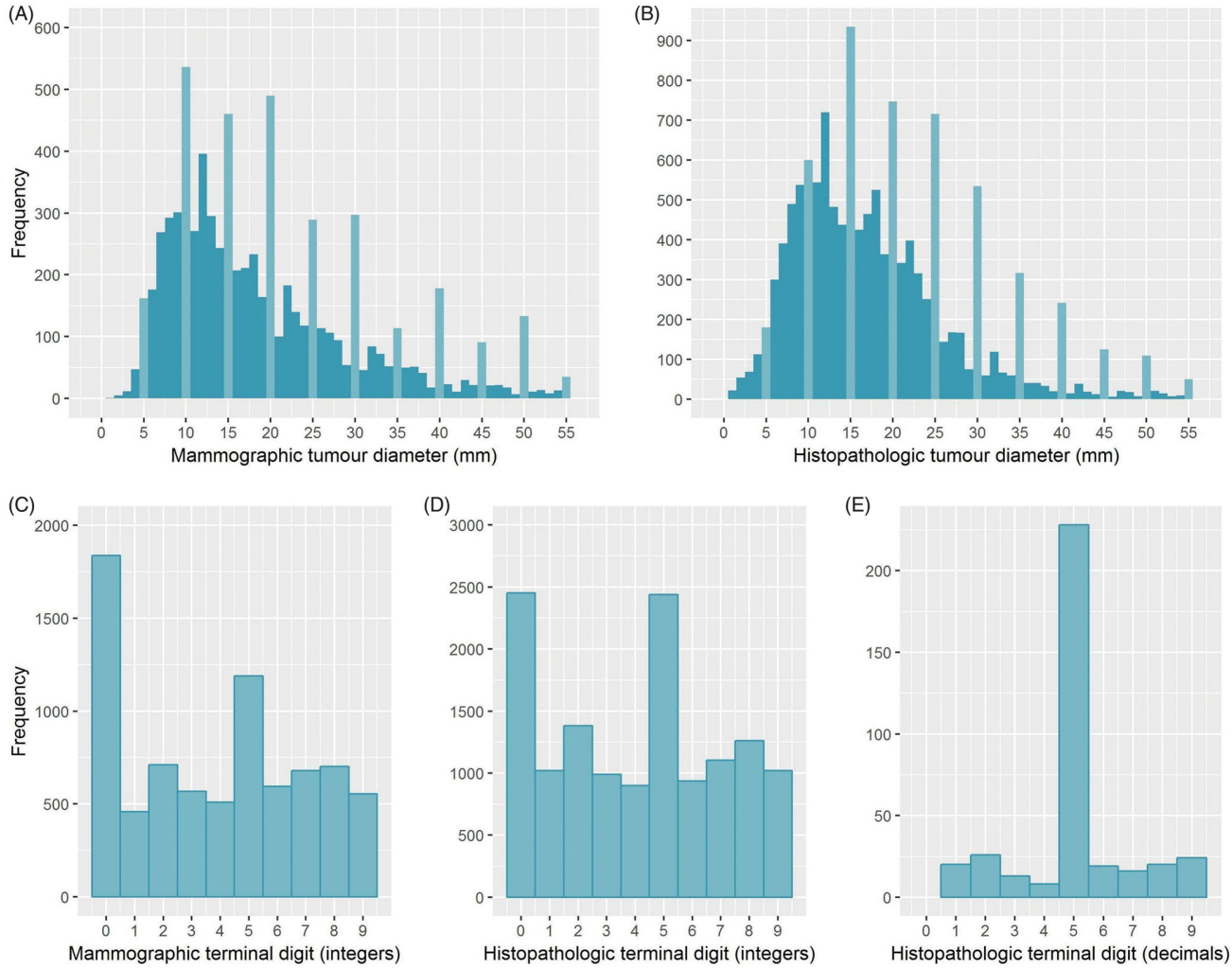


Figure 2. Longest measured tumor diameter recorded for T1–T3 breast cancers (≤55 mm) diagnosed during 2012–2016, based on (A) mammography ($n = 7792$) and (B) histopathology (whole numbers only; $n = 13{,}167$); light blue bars emphasize tumor diameters with zero or five as a terminal digit. The lower panels indicate the frequency of terminal digits from the longest measured tumor diameter (≤100 mm) for T1–T3 breast cancers diagnosed during 2012–2016, based on (C) mammography (whole numbers only; $n = 7792$), (D) histopathology (whole numbers only; $n = 13{,}167$), and (E) histopathology (decimals only; $n = 374$).

one, four, six, or nine (Figure 2(A,B)). A histogram of the histopathologic tumor diameters recorded with one decimal place also showed a high frequency had a terminal digit of five (Supplementary Figure S1).

Histograms of the terminal digits from mammographic and histopathologic tumor diameters recorded as whole numbers also showed a high frequency of zeroes and fives, compared to ones, fours, sixes, and nines (Figure 2(C,D)). A terminal digit of zero was more frequent than five among tumor diameters measured from mammography (23.4% and 15.3%, respectively), but not from histopathology (17.9% for both; Supplementary Table S1). The proportion of tumor diameters with terminal digit of five generally increased as mammographic or histopathologic tumor diameter increased. This was also observed for a terminal digit of zero, but only for tumors ≥20 mm (Supplementary Table S1). Among the histopathologic measurements recorded with one decimal, 61% of cases had a decimal value of five, and none had a decimal value of zero (Figure 2(E)).

Agreement between mammographic and histopathologic assessments

Among cases for which a mammographic and histopathologic tumor diameter were available ($n = 6865$), the median (range) tumor diameters were 16 mm (1–99) and 15 mm (0.1–90), respectively. We observed moderate correlation between the two assessments: $r = 0.52$ (95% CI: 0.51–0.54), and $ICC = 0.52$ (95% CI: 0.50–0.53). On average, mammographic tumor diameters were 1.26 mm (95% LOA: −22.29–24.73) larger than the corresponding histopathologic diameters. However, there was evidence of disagreement in both directions, and agreement between the two decreased as the average tumor diameter increased (Figure 3). On a relative scale, the mammographic tumor diameters were 1.06 times (6%) larger than the pathologic values (95% LOA: 0.34–3.28) (Supplementary Figure S2). A subgroup analysis stratified by histologic type showed that for invasive NST, lobular, or other carcinomas, the mammographic tumor diameters were, on average, 1.53 mm (95% LOA: −21.39–24.45) larger, 1.59 mm (95% LOA: −28.85–25.67) smaller, and 2.04 mm (95% LOA: −21.06–25.14) larger than the corresponding histopathologic measurements, respectively.

Ad hoc analysis of all cases with mammographic and histopathologic tumor diameter with a terminal digit of zero or five ($n = 3890$) displayed clear graphical patterns in the scatter plot (checkerboard pattern) and the Bland-Altman plots (lattice pattern; Supplementary Figure S3).

The mT and pT categories were the same in 5313 cases (77%), while the mT category was highest in 823 (53%) of the discordant cases ($p_{symmetry} < 0.0001$; Table 2). Overall, there was moderate agreement between the mT and pT categories: $\kappa_w = 0.50$, 95% CI: 0.48–0.53.

Discussion

Terminal digit preference in the measurement of breast tumors is not well-studied and only one of four previously published studies on this topic used population-based data [3–6]. Our results from nationwide cancer registry data showed evidence of terminal digit preference for zeroes and fives in the reporting of maximum mammographic and histopathologic tumor diameters of invasive breast cancers diagnosed during 2012–2016. This measurement error can lead to T-category misclassification and has the potential to

Table 2. Contingency table of mammographic vs histopathologic tumor categories (pT), $n = 6865$.

	Histopathologic tumor category (pT)		
	pT1	pT2	pT3
Mammographic tumor category			
mT1	3932 (57.3%)	615 (9.0%)	36 (0.5%)
mT2	657 (9.6%)	1335 (19.4%)	78 (1.1%)
mT3	87 (1.3%)	79 (1.1%)	46 (0.7%)

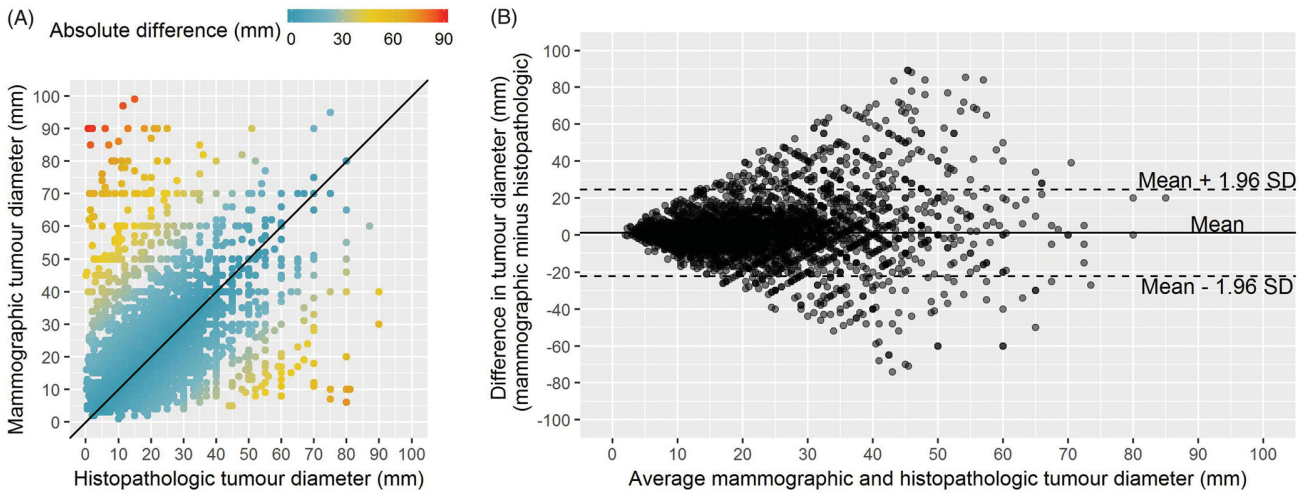


Figure 3. Maximum tumor diameters for T1–T3 breast cancers diagnosed in Norway during 2006–2012, $n = 6865$. (A) Scatterplot displaying the maximum histopathologic tumor diameter (mm; x-axis) versus the maximum mammographic tumor diameter (mm; y-axis), where the solid line indicates perfect agreement (B) Bland-Altman plot illustrating the difference in tumor size measurement (mm) between mammographic and histopathologic measurements, against the average of the two measurements (mm). The solid line represents the mean difference, while the top and bottom dashed lines represent the upper and lower limits of agreement, respectively.

impact patient treatment. Further, we observed moderate agreement between mammographic and pathologic tumor diameter. Mammographic tumor diameter both over- and underestimated histopathologic tumor diameter, and the absolute discrepancy between the two measurements increased with increasing tumor diameter.

