Journal of Clinical Epidemiology 148 (2022) 54−64

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Representativeness of trial participants: linking the EORTC boost-no boost trial to the Netherlands cancer registry

Anouk Neven[a,b,1], Marissa C. van Maaren[c,d,1,*], Kay Schreuder[c], Ries Kranse[c], Henk Struikmans[e], Philip M. Poortmans[f,g], Harry Bartelink[h], Laurence Collette[a,i], Lifang Liu[a], Sabine Siesling[c,d]

[a]*European Organisation for Research and Treatment of Cancer (EORTC), Brussels, Belgium*
[b]*Luxembourg Institute of Health, Competence Center for Methodology and Statistics, Strassen, Luxembourg*
[c]*Department of Research and Development, Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands*
[d]*Department of Health Technology and Services Research, Technical Medical Centre, University of Twente, Enschede, The Netherlands*
[e]*Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands*
[f]*Department of Radiation Oncology, Iridium Netwerk, Wilrijk-Antwerp, Belgium*
[g]*Faculty of Medicine and Health Sciences, University of Antwerp, Wilrijk-Antwerp, Belgium*
[h]*Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands*
[i]*International Drug Development Institute, Ottignies-Louvain-la-Neuve, Belgium*

Accepted 12 April 2022; Published online 15 April 2022

Abstract

Background and Objectives: To evaluate the representativeness of Dutch patients participating in the European Organization for Research and Treatment of Cancer EORTC boost-no-boost trial to the target breast cancer patient population.

Methods: All female breast cancer patients diagnosed between 1989 and 1996, aged ≤70 years, treated with breast-conserving surgery and radiation therapy, were selected from the Netherlands Cancer Registry (NCR) and linked to the EORTC trial database. Baseline characteristics were compared between trial and non-trial participants, for the Dutch population and according to seven participating institutions. Kaplan-Meier curves and multivariable Cox regression were used to explore potential heterogeneity in overall survival between low, medium and high-volume institutes.

Results: Overall, 20,880 patients were identified from the NCR: 2,445 of 2,602 (94%) trial participants could be linked, and 18,435 were treated outside the trial. Trial participants had similar age, morphology, topography, laterality and socioeconomic status as non-trial participants, but more often stage I (62.7% vs. 56.4%) tumours and less often adjuvant treatment (22.9% vs. 26.5%). Crude 20-year survival ranged from 52.5% to 57.4%, without significant differences in multivariable analyses.

Conclusion: This case study showed that participants in the boost-no-boost trial well represented the Dutch target population. Data linkage comes with challenges, but can close the gap between research and clinical practice.

Keywords: Data linkage; Boost radiation therapy; Breast cancer; Trial; Registry; Representativeness

1. Introduction

Trial populations may not reflect the entire target population, due to its strict inclusion criteria [1,2]. It is therefore important to analyze the generalizability of trial findings to daily practice. To achieve this, a linkage between trial and real-world registry data can be performed. Data linkages are described to be powerful tools for research, however, they are often time-consuming due to absence of or prohibition to use directly identifying variables [3] and data pre-processing [4]. Moreover, it comes with several challenges due to the possibility of

Declaration of Interests: The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Prof. Dr. Henk Struikmans, Prof. Dr. Philip Poortmans, Prof. Dr. Harry Bartelink and Laurence Colete were all involved in the EORTC boost-no-boost trial (Lancet Oncol 2015; 16: 47−56).

[1] These authors contributed equally to this work.

[*] Corresponding author. Netherlands Comprehensive Cancer Organisation (IKNL), P.O. Box 19079, 3501 DB Utrecht, The Netherlands. Tel.: +31-88-2346000.

E-mail address: m.vanmaaren@iknl.nl (M.C. van Maaren).

https://doi.org/10.1016/j.jclinepi.2022.04.014
0895-4356/

What is new?

Key findings

- Using a linkage between the EORTC boost-no boost trial database and the Netherlands Cancer Registry, we showed that patients included in the trial well represented the Dutch target population, including women diagnosed with early-stage breast cancer who received breast-conserving surgery and radiation therapy.
- Twenty-year survival rates did not differ between low, medium and high volume institutes, in which volume was based on two definitions: 1) based on the total number of patients treated in the RT dinstituted and 2) the number of included trial participants.

What this adds to what was known?

- This case study illustrates the opportunities for research into representativeness of trial results to the target population and shows its challenges.

What is the implication and what should change now?

- The described linkage procedure gives insight in challenges of linking data from different sources, and highlights research possibilities for collaborations between clinical research organisation and cancer registries.

incomplete linkage or mismatching [3,5,6]. Here, we perform a data linkage on an illustrative case study using European Organization for Research and Treatment of Cancer EORTC boost-no-boost trial data.