Among histopathologic tumor diameters, 97.3% of cases were registered as whole mm, and the majority of those reported with decimal values were <10 mm. This is a slight overestimate of the number of cases reported as whole mm by pathologists because decimal values of zero are truncated when registered at the CRN. More explicit national guidance may improve compliance in reporting tumor diameter to the nearest mm, as advised by international guidelines [7,8].

The overrepresentation of tumor diameters with a terminal digit of zero or five corresponded largely to whole and half cm values ≥1 cm on mammography and ≥1.5 cm from histopathology. The resulting measurement error appeared to be due in equal parts to rounding up and down to the nearest whole or half cm value. A recent Dutch study suggested that pathologists avoided reporting T category border values of 10 mm and 20 mm, however, these results have not been corroborated [3–6]. In our study, terminal digit preference led to an excess of tumors recorded as 10, 20, or 50 mm, which define the border values for T1c, T2, and T3 tumors [7–9]. Any tendency to underestimate tumor diameter by rounding down to the nearest cm increases the likelihood of under-staging tumors with respect to their T category and could lead to under-treatment. This limitation of the TNM system was pointed out as early as 2006, but even recent suggestions to simplify the TNM system have overlooked this shortcoming [4,24].

On the other hand, rounding up to 10, 20 or 50 mm does not change the T category and therefore does not directly lead to over-staging (e.g., an 18 mm tumor rounded up to 20 mm is classified as T1 in either situation). A tumor that is under a boundary value (e.g., 18 mm; T1) and is rounded up to over the boundary value (e.g., 21 mm; T3) would result in over-staging in terms of the T category, but our study does not provide evidence about whether this type of rounding occurs.

Without knowing which specific cases are affected by terminal digit preference, it is difficult to speculate the scope of the clinical implications of this measurement error. Studies that report treatment data and re-measure histopathologic samples are needed to quantify the number of women who may potentially be undertreated for their breast cancer due to under-staging and could have important consequences for decisions regarding neoadjuvant treatment and surgical planning, as well as the use of chemo and radiation therapy and response monitoring.

On average, we observed that mammographic tumor diameters were slightly larger than the corresponding histopathologic measurements. However, the 95% LOA for this estimate was wide and indicated that both under- and overestimation can occur, as has been observed in other studies [10,11,25]. We observed decreasing absolute agreement between mammographic and histopathologic tumor diameter measurements as tumor size increases, which supports previous findings [10,25]. Despite the measurement differences in mammographic and histopathologic tumor diameter, the mT and pT categories assigned to a case were the same in over 75% of cases. In the remaining cases, particularly the <2% categorized as mT3 and pT1 or vice-versa, more accurate mammographic measurement may have increased the likelihood of breast conserving surgery, or reduced the need for reoperation.

When considering histologic subtype, our study indicated that mammography underestimates histopathologic tumor diameter for lobular carcinomas, but results from two single-centre studies indicate the opposite [10,26]. This discrepancy may be due to sampling variability, as these two studies evaluated 99 [10] and 18 [26] cases of lobular carcinoma and diffuse tumors can be difficult to measure. Our findings corroborate those from a multicentre cohort study that analyzed 474 cases of lobular carcinoma, but do not corroborate their finding that mammography underestimates histopathologic tumor diameter for invasive carcinomas NST [25]. Both our study and that of Stein et al. [25] used retrospective data and had similar inclusion criteria, but the difference between mammographic and histopathologic tumor diameter was roughly 1.5 mm larger across all subgroups in our study, which would have been enough to change the direction of the association observed for the sub-analysis of invasive carcinomas NST. Because the study samples were similar, we believe that differences in measurement practices between countries can explain this discrepancy.

Discrepancies between mammographic and histopathologic measurements can occur due to differential terminal digit preference between radiologists and pathologists, or because the measurements are taken at different times and potentially from different axes of the tumor. They may also be attributed to measurement challenges associated with endocrine responsiveness, mammographic breast density, diffuse or multifocal lesions, or shrinkage as a result of neoadjuvant treatment or formalin fixation [10,11,25,26]. Moreover, histopathologic measurements can be influenced by the degree of sampling, particularly from the tumor periphery and surrounding tissues. We included only invasive T1-T3 breast cancer cases and aimed to exclude those who received neoadjuvant treatment (pTX) to mitigate some of these challenges.

Our study design did not allow us to determine whether terminal digit preference is associated with factors that complicate accurate tumor measurement; this was a limitation of our study. We hypothesize that only factors that obscure the tumor periphery (e.g., multifocal or diffuse lesions) may be associated with increased terminal digit preference, and future studies might investigate this topic. With respect to our secondary outcome – agreement between mammographic and histopathologic tumor diameter measurements – determining whether factors such as breast density confounded the relationships we observed was outside the scope of this study. This was another limitation of our study.

To the best of our knowledge, studies measuring the agreement between mammographic and histopathologic

tumor diameter have not considered the potential for measurement error due to terminal digit preference. In our study, the effects of this preference can be seen in the checkerboard pattern arising in the scatterplot and lattice pattern arising in the Bland-Altman plots. These patterns are also visible in the plots of other agreement studies, which suggests that terminal digit preference is prevalent in the measurement of tumor diameter from mammography and histopathology, as well as from clinical examination, ultrasound and MRI [10,11,25,27]. Terminal digit preference is therefore an important source of measurement error in cT-category staging. This source of measurement error should be more widely discussed and potentially taken into account when making neoadjuvant and surgical treatment decisions, particularly for borderline cases.

The use of prospectively collated data from a population-based registry is a strength of our study. Reporting to the CRN is mandated by law, and clinicians working in oncology do so as a part of routine clinical practice. The data included in this study are continually used for research and surveillance of BreastScreen Norway, and are constantly assured by clinicians working in the program. Moreover, pathologic tumor diameter data at the CRN are subject to annual quality assurance against the original pathology report, as described in the appendix. This validation work has not been documented extensively, but studies using colorectal cancer data from the CRN and registry-based breast cancer data in Denmark and Sweden support the notion that pathological breast cancer tumor diameter data at the CRN reflects that from the original pathology report [28–30]. Furthermore, the data used in this study are not affected by clinicians' awareness of being studied (the Hawthorne effect) [31]. Our results therefore reflect national standard clinical practice.

Although our sample includes tumor measurements from a national pool of radiologists and pathologists, our study did not include information about the individual clinicians who performed the measurements, and it was not possible to investigate inter- or intra-observer trends. Moreover, we did not have information about the specific conditions in which the mammographic tumor diameter measurements took place, for example, whether standard 2D images or tomographic images were used. This is unlikely to be a major limitation since tumor diameter measurements from both techniques are relatively similar compared to histopathology [32]. Additionally, the reasons for missing tumor diameter data are unclear: data could have been missing because the clinicians were unable to measure the tumor (e.g., due to it being mammographically occult), because they failed to record the measured value, or because the recorded value was not sent to the CRN. The latter is the most likely explanation for the majority of missing mammographic information as this data cannot be submitted by centers that are not affiliated with BreastScreen Norway [33]. Nonetheless, the overall reporting rate has improved since 2012, and was 72% in 2016 [33,34]. Only 3% of histopathologic tumor diameter information was missing for women who had surgery for breast cancer during 2009–2011 [35], thus missing histopathologic tumor diameter information in our study is likely due to women receiving neoadjuvant therapy (where post-treatment staging, ypT, is reported instead of pT). No reason for missing data seems likely to have caused any systematic bias in the terminal digits of tumor diameters in the final sample.

In this population-based study, we observed a preference for reporting tumor diameters with terminal digits of zero or five, corresponding to whole and half centimeter values. Further, our results support the notion that absolute agreement between mammographic and histopathologic tumor diameter is moderate and decreases with increasing tumor size. The current guidelines for TNM staging do not consider terminal digit preference and histopathological review studies are needed to investigate the potential extent of understaging and under-treatment resulting from this source of measurement error.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Norwegian ExtraFoundation for Health and Rehabilitation via the Norwegian Breast Cancer Society under Grant 2018/FO201362 (to Kaitlyn Tsuruda).

ORCID

Kaitlyn M. Tsuruda http://orcid.org/0000-0001-8749-6237
Solveig Hofvind http://orcid.org/0000-0003-0178-8939
Lars A. Akslen http://orcid.org/0000-0003-2710-9543
Solveig R. Hoff http://orcid.org/0000-0001-7928-3821
Marit B. Veierød http://orcid.org/0000-0002-2083-2758

References

[1] Hayes SJ. Terminal digit preference occurs in pathology reporting irrespective of patient management implication. J Clin Pathol. 2008;61(9):1071–1072.
[2] Veierød MB, Page CM, Aaserud S, et al. Melanoma staging: varying precision and terminal digit clustering in Breslow thickness data is evident in a population-based study. J Am Acad Dermatol. 2018;79(1):118–125.e1.
[3] den Bakker MA, Damhuis RAM. Pentameric last-digit preference and stage border avoidance in pathology measurement. Histopathology. 2018;73(3):510–513.
[4] Coburn NG, Clarke-Pearson E, Chung MA, et al. A novel approach to T classification in tumor-node-metastasis staging of breast cancer. Am J Surg. 2006;192(4):434–438.
[5] Tabár L, Tucker L, Davenport RR, et al. The use of mammographic tumour feature significantly improves outcome prediction of breast cancers smaller than 15 mm: a reproducibility study from two comprehensive breast centres. Memo. 2011;4(3):149–157.
[6] Verschraegen C, Vinh-Hung V, Cserni G, et al. Modeling the effect of tumor size in early breast cancer. Ann Surg. 2005;241(2):309–318.
[7] Badve SS, Beitsch PD, Bose S, et al. Part XI, Breast In: Amin MB, Edge S, Greene F, et al. editors. AJCC Cancer Staging Manual. 8th ed. Heidelberg, (Germany): Springer International Publishing; 2017.
[8] Union for International Cancer Control. TNM classification of malignant tumours In: Brierley JD, Gospodarowicz MK, Wittekind

C, editors. 8th ed. Oxford, UK & Hoboken, NJ: John Wiley & Sons; 2016.
[9] Hayes DF, Allred C, B.O A, et al. Part VII, Breast In: Edge S, Byrd D, Compton CC, et al. editors. AJCC Cancer Staging Manual. 7th ed. Heidelberg, (Germany): Springer International Publishing; 2010.
[10] Heusinger K, Lohberg C, Lux MP, et al. Assessment of breast cancer tumor size depends on method, histopathology and tumor size itself*. Breast Cancer Res Tr. 2005;94(1):17–23.
[11] Marinovich ML, Macaskill P, Irwig L, et al. Agreement between MRI and pathologic breast tumor size after neoadjuvant chemotherapy, and comparison with alternative tests: individual patient data meta-analysis. BMC Cancer. 2015;15(1):662.
[12] Brystkreft – handlingsprogram [Internet]. Oslo: Norwegian Directorate of Health; 2019. [updated 2019/01/17; cited 2019/04/05]. Available from: https://helsedirektoratet.no/retningslinjer/nasjonalt-handlingsprogram-med-retningslinjer-for-diagnostikk-behandling-og-oppfolging-av-pasienter-med-brystkreft
[13] Kvalitetsmanual i Mammografiprogrammet – Patologi. Oslo: Mammografiprogrammet, Kreftregisteret; 2018. Available from: https://www.kreftregisteret.no/globalassets/publikasjoner-og-rapporter/mammografiprogrammet/20180206_kvalitetsmanual-patologi—nettversjon_med-forside.pdf
[14] Larsen IK, Smastuen M, Johannesen TB, et al. Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. Eur J Cancer. 2009;45(7):1218–1231.
[15] Cancer Registry of Norway. Cancer in Norway 2016 – cancer incidence, mortality, survival and prevalence in Norway. Oslo Cancer Registry of Norway; 2017.
[16] Hofvind S, Tsuruda K, Mangerud G, et al. The Norwegian Breast Cancer Screening Program, 1996-2016: celebrating 20 years of organised screening in Norway. Oslo: Cancer Registry of Norway; 2017.
[17] Forskrift om innsamling og behandling av helseopplysninger i Kreftregisteret (Kreftregisterforskriften) [Regulations on the collection and processing of personal health data in the Cancer Registry of Norway (Cancer Registry Regulations)], FOR-2001-12-21-1477 2002. Norwegian.
[18] Larsen IK, Myklebust TÅ, Johannesen TB, et al. Stage-specific incidence and survival of breast cancer in Norway: The implications of changes in coding and classification practice. Breast. 2018;38:107–113.
[19] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–163.
[20] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327(8476):307–310.
[21] R Core Team. R: a language and environment for statistical computing. version 3.5.0. Vienna, Austria: R Foundation for Statistical Computing; 2018.
[22] Gamer M, Lemon J, Singh IFP. irr: Various coefficients of interrater reliability and agreement. Version 0.84. 2012.
[23] Signorell A, Aho K, Alfons A, et al. DescTools: tools for descriptive statistics. Version 0.99.26. 2018.
[24] Pineros M, Parkin DM, Ward K, et al. Essential TNM: a registry tool to reduce gaps in cancer staging information. Lancet Oncol. 2019;20(2):e103–e11.
[25] Stein RG, Wollschlager D, Kreienberg R, et al. The impact of breast cancer biological subtyping on tumor size assessment by ultrasound and mammography – a retrospective multicenter cohort study of 6543 primary breast cancer patients. BMC Cancer. 2016;16(1):459.
[26] Gruber IV, Rueckert M, Kagan KO, et al. Measurement of tumour size with mammography, sonography and magnetic resonance imaging as compared to histological tumour size in primary breast cancer. BMC Cancer. 2013;13(1):328.
[27] Hieken TJ, Harrison J, Herreros J, et al. Correlating sonography, mammography, and pathology in the assessment of breast cancer size. Am J Surg. 2001;182(4):351–354.
[28] Cronin-Fenton DP, Kjaersgaard A, Ahern TP, et al. Validity of Danish Breast Cancer Group (DBCG) registry data used in the predictors of breast cancer recurrence (ProBeCaRe) premenopausal breast cancer cohort study. Acta Oncol. 2017;56(9):1155–1160.
[29] Sakkestad ST, Olsen BC, Karliczek A, et al. Validity of Norwegian Rectal Cancer Registry data at a major university hospital 1997-2005. Acta Oncol. 2015;54(10):1723–1728.
[30] Lofgren L, Eloranta S, Krawiec K, et al. Validation of data quality in the Swedish National Register for Breast Cancer. BMC Public Health. 2019;19(1):495
[31] Porta M. A Dictionary of Epidemiology. Oxford, UK: Oxford University Press; 2008.
[32] Marinovich ML, Macaskill P, Bernardi D, et al. Systematic review of agreement between tomosynthesis and pathologic tumor size for newly diagnosed breast cancer and comparison with other imaging tests. Expert Rev Med Devices. 2018;15(7):489–496.
[33] Cancer Registry of Norway. Årsrapport 2017 med resultater og forbedringstiltak fra Nasjonalt kvalitetsregister for brystkreft. [National quality registry for breast cancer annual report, 2017]. Oslo: Cancer Registry of Norway; 2018. Norwegian.
[34] Cancer Registry of Norway. Årsrapport 2016 med resultater og forbedringstiltak fra Nasjonalt kvalitetsregister for brystkreft. [National quality registry for breast cancer annual report, 2016]. Oslo: Cancer Registry of Norway; 2017. Norwegian.
[35] Cancer Registry of Norway. Patologirapport. [Pathology report]. Oslo: Cancer Registry of Norway; 2015. Norwegian.

Table S1: Terminal digits of mammographic and histopathologic tumour diameters among women diagnosed with invasive T1–T3 breast cancer in Norway during 2012-2016

Tumour diameter (mm)	**n**	**Percentage with terminal digit**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**
Mammographic											
1-10	1801	0.1	0.3	0.7	2.6	9.0	9.8	14.9	16.2	16.7	29.8
11-20	2970	9.1	13.3	9.9	8.2	15.5	7.0	7.1	7.8	5.5	16.5
21-30	1495	6.7	12.2	9.4	7.9	19.3	7.6	7.1	6.3	3.6	19.9
31-40	706	6.5	11.9	10.2	7.4	16.1	7.1	7.2	5.8	2.5	25.2
41-50	378	6.1	2.9	7.9	5.8	24.1	5.6	5.8	4.8	1.9	35.2
> 50	442	3.8	7.0	4.5	6.3	16.5	5.9	4.5	5.7	2.3	43.4
Total	7792	5.9	9.1	7.3	6.5	15.3	7.6	8.7	9.0	7.1	23.4
Histopathologic (whole numbers)											
1-10	2753	0.8	2.0	2.5	4.1	6.5	10.9	14.2	17.8	19.5	21.8
11-20	5641	9.6	12.8	8.5	7.7	16.6	7.5	8.2	9.3	6.4	13.2
21-30	3107	11.0	12.8	10.1	8.1	23.0	4.6	5.4	5.3	2.4	17.2
31-40	997	5.9	11.9	6.7	5.9	31.7	4.1	4.1	3.3	2.0	24.3
41-50	369	3.8	10.3	5.1	3.3	33.9	1.6	5.7	4.9	1.9	29.5
> 50	300	7.7	6.0	4.0	4.0	28.7	1.7	2.0	3.0	0.7	42.3
Total	13,167	7.6	10.2	7.3	6.7	17.9	7.0	8.3	9.4	7.6	17.9

Histopathologic (with decimal)											
1-10.0	276	5.1	5.8	4.0	1.4	63.4	4.7	4.0	5.4	6.2	0.0
10.1-20.0	82	3.7	12.2	1.2	4.9	52.4	6.1	6.1	6.1	7.3	0.0
20.1-30.0	13	15.4	0.0	7.7	0.0	61.5	7.7	0.0	0.0	7.7	0.0
30.1-40.0	1	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
40.1-50.0	1	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
> 50.0	1	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	374	5.3	7.0	3.5	2.1	61.0	5.1	4.3	5.3	6.4	0.0

Figure S1

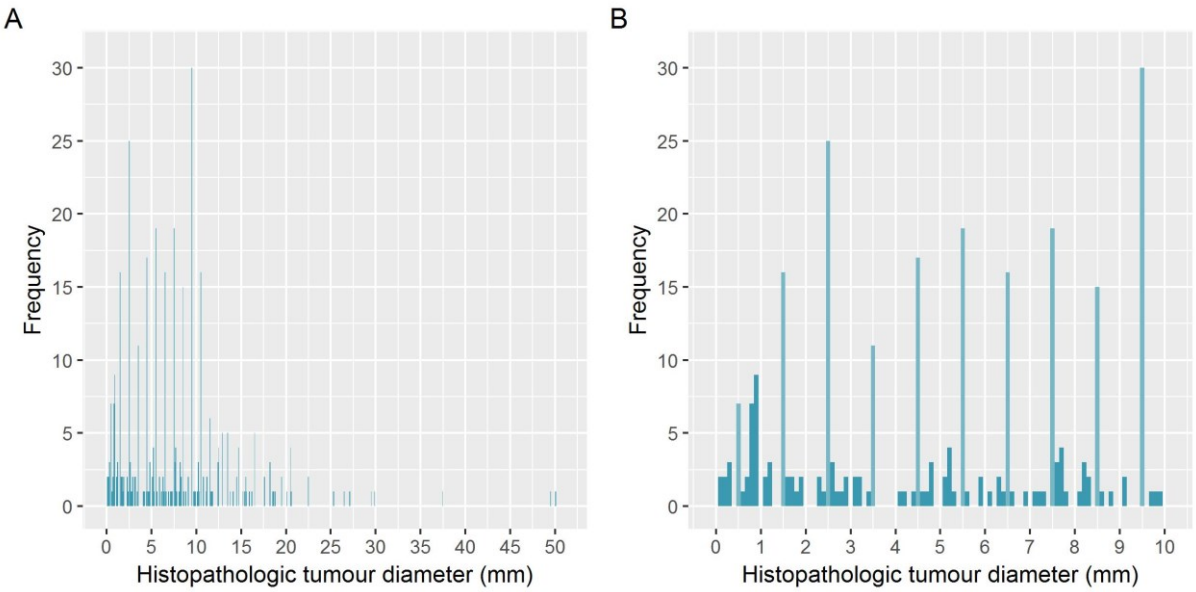
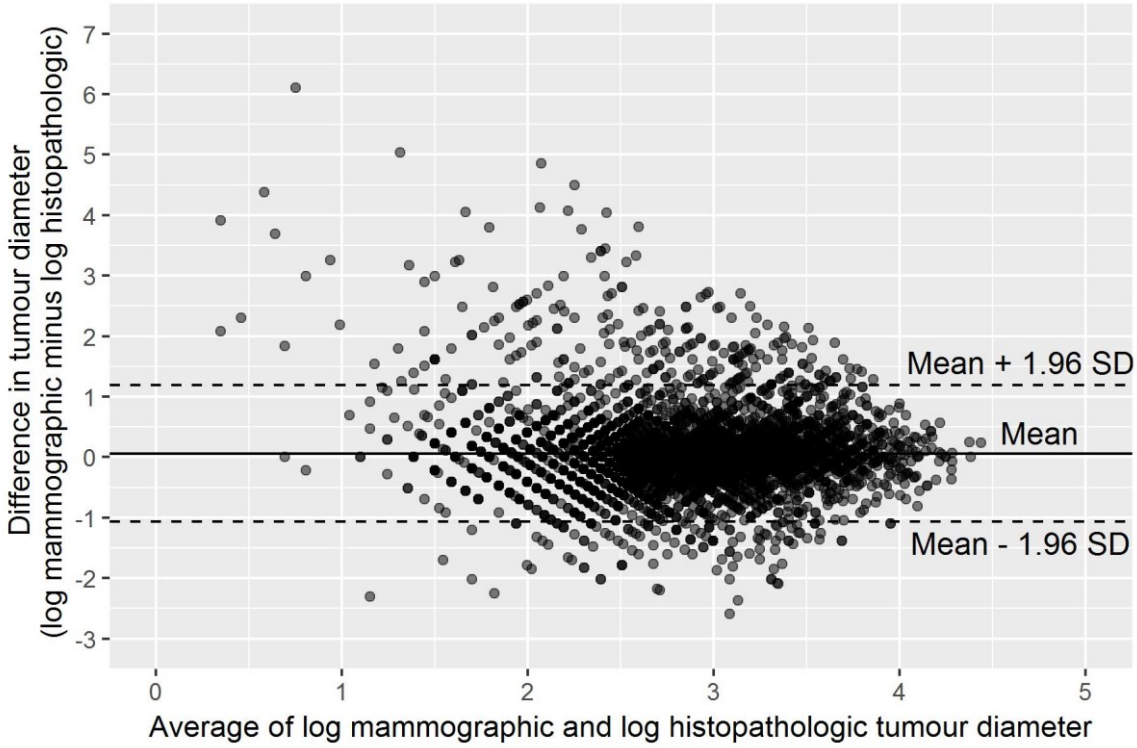