The EORTC boost-no-boost multinational study (EORTC trial 22,881-10,882, more information at http://clinicaltrials.gov/ct/show/NCT02295033) randomized women with early-stage breast cancer who received breast-conserving surgery (BCS) and radiation therapy (RT) between boost irradiation and no boost between 1989 and 1996. The long-term follow-up revealed a 4.4% reduced 20-year local recurrence risk in the boost arm, however with a negative impact on cosmetic results [7,8]. Overall survival (OS) was not significantly different between the groups [8]. The Dutch treatment guidelines [9] recommend, mostly based on the boost-no boost trial, that the benefit of a boost should be evaluated and weighted based on age, comorbidity and the chance of reduced cosmetic results [10]. After implementation of these guidelines in 2012, the variability in use of boost irradiation between institutes decreased but remained significant [11].

Here, we aimed to gain experience in assessing extrapolation of trial results to the daily breast cancer population. We therefore examined the representativeness of the Dutch EORTC boost-no-boost trial participants to the contemporaneous Dutch target population. Furthermore, we evaluated the process of linking two databases from different sources and described its challenges and opportunities.

2. Methods

2.1. Study design

In this historic cohort study, two databases were linked: the Netherlands Cancer Registry (NCR) and the EORTC boost-no-boost trial [8] database. The NCR is hosted by the Netherlands Comprehensive Cancer Organization (IKNL) where trained and dedicated data managers register data on patient-, tumor- and treatment-related characteristics of all newly diagnosed malignancies from 1989 onwards. Data are extracted directly from patient records. The main source of notification is the nationwide network and registry of histo- and cytopathology in the Netherlands (PALGA). The NCR is described to be complete for over 96% [12]. Vital status is obtained through regular linkage with the Municipal Personal Records Database.

The current analysis focuses first on the general Dutch population (treated in any Dutch institute) and then on patients treated in any of the seven Dutch institutes that included patients into the trial: The Netherlands Cancer Institute in Amsterdam, Leiden University Medical Center, Institute Verbeeten in Tilburg, Radboud University Medical Center Nijmegen, Erasmus University Medical Center Rotterdam, University Medical Center Utrecht and Maastricht Radiation Oncology Institute. All data analyzed in this study was extracted from the NCR, the EORTC trial data were solely used for identification of trial and non-trial participants.

2.2. Data linkage procedure and validation

To match the eligibility criteria of the EORTC trial, all female breast cancer patients diagnosed between 1989 and 1996, aged ≤70 years and treated with BCS and RT were selected from the NCR. The NCR did not record whether or not a patient had been included in a trial, and the EORTC database neither includes the exact date of diagnosis nor full patient name (privacy-related).

First, we aimed to identify the trial participants in the NCR and created a binary variable to indicate trial participation. The first linkage attempt was based on matching dates of birth and start of RT. The percentage of complete matches was poor (28%) primarily due to the frequency of breast cancer (many patients with same date of birth resulting in one-to-many matches) and the lack of information on date of RT in the NCR (in many instances information on

RT was present in text fields with the exact date missing in the electronic files). A second linkage attempt was based on date of birth and patient codes (first letter of first name followed by first two letters of last name given at birth). This second linkage reached over 90% complete matches.

Second, we tried to identify in which Dutch EORTC RT institute each included patient was irradiated, to analyze variation between institutes. As the NCR only fully collected information on treating RT institutes from 1990 on, we generated list of hospitals which were referring patients to any of the seven RT institutes. This proxy variable allowed us to estimate the RT institute in case it was unknown.

The linkage was validated by comparing the date of RT and surgery in both databases and plotting the date of diagnosis (NCR) against the date of randomization (EORTC), which was deemed satisfying. The actual linkages were executed by IKNL (RK) in close collaboration with the EORTC (AN). The linkage procedure was approved by the local privacy committee of the NCR (request number K15.183).

2.3. Statistical analysis

Patient-, tumor- and treatment characteristics at baseline were analyzed according to trial participation in the general Dutch population and then among patients treated in any of the seven EORTC Dutch RT institutes. Trial and non-trial participants were compared using the Chi-squared test, two-tailed Cochran-Armitage trend test or two-tailed t-test for binary, ordinal and continuous variables, respectively. Patients with missing or unknown values were excluded from the P-value computation. Due to the large dataset, a nominal significance level of 0.001 was chosen. Given the exploratory nature of this study, no further adjustment for multiplicity has been performed. Sensitivity analyses restricted to patients diagnosed between 1991 and 1995 (period during which the EORTC trial was fully recruiting) were conducted.