Figure S2


Difference in tumour diameter
(log mammographic minus log histopathologic)
7
6
5
4
3
2
1
0
-1
-2
-3
Mean + 1.96 SD
Mean
Mean - 1.96 SD
0
1
2
3
4
5
Average of log mammographic and log histopathologic tumour diameter

Paper 2

Tsuruda KM, Hovda T, Bhargava S, Veierød MB, Hofvind S. Survival among women diagnosed with screen-detected or interval breast cancer classified as true, minimal signs, or missed through an informed radiological review. Eur Radiol. 12 November 2020. https://doi.org/10.1007/s00330-020-07340-4. Online ahead of print.

II

European Radiology
https://doi.org/10.1007/s00330-020-07340-4

BREAST

Survival among women diagnosed with screen-detected or interval breast cancer classified as true, minimal signs, or missed through an informed radiological review

Kaitlyn M. Tsuruda[1,2] · Tone Hovda[3,4] · Sameer Bhargava[5] · Marit B. Veierød[2] · Solveig Hofvind[1,6]

Received: 30 March 2020 / Revised: 11 August 2020 / Accepted: 22 September 2020

Abstract

Objectives "True" breast cancers, defined as not being visible on prior screening mammograms, are expected to be more aggressive than "missed" cancers, which are visible in retrospect. However, the evidence to support this hypothesis is limited. We compared the risk of death from any cause for women with true, minimal signs, and missed invasive screen-detected (SDC) and interval breast cancers (IC).

Methods This nation-wide study included 1022 SDC and 788 IC diagnosed through BreastScreen Norway during 2005–2016. Cancers were classified as true, minimal signs, or missed by five breast radiologists in a consensus-based informed review of prior screening and diagnostic images. We used multivariable Cox regression to estimate hazard ratios (HRs) and 95% confidence intervals (CIs) for the risk of death from any cause associated with true, minimal signs, and missed breast cancers, adjusting for age at diagnosis, histopathologic tumour diameter and grade, and subtype. Separate models were created for SDC and IC.

Results Among SDC, 463 (44%) were classified as true and 242 (23%) as missed; among IC, 325 (39%) were classified as true and 235 (32%) missed. Missed SDC were associated with a similar risk of death as true SDC (HR = 1.20, 95% CI (0.49, 2.46)). Similar results were observed for missed versus true IC (HR = 1.31, 95% CI (0.77, 2.23)).

Conclusions We did not observe a statistical difference in the risk of death for women diagnosed with true or missed SDC or IC; however, the number of cases reviewed and follow-up time limited the precision of our estimates.

Key Points

- *An informed radiological review classified screen-detected and interval cancers as true, minimal signs, or missed based on prior screening and diagnostic mammograms.*
- *It has been hypothesised that true cancers, not visible on the prior screening examination, may be more aggressive than missed cancers.*
- *We did not observe a statistical difference in the risk of death from any cause for women with missed versus true screen-detected or interval breast cancers.*

Keywords Breast neoplasms · Mammography · Mass screening · Survival rate · Early detection of cancer

✉ Solveig Hofvind
Solveig.Hofvind@kreftregisteret.no

[1] Section for Breast Cancer Screening, Cancer Registry of Norway, PO Box 5313, Majorstuen, 0304 Oslo, Norway

[2] Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, PO Box 1122, Blindern, 0317 Oslo, Norway

[3] Department of Radiology, Vestre Viken Hospital Trust, PO Box 800, 3004 Drammen, Norway

[4] Institute of Clinical Medicine, University of Oslo, PO Box 1171, Blindern, 0318 Oslo, Norway

[5] Division of Oncology, Department of Medicine, Bærum Hospital, Vestre Viken Hospital Trust, PO Box 800, 3004 Drammen, Norway

[6] Faculty of Health Sciences, Oslo Metropolitan University, Pilestredet Campus, PO Box 4 St. Olavs plass, N-0130 Oslo, Norway

Published online: 12 November 2020

Abbreviations

CI	Confidence interval
CRN	Cancer Registry of Norway
DCIS	Ductal carcinoma in situ
ER	Estrogen receptor
HER2	Human epidermal growth factor receptor 2
PACS	Picture archiving and communication system
PR	Progesterone receptor

Introduction

Mammographic screening is considered the best approach to detect breast cancer at an early stage and thereby reduce breast cancer mortality [1, 2]. For mass screening to be effective, radiologist sensitivity, the ability to correctly identify women *with* breast cancer, must be balanced against specificity, the ability to correctly identify women *without* breast cancer. False-positive screening exams are associated with temporary uncertainty and anxiety [3–5], and healthcare costs for further assessment [6]; however, this follow-up can provide confirmation that a woman does not have breast cancer. On the other hand, false negatives may lead to a delayed diagnosis of breast cancer and can lower women's confidence in mammographic screening [7].

Retrospective radiologic reviews of prior screening mammograms can give insights into the effectiveness and quality of mammographic screening [1, 8]. These reviews are typically limited to the prior screening mammograms of interval breast cancers, but can also include the prior screening mammograms of screen-detected breast cancers and are often performed with access to diagnostic mammograms [9–11]. Typically, reviewing radiologists classify cancers as "true", "minimal signs", "missed", or "occult" [1]. Cancers that are classified as not visible on prior screening mammograms but that develop following a true-negative screening examination are considered true. Minor abnormalities that are regarded as visible on the prior screening mammograms but did not lead to a diagnosis of breast cancer are considered minimal signs. Cancers that are retrospectively visible on prior screening mammograms are considered missed at the prior screen and the prior screening examination is then considered a false negative. Occult cancers are those that are not regarded as mammographically visible at diagnosis but may be symptomatic or diagnosed through other modalities such as ultrasound. The rates of these types of cancers in an organised screening program are associated with the sensitivity of the interpreting radiologists and that of any follow-up assessment, as well as the review process [12–15].

The histopathology of true, minimal signs, and missed screen-detected breast cancers is not well described [9, 16]. True interval breast cancers have more often been histopathologic grade 3 with a smaller tumour diameter than missed breast cancers, but other aspects of histopathology have not demonstrated consistent results [17–22]. It has been hypothesised that the short sojourn time of true breast cancers indicates that they are more aggressive than missed breast cancers [9, 23]. However, we are not aware of any studies that report the survival of women with true and missed screen-detected breast cancer, and three of four studies did not observe a difference in the overall survival of women with true and missed interval breast cancers [18, 19, 21, 23]. These survival results are based on decades-old data, and breast cancer screening, diagnosis, and treatment have since improved considerably [8, 24].

Radiologic reviews often aim to understand the distribution of true, minimal signs, and missed cancers and reduce the rate of missed cancers in order to improve the quality of mammographic screening. However, it is also important to evaluate whether having a breast cancer classified as true, minimal signs, or missed has prognostic implications for women attending screening. The objective of this retrospective study was to re-use data from a completed informed radiologic review to describe the histopathologic findings and survival associated with these three classifications of screen-detected and interval breast cancers within a population-based breast cancer screening program.

Methods

This study was part of a project approved by the Oslo University Hospital data protection official for research (PVO 2016-4696).

Retrospective radiological review

The radiologic classifications used in this study were obtained during a nationwide, multicentre informed review of ductal carcinoma in situ (DCIS) and invasive breast cancer diagnosed in BreastScreen Norway. The Cancer Registry of Norway (CRN) administers this population-based screening program and also administered this review, which was performed between September 2016 and April 2017. The review included digital mammograms from consecutive round screen-detected breast cancers and interval breast cancers. BreastScreen Norway offered screening with digital mammography in a single-centre study starting in 2000, and this technology was implemented at all 16 centres in the program by the fall of 2011 [25]. Screen-detected cancer was defined as breast cancer diagnosed after a recall for further assessment due to abnormal mammographic findings, and interval cancer was defined as breast cancer diagnosed within 24 months of a negative screen, or 6–24 months after a false-positive screening result.

The review was designed to include a stratified sample of 75 screen-detected and 75 interval cancers diagnosed at each of the 16 breast centres. With respect to the national distribution of

breast cancers, this non-proportional sampling method over-represented smaller centres and under-represented larger centres. However, this gave participating radiologists an equal opportunity to review and learn from cases diagnosed within their own centres. Screen-detected cancers were oversampled at low-volume centres with too few interval cancers diagnosed after screening with digital mammography. Recently diagnosed cancer cases were preferred over older cases to facilitate retrieval from the picture archiving and communication systems (PACS). The review is described elsewhere [26].

Briefly, the review was performed at the breast centres by a pool of 37 radiologists who had read at least 5000 mammograms during the past 2 years. The centres were randomly paired and radiologists from paired centres reviewed each other's cases. A panel of five radiologists reviewed each case: two from the reviewing centre, two from the paired centre, and one (T.H.) was present at every session. The panel had access to screening and diagnostic images. Through consensus, or a majority vote if consensus could not be reached, the panel of radiologists classified all cancer cases as "true", "minimal signs, actionable", "minimal signs, non-actionable", "missed", or "occult" indicating whether a cancer was visible and/or perceived at the prior screening examination (Table 1). Information on surgical treatment and histopathology was provided after a case was reviewed.

Study sample

The review included mammograms from 1227 screen-detected and 1015 interval breast cancers (both DCIS and invasive). However, in this study, we sequentially excluded DCIS, occult cases, and cases for whom information about postoperative histopathological tumour diameter was unavailable (Fig. 1). The proportion of DCIS in the review reflected population averages in BreastScreen Norway [25]. Women with DCIS in Norway are typically treated with breast-conserving surgery and adjuvant radiation therapy, which is associated with low long-term rates of local recurrence [27]. We excluded women with DCIS because the excellent survival in this group makes it difficult to conduct an informative survival analysis. The largest tumour was included for multifocal (n = 44 screen-detected and 36 interval) and bilateral cancers (n = 21 screen-detected and 15 interval).

Data sources and variables

Three radiological review classifications were used in this study: true, minimal signs, and missed. The minimal signs classification included both actionable and non-actionable cases (Table 1).

The CRN provided information about cancer diagnoses and prior screening exams, including women's age at diagnosis, date of screening and diagnosis, screening location, and mode of detection (screen-detected or interval). The CRN also provided information on histopathologic type (invasive no special type (NST), lobular, other), histopathologic tumour diameter (mm), histopathologic grade (1, 2, 3), lymph node status, estrogen receptor (ER) and progesterone receptor (PR) status, Ki67 expression, and human epidermal growth factor receptor 2 (HER2) status. Breast cancer subtype (Luminal A-like, Luminal B-like, HER2+, triple negative) was determined by applying a clinicopathologic surrogate definition of intrinsic subtypes to ER, PR, Ki67, and HER2 information [28]. These variables are described in detail in Table S1.

Women were followed from date of histologically verified invasive breast cancer until death, emigration, or December 31, 2018. Information on death and emigration was obtained from the CRN, which regularly receives information from the national Cause of Death Registry [29].

Statistical analysis

Descriptive results were presented as proportions (95% confidence intervals, CIs, calculated using the Wilson score interval [30]), means (standard deviations, SDs), and medians (95% CIs from quantile regression with standard errors based on 100 bootstrap replications).

Kaplan-Meier estimates were used for overall survival in true, minimal signs, and missed cancers. Nelson-Aalen cumulative hazard estimates were used to estimate the risk of breast cancer death. Differences between true, minimal signs, and missed cancers were tested using the log-rank test. We used Cox regression with time since diagnosis as the time variable to estimate hazard ratios (HRs) with 95% CIs for the risk of death from all causes in true, minimal signs, and missed cancers. We adjusted for age at diagnosis and included tumour diameter, grade, and subtype as confounders based on a priori knowledge of their relationship with the exposure and outcome in interval cancers [17–22]. All analyses were conducted separately for screen-detected and interval cancers due to the potential for lead time bias in combined analyses. The proportional hazards assumption was checked using graphical methods and Schoenfeld residuals [31]. This assumption was initially violated in the analysis of interval cancers, but was satisfied after splitting the follow-up time after 3 years, and stratifying on subtype.

Multiple imputation with chained equations was used to impute missing data for grade; lymph node status; ER, PR, and HER2 positivity; and Ki67 expression. Subtype was determined after imputation. Given detection mode and year of diagnosis, data were assumed to be approximately missing completely at random. To increase predictive power, the imputation models also included the radiological classification (true, minimal signs, and missed), tumour diameter, screening centre, age at diagnosis, information about whether women were alive at the end of follow-up, and the Nelson-Aalen cumulative hazard estimator for overall survival [32]. We presented results based on 40 imputed data sets.

Table 1 Definitions of radiological and study classifications of true, minimal signs, and missed screen-detected and interval breast cancers

Radiological classification	Study classification	Definition
True	True	No abnormalities visible on prior screening mammograms at the cancer site (true-negative prior screen), followed by a diagnosis of interval breast cancer, or screen-detected breast cancer during the subsequent screening round
Minimal signs, actionable	Minimal signs	Minor abnormalities visible on prior screening mammograms at the cancer site. Recall would have warranted, but was not expected within the screening program
Minimal signs, non-actionable	Minimal signs	Non-specific findings visible on prior screening mammograms at the cancer site. Recall not possible or expected within the screening program
Missed	Missed	Obvious abnormalities visible on prior screening mammograms at the cancer site (false negative prior screen) that resulted in interval breast cancer or screen-detected breast cancer during the subsequent screening round
Occult	Excluded	No mammographically visible findings at diagnosis

We conducted a sensitivity analysis in which we included women without tumour diameter information, and did not use tumour diameter as a covariate in the imputation or Cox regression models.

All analyses were performed using Stata version 16.0 (StataCorp).

Results

The final study sample consisted of 1022 screen-detected and 788 interval cancers with prior screening examinations between January 2005 and March 2016 (Fig. 1). Among screen-detected cancers, 457 (45%) were classified as true, 336 (33%) as minimal signs, and 229 (22%) as missed. Among interval cancers, 310 (39%) were classified as true, 254 (32%) as minimal signs, and 224 (28%) as missed.

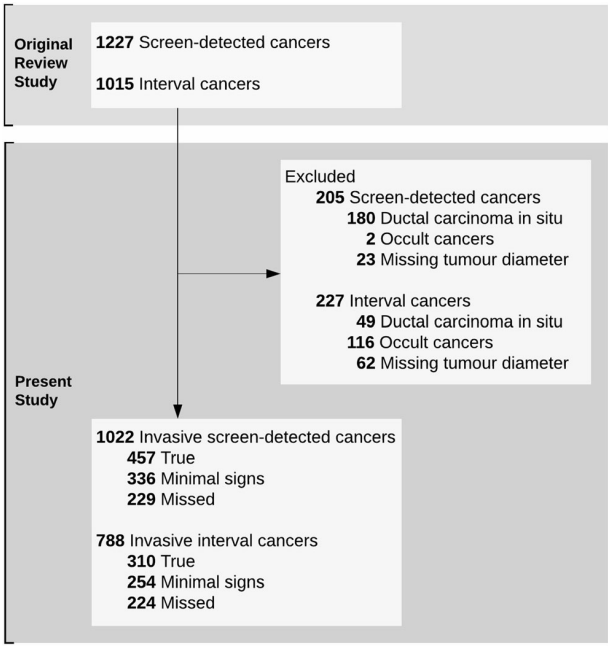


Fig. 1 Number of individuals included and excluded. Individuals were excluded sequentially using the exclusion criteria

Mean (SD) age at diagnosis did not differ by more than 2 years for women with true (62 (5.1)), minimal signs (62 (4.7)), or missed (63 (4.8)) screen-detected cancer, or for women with true (59 (5.8)), minimal signs (60 (5.7)), or missed (61 (5.2)) interval cancer.

Histopathologic findings

True screen-detected cancers had less favourable histopathology than minimal signs and missed cancers, which had comparable histopathology (Table 2). In particular, true screen-detected cancers had a higher proportion of grade 3 tumours (30.0%) than minimal signs (14.9%), or missed cancers (13.7%), and were more likely to be triple negative (9.8% versus 2.3% and 2.9%). True interval cancers also had less favourable histopathology than minimal signs and missed interval cancers, which generally had comparable histopathologic characteristics. True interval cancers were more likely to be grade 3 (46.7%) than minimal signs (36.1%), or missed cancers (35.9%). The proportion of triple negative cancers was 18.1% among true interval cancers, 14.5% among minimal signs, and 9.6% among missed cancers.

Survival

Median follow-up was 5.4 years (range 0.2–12.8) for women with screen-detected cancer; 43 (4.2%) died from any cause and 10 (1.0%) died from breast cancer. Among women with interval cancer, median follow-up was 5.6 years (range

Table 2 Histopathologic and clinicopathologic tumour characteristics of true, minimal signs, and missed screen-detected and interval breast cancers (proportions with 95% confidence intervals, CIs, unless otherwise specified)

Tumour characteristic	Screen-detected breast cancers						Interval breast cancers					
	True		Minimal signs		Missed		True		Minimal signs		Missed	
	n = 457		*n* = 336		*n* = 229		*n* = 310		*n* = 254		*n* = 224	
	n	% (95% CI)	*n*	% (95% CI)	*n*	% (95% CI)	*n*	% (95% CI)	*n*	% (95% CI)	*n*	% (95% CI)
Type												
Invasive NST[a]	408	89.9 (86.7, 92.3)	280	84.1 (79.8, 87.6)	198	86.8 (81.8, 90.6)	265	86.6 (82.3, 90.0)	212	83.8 (78.8, 87.8)	186	83.4 (78.0, 87.7)
Lobular	35	7.7 (5.6, 10.5)	37	11.1 (8.2, 14.9)	20	8.8 (5.8, 13.2)	35	11.4 (8.3, 15.5)	33	13.0 (9.4, 17.8)	26	11.7 (8.1, 16.5)
Other	11	2.4 (1.4, 4.3)	16	4.8 (3.0, 7.7)	10	4.4 (2.4, 7.9)	6	2.0 (0.9, 4.2)	8	3.2 (1.6, 6.1)	11	4.9 (2.8, 8.6)
Information not available	3	-	3	-	1	-	4	-	1	-	1	-
Tumour diameter (mm)												
Median (95% CI)	457	12 (11, 13)	336	13 (12, 14)	229	13 (12, 14)	310	19 (18, 20)	254	20 (18, 22)	224	20 (18, 22)
Histologic grade												
1	103	22.9 (19.2, 27.0)	111	33.0 (28.2, 38.2)	73	32.3 (26.5, 38.6)	26	8.5 (5.9, 12.2)	43	17.3 (13.1, 22.5)	30	13.8 (9.9, 19.0)
2	212	47.1 (42.5, 51.7)	175	52.1 (46.7, 57.4)	122	54.0 (47.5, 60.4)	137	44.8 (39.3, 50.4)	116	46.6 (40.5, 52.8)	109	50.2 (43.6, 56.8)
3	135	30.0 (25.9, 34.4)	50	14.9 (11.5, 19.1)	31	13.7 (9.8, 18.8)	143	46.7 (41.2, 52.3)	90	36.1 (30.4, 42.3)	78	35.9 (29.9, 42.5)
Information not available	7	-	0	-	3	-	4	-	5	-	7	-
Lymph node status												
Negative	352	78.0 (74.0, 81.6)	264	80.7 (76.1, 84.6)	181	80.4 (74.8, 85.1)	164	53.9 (48.3, 59.5)	153	61.0 (54.8, 66.8)	126	57.5 (50.9, 63.9)
Information not available	6	-	9	-	4	-	6	-	3	-	5	-
Subtype												
Luminal A-like	121	29.0 (24.9, 33.5)	114	37.4 (32.1, 42.9)	76	36.4 (30.1, 43.1)	31	11.0 (7.9, 15.2)	45	19.7 (15.1, 25.4)	52	25.0 (19.6, 31.3)
Luminal B-like	239	57.3 (52.5, 62.0)	174	57.0 (51.4, 62.5)	123	58.9 (52.1, 65.3)	175	62.3 (56.5, 67.7)	141	61.8 (55.4, 67.9)	118	56.7 (49.9, 63.3)
HER2+[b] (non-luminal)	16	3.8 (2.4, 6.1)	10	3.3 (1.8, 5.9)	4	1.9 (0.7, 4.8)	24	8.5 (5.8, 12.4)	9	3.9 (2.1, 7.3)	18	8.7 (5.5, 13.3)
Triple negative (ductal)	41	9.8 (7.3, 13.1)	7	2.3 (1.1, 4.7)	6	2.9 (1.3, 6.1)	51	18.1 (14.1, 23.1)	33	14.5 (10.5, 19.6)	20	9.6 (6.3, 14.4)
Information not available	40	-	31	-	20	-	29	-	26	-	16	-

Note: 95% confidence intervals were calculated for proportions the Wilson score interval, and for medians using quantile regression with standard errors based on 100 bootstrap replications

[a] *NST* no special type

[b] Human epidermal growth factor receptor 2 positive

0.3–14.8); 81 (10.3%) died from any cause and 39 (4.9%) died from breast cancer.

The Kaplan-Meier estimates for overall survival did not differ between true, minimal signs, and missed cancers, whether they were screen-detected (p = 0.82, Fig. 2a) or interval cancers (p = 0.43, Fig. 2b). We did not examine the Nelson-Aalen cumulative hazard for the risk of screen-detected breast cancer death because of the small number of deaths (5, 3, and 2 deaths among true, minimal signs, and missed cancers). The Nelson-Aalen cumulative hazard estimates of the risk of interval breast cancer death did not differ for true (16 deaths), minimal signs (11 deaths), or missed cancers (12 deaths; p = 0.80, Fig. 3).

The distribution of the imputed variables was comparable with that observed in complete case data (Table S2), and the results for complete case and multiple imputation analyses were similar (Table 3). We report the multiple imputation results here. In the multivariable Cox regression (Table 3), risk of death from any cause did not differ between minimal signs and true screen-detected cancers (HR = 1.04, 95% CI (0.51, 2.13)), or between missed and true screen-detected cancers (HR = 1.10, 95% CI (0.49, 2.46)). Similarly, the average risk of death from any cause during the entire follow-up period did not differ between minimal signs and true interval cancers (HR = 0.80, 95% CI (0.46, 1.37)), or missed and true interval cancers (HR = 1.31, 95% CI (0.77, 2.23)). Due to lack of proportional hazards, follow-up time was split for interval cancers. The risk of death among women with minimal signs interval cancers was lower than for women with true interval cancers during the first 3 years after diagnosis (HR = 0.29, 95% CI (0.10, 0.86)), but did not differ after the first 3 years (HR = 1.40, 95% CI (0.70, 2.80)). The risk of death from any cause did not differ for missed and true interval cancers before or after 3 years of follow-up.

Results from the sensitivity analysis did not change our main conclusions (Table S3). This analysis included women without tumour diameter information, and did not use tumour diameter as a covariate in the imputation or Cox regression models.

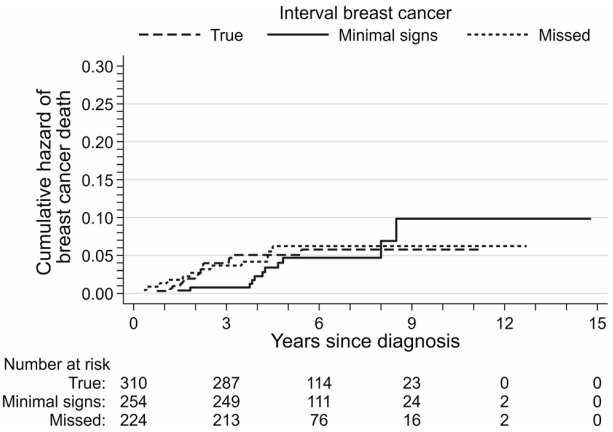


Fig. 3 Nelson-Aalen cumulative hazard estimates of breast cancer death for true, minimal signs, and missed interval breast cancers

Discussion

We observed that true screen-detected and interval cancers had less favourable histopathologic characteristics than minimal signs and missed cancers in BreastScreen Norway. However, we did not observe any differences in the overall survival between these groups 3 years after diagnosis within each mode of detection after adjusting for age at diagnosis, tumour diameter, grade, and subtype. Our study included only women whose histopathologic tumour diameter was available (97.7% of screen-detected and 92.7% of interval cancers). For these women, there may not be substantial inherent prognostic differences associated with the classification of a true versus a missed or minimal signs cancer within a given mode of diagnosis (screen-detected or interval cancer). Our study is, as far as we know, the first to report overall survival among true, minimal signs, and missed screen-detected and interval breast cancers detected exclusively with digital mammography.

Missed screen-detected breast cancers could represent underdiagnosis if they have aggressive tumour characteristics at diagnosis, or overdiagnosis if they are indolent tumours. In our study, missed screen-detected cancers were often invasive NST, Luminal B-like, without lymph node involvement.

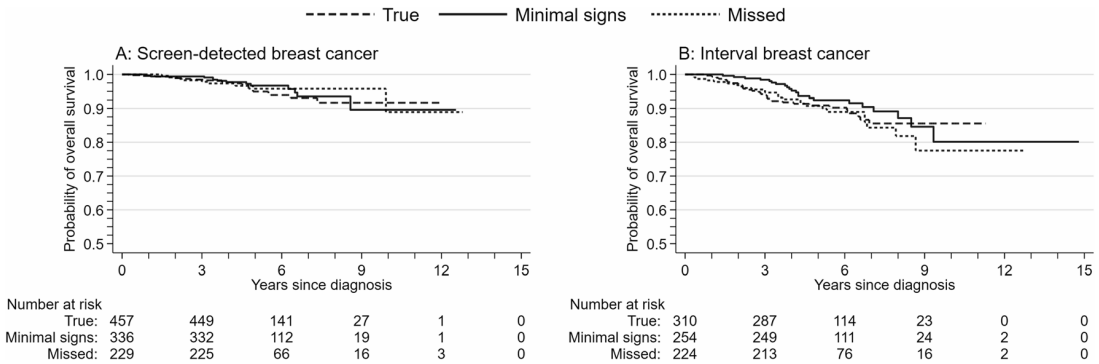


Fig. 2 Kaplan-Meier estimates of overall survival for true, minimal signs, and missed (**a**) screen-detected and (**b**) interval breast cancers

Table 3 Hazard ratios (HRs) with 95% confidence intervals (CIs) for death due to any cause among women diagnosed with screen-detected and interval breast cancers

	Complete case						Multiple imputation[b]			
	No. of subjects	No. of deaths	Age adjusted		Multivariable[a]		Age adjusted		Multivariable[c]	
			HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
Screen-detected breast cancers	921	35								
True	410	16	1.00	-	1.00	-	1.00	-	1.00	-
Minimal signs	305	11	0.93	(0.43, 2.00)	1.05	(0.48, 2.31)	0.87	(0.43, 1.73)	1.04	(0.51, 2.13)
Missed	206	8	1.12	(0.48, 2.63)	1.28	(0.53, 3.07)	0.90	(0.41, 1.97)	1.10	(0.49, 2.46)
Interval breast cancers, overall	702	73								
True	278	32	1.00	-	1.00	-	1.00	-	1.00	-
Minimal signs	223	18	0.69	(0.39, 1.23)	0.76	(0.42, 1.36)	0.77	(0.45, 1.32)	0.80	(0.46, 1.37)
Missed	201	23	1.02	(0.59, 1.76)	1.23	(0.71, 2.14)	1.13	(0.67, 1.90)	1.31	(0.77, 2.23)
Interval breast cancers, first 3 years[d]	702	29								
True	278	18	1.00	-	1.00	-	1.00	-	1.00	-
Minimal signs	223	3	0.21	(0.06, 0.70)	0.23	(0.07, 0.78)	0.27	(0.09, 0.79)	0.29	(0.10, 0.86)
Missed	201	8	0.65	(0.28, 1.51)	0.83	(0.35, 1.96)	0.83	(0.38, 1.81)	1.01	(0.45, 2.25)
Interval breast cancers, after 3 years[d]	673	44								
True	260	14	1.00	-	1.00	-	1.00	-	1.00	-
Minimal signs	220	15	1.31	(0.63, 2.73)	1.46	(0.70, 3.05)	1.34	(0.67, 2.66)	1.40	(0.70, 2.80)
Missed	193	15	1.51	(0.72, 3.17)	1.76	(0.83, 3.72)	1.50	(0.74, 3.07)	1.67	(0.81, 3.44)

[a] Adjusted for age at diagnosis, histopathologic tumour diameter and grade, and subtype

[b] Multiple imputation analyses conducted using chained equations and 40 generated data sets using 1022 screen-detected cancers (43 deaths) and 788 interval breast cancers (81 deaths)

[c] Model for screen-detected cancer adjusted for age at diagnosis, histopathologic tumour diameter and grade, and subtype. Models for interval cancer adjusted for age at diagnosis, histopathologic tumour diameter, and histopathologic grade, and stratified by subtype

[d] Follow-up time was split at 3 years due to lack of proportional hazards, the multiple imputation analyses included 788 interval cancers (32 deaths) in the model for the first 3 years, and 749 interval cancers (49 deaths) in the model for after 3 years

Assuming the majority of these cancers never displayed any clinical symptoms, we conject that women with missed screen-detected cancer were not underdiagnosed and would not have benefitted from an earlier diagnosis at the prior screen. On the other hand, if these missed screen-detected cancers were overdiagnosed, they would have a longer lead time and higher survival than true screen-detected cancers. We did not observe higher overall survival for missed screen-detected cancers. Breast cancer–specific survival outcomes will provide more information about potential overdiagnosis, but longer follow-up is needed to obtain sufficient statistical power.

The literature suggests that true interval cancers are more likely to be smaller [17, 18, 20–22] and histologic grade 3 [17–19, 21, 22] than missed interval cancers. Our results confirmed that true interval cancers are more likely to be grade 3, but found no more than a 1 mm difference in the median histopathologic tumour diameter of true, minimal signs, and missed cancers. In our study and others, such findings about tumour diameter only apply to women for whom this information was available. Contemporary use of neoadjuvant therapy may have narrowed the observed range of tumour diameters in our study compared with older studies that took place when neoadjuvant therapy was less common. Moreover, our results may differ from those in previous studies that calculated a mean [17, 21], which is sensitive to the skewed distribution of tumour diameter, or that used a categorical variable [18, 21, 22], which may be misclassified at commonly used cut-points like 10 and 20 mm [33–35].

The tumour histopathology for true interval cancers indicated these were less favourable than minimal signs or missed cancers. However, we did not observe any differences in the overall survival or risk of death from breast cancer between true and minimal signs cancers 3 or more years following diagnosis, or between true and missed cancers during the entire follow-up period. We observed that minimal signs cancers were associated with a lower risk of death from any cause than true cancers during the first 3 years following diagnosis, even after adjustment for tumour histopathology. Our study is the first to report this finding and further studies are needed to confirm this result. Effective treatment options for advanced cancers may partially explain why the longer-term survival

was similar for all classifications, despite differences in tumour histopathology. Our study did not include information about treatment regimens or long-term side effects which have the potential to highlight quality of life differences. Data completeness for oncological treatment is increasing at the national quality registry for breast cancer, and improved reporting of this information to the CRN may facilitate this type of analysis in the future [36]. In the absence of treatment data, longer follow-up may help us understand whether this survival profile for true, minimal signs, and missed interval cancers persists over a longer period, or whether any potential "treatment effect" is temporary.

Reviews are usually performed in a study setting with an artificially high volume of cancers, and radiologists are aware that they are being studied. This was also the case in our study and may limit the external generalizability of our results. Moreover, the distribution of true and missed cancers is sensitive to the review design used: higher proportions of missed interval cancers are associated with informed [13–15], and non-mixed reviews where cancer cases are not mixed in with negative screening examinations [11]. Lower proportions of missed interval cancers are associated with consensus-based reviews [15]. In our study, a panel of internal and external radiologists conducted a consensus-based informed review with one radiologist (T.H.) present during all classification activities to ensure methodological consistency. The panel had access to information about tumour localisation and features from diagnostic imaging, which may have led to a higher proportion of missed cancers in our study compared with studies with alternative review designs [21–23, 37].

Our study did not provide information about whether missed interval cancers would have a favourable prognosis had they been detected earlier as screen-detected breast cancers. Moreover, the statistical power was limited by the number of cases that the radiologists were able to review because review studies are resource intensive. Deep learning algorithms have the potential to interpret digital mammograms with a sensitivity and specificity comparable with radiologists, and there is now a focus on the potential for such algorithms to triage or identify true-negative screens so expert radiologists can focus on more challenging cases [38]. True interval cancers are the most frequent classification assigned in review studies [39] and using this technology to classify prior screening mammograms could substantially reduce the review workload for radiologists and facilitate larger studies than those conducted to date.

Unavailable histopathology data further limited the amount of information available for analysis—this is a common limitation of regression-based analyses because statistical software typically handles missing data by deleting the associated case [40]. We used multiple imputation to overcome the challenge of missing data and observed similar results from complete case and multiple imputation analyses. We could not impute tumour diameter information because it was not missing completely at random; therefore, we excluded women without this information from our sample. By excluding women whose histopathologic tumour diameter was not recorded at the CRN, we likely excluded women who received neoadjuvant therapy to downstage their tumour prior to surgery, thereby excluding women with the most aggressive tumours, particularly for interval cancers. Indeed, women without tumour diameter information in our study were more likely to have died during the follow-up period than women for whom this information was available. We performed a sensitivity analysis in which we included women without tumour diameter information, and did not use tumour diameter as a covariate in the imputation or Cox regression models. The results of this analysis did not change our main conclusions. However, we caution against generalising the results of our study to women who undergo neoadjuvant therapy.

The overall Cox regression model for interval cancers provided an estimate of the average risk of death over time. The proportional hazards assumption was violated in that model, which indicated that the risk of death in our sample was not constant over time. The models with split follow-up time indicated how that risk changed over time, but included fewer cases and had less statistical power than the overall model. Nonetheless, our study is one of the largest to evaluate the overall survival associated with true, missed, and minimal signs cancers in population-based screening. This is an important methodological strength of our work, as the adjusted complete case analyses omitted roughly 10% the available observations. Another strength of our study is that our sample included only cases detected with standard digital mammography, which is the current standard of care.

Conclusion

We did not observe any differences in the longer-term overall survival between women classified as having true, minimal signs or missed screen-detected or interval cancers. However, the number of cases reviewed and follow-up time limited the precision of our estimates. In the future, deep learning algorithms may increase the number of prior screening mammograms that can be reviewed and thereby facilitate the analysis of breast cancer–specific survival associated with these classifications. This could provide additional information about the potential for "under" or "over" diagnosed breast cancer.

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/s00330-020-07340-4).

Funding Open Access funding provided by the University of Oslo (incl Oslo University Hospital). This study received funding from the Dam Foundation via the Norwegian Breast Cancer Society (2018/FO201362).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Solveig Hofvind.

Conflict of interest SH is the head of BreastScreen Norway. The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise (Marit B. Veierød).

Informed consent Written informed consent was not required for this study because the Cancer Registry Regulations waive the requirement for informed consent for surveillance and quality assurance projects based on data collected as a part of invitation to and/or participation in BreastScreen Norway.

Ethical approval Institutional Review Board approval was not required because this was a quality assurance project. Approval was obtained from the Oslo University Hospital data protection official for research (PVO 2016-4696).

Study subjects or cohorts overlap Some study subjects may have been previously reported in other studies using population-based breast cancer data from the Cancer Registry of Norway.

The study sample from the submitted manuscript partially overlaps with a study submitted in March 2020 to European Radiology by Hovda T, Tsuruda KM, Hoff SR et al (2020) Radiological review of prior screening mammograms of screen-detected breast cancer. Eur Radiol. https://doi.org/10.1007/s00330-020-07130-y. The article by Hovda et al focuses on the mammographic features of true, missed, and minimal signs screen-detected breast cancer.

This article focuses on the survival associated with true, missed, and minimal signs breast cancer, both screen-detected and interval breast cancer, and does not discuss the mammographic features associated with these classifications.

Methodology
- retrospective
- observational
- multicentre study

References

1. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L (2006) European guidelines for quality assurance in breast cancer screening and diagnosis, 4th edn. Office for Official Publications of the European Communities, Luxembourg
2. Lauby-Secretan B, Scoccianti C, Loomis D et al (2015) Breast-cancer screening–viewpoint of the IARC Working Group. N Engl J Med 372:2353–2358
3. Solbjør M, Forsmo S, Skolbekken JA, Siersma V, Brodersen J (2018) Psychosocial consequences among women with false-positive results after mammography screening in Norway. Scand J Prim Health Care 36:380–389
4. Long H, Brooks JM, Harvie M, Maxwell A, French DP (2019) How do women experience a false-positive test result from breast screening? A systematic review and thematic synthesis of qualitative studies. Br J Cancer 121:351–358
5. Salz T, Richman AR, Brewer NT (2010) Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. Psychooncology 19:1026–1034
6. Moger TA, Kristiansen IS (2012) Direct and indirect costs of the Norwegian Breast Cancer Screening Program. University of Oslo, Oslo
7. Solbjør M, Skolbekken JA, Saetnan AR, Hagen AI, Forsmo S (2012) Mammography screening and trust: the case of interval breast cancer. Soc Sci Med 75:1746–1752
8. International Agency for Research on Cancer (2016) Breast Cancer Screening. IARC Handbook of Cancer Prevention, vol 15. IARC Handbook of Cancer Prevention Lyon, France
9. Saarenmaa I, Salminen T, Geiger U et al (1999) The visibility of cancer on earlier mammograms in a population-based screening programme. Eur J Cancer 35:1118–1122
10. Hoff SR, Abrahamsen AL, Samset JH, Vigeland E, Klepp O, Hofvind S (2012) Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography–results from a retrospective review. Radiology 264:378–386
11. Moberg K, Grundstrom H, Lundquist H, Svane G, Havervall E, Muren C (2000) Radiological review of incidence breast cancers. J Med Screen 7:177–183
12. Houssami N, Irwig L, Ciatto S (2006) Radiological surveillance of interval breast cancers in screening programmes. Lancet Oncol 7: 259–265
13. Lekanidi K, Dilks P, Suaris T, Kennett S, Purushothaman H (2017) Breast screening: What can the interval cancer review teach us? Are we perhaps being a bit too hard on ourselves? Eur J Radiol 94:13–15
14. Ciatto S, Catarzi S, Lamberini MP et al (2007) Interval breast cancers in screening: the effect of mammography review method on classification. Breast 16:646–652
15. Hofvind S, Skaane P, Vitak B et al (2005) Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology 237:437–443
16. Daly CA, Apthorp L, Field S (1998) Second round cancers: How many were visible on the first round of the UK national breast screening programme, three years earlier? Clin Radiol 53:25–28
17. Brekelmans CT, van Gorp JM, Peeters PH, Collette HJ (1996) Histopathology and growth rate of interval breast carcinoma: characterization of different subgroups. Cancer 78:1220–1228
18. Vitak B, Olsen KE, Manson JC, Arnesson LG, Stal O (1999) Tumour characteristics and survival in patients with invasive interval breast cancer classified according to mammographic findings at

the latest screening: a comparison of true interval and missed interval cancers. Eur Radiol 9:460–469
19. Porter GJ, Evans AJ, Burrell HC, Lee AH, Ellis IO, Chakrabarti J (2006) Interval breast cancers: prognostic features and survival by subtype and time since screening. J Med Screen 13:115–122
20. Hofvind S, Geller B, Skaane P (2008) Mammographic features and histopathological findings of interval breast cancers. Acta Radiol 49:975–981
21. Weber RJ, van Bommel RM, Louwman MW et al (2016) Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. Breast Cancer Res Treat 158:471–483
22. Sala M, Domingo L, Louro J et al (2018) Survival and disease-free survival by breast density and phenotype in interval breast cancers. Cancer Epidemiol Biomarkers Prev 27:908–916
23. Fong Y, Evans J, Brook D, Kenkre J, Jarvis P, Gower Thomas K (2014) The incidence and 10-year survival of interval breast cancers in Wales. Clin Radiol 69:e168–e172
24. Zurrida S, Veronesi U (2015) Milestones in breast cancer treatment. Breast J 21:3–12
25. Hofvind S, Tsuruda K, Mangerud G et al (2017) The Norwegian Breast Cancer Screening Program, 1996-2016: celebrating 20 years of organised screening in Norway Cancer in Norway 2016 - cancer incidence, mortality, survival and prevalence in Norway. Cancer Registry of Norway, Oslo
26. Hovda T, Tsuruda KM, Hoff SR, et al (2020) Radiological review of prior screening mammograms of screen-detected breast cancer. Eur Radiol. https://doi.org/10.1007/s00330-020-07130-y
27. Stuart KE, Houssami N, Taylor R, Hayen A, Boyages J (2015) Long-term outcomes of ductal carcinoma in situ of the breast: a systematic review, meta-analysis and meta-regression analysis. BMC Cancer 15:890
28. Goldhirsch A, Winer EP, Coates AS et al (2013) Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 24:2206–2223
29. Cancer Registry of Norway (2019) Cancer in Norway 2018 - cancer incidence, mortality, survival and prevalence in Norway. Cancer Registry of Norway, Oslo
30. Lydersen S, Fagerland MW, Laake P (2012) Categorical data and contingency tables. In: Veierød MB, Lydersen S, Laake P (eds) Medical statistics in clinical and epidemiological research. Gylendal Norsk Forlag, Oslo, pp 48–89
31. Kleinbaum DG, Klein M (2006) Survival analysis: a self-learning text. Springer Science & Business Media
32. White IR, Royston P (2009) Imputing missing covariate values for the Cox model. Stat Med 28:1982–1998
33. Tsuruda KM, Hofvind S, Akslen LA, Hoff SR, Veierød MB (2020) Terminal digit preference: a source of measurement error in breast cancer diameter reporting. Acta Oncol 59:260–267
34. den Bakker MA, Damhuis RAM (2018) Pentameric last-digit preference and stage border avoidance in pathology measurement. Histopathology 73:510–513
35. Tabár L, Tucker L, Davenport RR et al (2011) The use of mammographic tumour feature significantly improves outcome prediction of breast cancers smaller than 15 mm: a reproducibility study from two comprehensive breast centres. Memo 4:149–157
36. Cancer Registry of Norway (2019) 2018 Annual report with results from the National Breast Cancer Registry. Cancer Registry of Norway, Oslo
37. Domingo L, Salas D, Zubizarreta R et al (2014) Tumor phenotype and breast density in distinct categories of interval cancer: results of population-based mammography screening in Spain. Breast Cancer Res 16:R3
38. Sechopoulos I, Mann RM (2020) Stand-alone artificial intelligence - the future of breast cancer screening? Breast 49:254–260
39. Houssami N, Hunter K (2017) The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. NPJ Breast Cancer 3:12
40. Harrell FE Jr (2015) Regression Modeling Strategies, 2nd edn. Springer International Publishing

Supplemental material

Table S1: Definitions of tumour characteristics used in this study

Variable	Definition
Histopathologic tumour diameter	Measured to the nearest mm using a transparent ruler and described the distance between the outermost boundaries of an invasive lesion from microscopic slides [1]. If a measurement could not be obtained using the microscopic slide, one was taken from the macroscopic examination (formalin fixed specimen). This measurement was taken either from a single tissue slice, or estimated across all tissue slices containing microscopically verified invasive tumour tissue [2].
Lymph node status	Positive if micrometastases or metastases were detected in one or more regional or axillary lymph nodes. We prioritized histopathologic results from the diagnostic biopsy and used histopathologic results from the surgical specimen if the former were not available.
Estrogen receptor (ER) status	Positive if the sample displayed ≥10% reactivity, and negative otherwise. We prioritized histopathologic samples from the diagnostic biopsy and used histopathologic results from the surgical specimen if the former were not available.
Progesterone receptor (PR) status	Positive if the sample displayed ≥10% reactivity, and negative otherwise. We prioritized histopathologic samples from the diagnostic biopsy and used histopathologic results from the surgical specimen if the former were not available.
HER2[a] positivity	If in situ hybridization (ISH) was performed, a borderline or amplified result was considered positive. If ISH was not performed, positivity was determined using immunohistochemistry (IHC).
Ki-67 expression	Percentage of Ki-67 positive cells from a sample of 500 tumour cells from the surgical specimen [1]
Subtype	Clinico-pathologic surrogate definition of intrinsic subtype based on the St Gallen consensus [3] Luminal A: ER and PR positive, HER2 negative, and Ki67 proliferation < 20% Luminal B: ER positive, PR negative and HER2 negative; or ER positive, HER2 negative and Ki67 proliferation ≥ 20%; or ER positive and HER2 positive HER2 positive (non-luminal): ER and PR negative, and HER2 positive Triple negative (ductal): ER, PR, and HER2 negative

[a] Human epidermal growth factor receptor 2

[1] Brystkreft – handlingsprogram [Internet]. Oslo: Norwegian Directorate of Health; 2019. [updated 2019/01/17; cited 2019/04/05]. Available from: https://helsedirektoratet.no/retningslinjer/nasjonalt-handlingsprogram-med-retningslinjer-for-diagnostikk-behandling-og-oppfolging-av-pasienter-med-brystkreft

[2] Badve SS, Beitsch PD, Bose S, et al. Part XI, Breast In: Amin MB, Edge S, Greene F, et al. editors. AJCC Cancer Staging Manual. 8th ed. Heidelberg, (Germany): Springer International Publishing; 2017.

[3] Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol. 2013;24(9):2206-23.

Table S2: Distribution of histopathologic and clinico-pathologic tumour characteristics (proportions with 95% confidence intervals, CIs[a], unless otherwise specified) for true, minimal signs, and missed screen-detected and interval breast cancers based on 40 imputations

	Screen-detected breast cancers			Interval breast cancers		
	True	Minimal signs	Missed	True	Minimal signs	Missed
Tumour characteristic	**n = 457**	**n = 336**	**n = 229**	**n = 310**	**n = 254**	**n = 224**
Histologic grade						
1	23.1 (19.2, 27.0)	33.0 (28.0, 38.1)[c]	32.5 (26.4, 38.6)	8.5 (5.4, 11.7)	17.2 (12.5, 21.8)	13.8 (9.2, 18.4)
2	47.2 (42.6, 51.8)	52.1 (46.7, 57.4)[c]	53.9 (47.4, 60.4)	44.8 (39.2, 50.4)	46.3 (40.1, 52.5)	50.1 (43.4, 56.7)
3	29.7 (25.5, 33.9)	14.9 (11.1, 18.7)[c]	13.6 (9.2, 18.1)	46.6 (41.0, 52.2)	36.6 (30.6, 42.5)	36.1 (29.8, 42.5)
Lymph node status						
Negative	78.1 (74.3, 81.9)	80.8 (76.6, 85.1)	80.5 (75.3, 85.7)	53.8 (48.1, 59.4)	61.0 (55.0, 67.0)	57.6 (51.1, 64.2)
Subtype						
Luminal A-like	34.3 (29.6, 39.0)	43.1 (37.3, 48.9)	43.9 (36.9, 50.8)	17.6 (12.8, 22.4)	26.8 (20.4, 33.1)	31.2 (24.5, 37.9)
Luminal B-like	51.2 (46.3, 56.2)	51.5 (45.6, 57.4)	51.3 (44.3, 58.2)	57.3 (51.3, 63.3)	55.3 (48.2, 62.3)	49.9 (42.7, 57.1)
HER2+[b] (non-luminal)	4.3 (2.3, 6.3)	3.1 (1.2, 5.0)	2.0 (0.1, 3.8)	8.1 (5.0, 11.2)	4.3 (1.7, 7.0)	8.9 (5.1, 12.8)
Triple negative (ductal)	10.2 (7.3, 13.0)	2.3 (0.6, 3.9)	2.9 (0.7, 5.1)	17.1 (12.8, 21.3)	13.6 (9.3, 17.9)	10.0 (5.9, 14.0)

[a] *Calculated using the Wilson score interval*
[b] *Human epidermal growth factor receptor 2 positive*
c These proportions are the same as in Table 2 because there was no missing information about histopathologic grade among women with minimal signs screen-detected breast cancer

Table S3: Hazard ratios (HRs) with 95% confidence intervals (CIs) for death due to any cause among women diagnosed with screen-detected and interval breast cancers included in the original analysis in the main text (Table 3) and in a sensitivity analysis that additionally included those that without tumour diameter information

	Multivariable[a,b] results from original analysis (*excluding* women without tumour diameter information)		**Multivariable[a] results from sensitivity analysis (*including* women with without tumour diameter information)**	
	HR	**95% CI**	**HR**	**95% CI**
Screen-detected breast cancers				
True	1.00	-	1.00	-
Minimal signs	1.05	(0.48, 2.31)	1.09	(0.55, 2.15)
Missed	1.28	(0.53, 3.07)	1.15	(0.54, 2.46)
Interval breast cancers, overall				
True	1.00	-	1.00	-
Minimal signs	0.76	(0.42, 1.36)	0.90	(0.56, 1.45)
Missed	1.23	(0.71, 2.14)	1.40	(0.88, 2.22)
Interval breast cancers, first 3 years				
True	1.00	-	1.00	-
Minimal signs	0.23	(0.07, 0.78)	0.54	(0.27, 1.10)
Missed	0.83	(0.35, 1.96)	0.97	(0.51, 1.86)
Interval breast cancers, after 3 years				
True	1.00	-	1.00	-
Minimal signs	1.46	(0.70, 3.05)	1.47	(0.74, 2.91)
Missed	1.76	(0.83, 3.72)	2.13	(1.07, 4.24)

[a] Model for screen-detected cancer adjusted for age at diagnosis, and grade, and subtype. Models for interval cancer adjusted for age at diagnosis and histopathologic grade, and stratified by subtype

[b] Models for screen-detected and interval cancer additionally adjusted for histopathologic tumour diameter