Next, to investigate potential heterogeneity between institute volumes in OS rates, each EORTC Dutch RT institute was categorized in low, medium or high volume according to two definitions determined a priori. First, 'volume' was defined based on the total number of patients treated in the RT institute, with low volume being defined as $\leq$1,000 patients, medium as 1,001−1,500 patients, and high as $>$1,500 patients. Second, 'volume' was defined based on the number of included trial participants, with low volume being defined as $<$100 patients, medium as 101−300 patients and high as $>$300 patients. OS was defined from date of diagnosis to date of death or last recorded date of being alive. Kaplan−Meier curves of OS rates by volume were displayed. The volume effect hazard ratio (HR) with its 95% confidence interval (CI) was estimated using univariable and multivariable Cox models including age, clinical stage, morphology, adjuvant systemic treatment and socioeconomic status (SES, based on the first four numbers of the postal code at diagnosis, extracted from the Netherlands Institute for Social Research, classified as low, medium and high). The proportional hazard (PH) assumption was checked. Whereas the PH assumption was not rejected for the volume effect (independent of the definition), evidence for non-PH was observed for some covariates. Therefore, the multivariable Cox model was stratified for age and clinical stage and adjusted for adjuvant systemic treatment, SES and morphology assuming a piece-wise constant hazard rate in 5-year time intervals for the latter co-variate.

We assessed homogeneity of the results across RT institutes using a forest plot and conducted an interaction test between RT institute and trial participation indicator in a Cox model. All analyses were conducted in accordance with a statistical analysis plan defined prior to data linkage and analysis. Analyses were performed in the SAS version 9.4.

3. Results

3.1. Patient selection and linkage results

In total, 20,880 patients were selected from the NCR to perform the linkage with the EORTC database, which included 2,602 patients treated in any of the seven Dutch RT institutes.

The first linkage attempt, using dates of birth and RT as linkage variables, resulted in only 28% coverage (729 of 2,602 patients). The second linkage, using date of birth and patient codes as linkage variables, resulted in 94% complete matches (2,445 of 2,602 patients). Of the 20,880 patients selected from the NCR 2,455 could eventually be linked to the EORTC trial, whereas 18,435 patients were identified as treated in the general clinical setting (no other clinical studies were open for accrual at that time) (Table 1).

3.2. Representativeness of trial participants to the general Dutch population

Baseline characteristics of the included trial and non-trial population are shown in Table 2. All variables were

Table 1. Trial participation and EORTC Dutch RT institutes

Number of patients	Treated in any of the seven EORTC Dutch RT institutes		
EORTC trial participation	Yes	No	Total
Yes	2,153	292	2,445[a]
No	7,522	10,913	18,435
Total	9,675	11,205	20,880

Abbreviations: EORTC, European Organization for Research and Treatment of Cancer; RT, radiation therapy.

All the information is based on data reported in the NCR.

[a] Linkage rate NCR and EORTC trial database: 94% (2,445/2,602).

Table 2. Patient-, tumor- and treatment-related characteristics at baseline in the general Dutch population

	Patients			
	Non-trial population	Trial population		Total number of patients
	(*N* = 18,435)	(*N* = 2,445)		(*N* = 20,880)
Baseline characteristics	***N* (%)**	***N* (%)**	***P*-value**	***N* (%)**
Age at diagnosis (years)			0.0070[c]	
N observed	18,435	2,445		20,880
Median	53	54		53
Range	20–70	22–70		20–70
Interquartile range	45–62	46–62		46–62
Morphology			0.0073[a]	
N observed	18,435 (100.0)	2,445 (100.0)		20,880 (100.0)
Invasive ductal carcinoma	14,492 (78.6)	1,903 (77.8)		16,395 (78.5)
Invasive lobular carcinoma	1,613 (8.7)	230 (9.4)		1,843 (8.8)
Mixed invasive pattern	591 (3.2)	75 (3.1)		666 (3.2)
Tubular carcinoma	775 (4.2)	130 (5.3)		905 (4.3)
Medullary carcinoma	364 (2.0)	54 (2.2)		418 (2.0)
Colloid/Mucinous carcinoma	315 (1.7)	33 (1.3)		348 (1.7)
Other	285 (1.5)	20 (0.8)		305 (1.5)
Topography			0.2538[a]	
N observed	17,973 (97.5)	2,394 (97.9)		20,367 (97.5)
Nipple	79 (0.4)	14 (0.6)		93 (0.4)
Central portion	972 (5.4)	110 (4.6)		1,082 (5.3)
Upper medial quadrant	2,744 (15.3)	350 (14.6)		3,094 (15.2)
Lower medial quadrant	1,253 (7.0)	185 (7.7)		1,438 (7.1)
Upper lateral quadrant	7,766 (43.2)	1,066 (44.5)		8,832 (43.4)
Lower lateral quadrant	1,536 (8.5)	205 (8.6)		1,741 (8.5)
Axillary tail of breast	253 (1.4)	25 (1.0)		278 (1.4)
Overlapping	3,370 (18.8)	439 (18.3)		3,809 (18.7)
Laterality			0.7184[a]	
N observed	18,423 (99.9)	2,444 (100.0)		20,867 (99.9)
Left	9,434 (51.2)	1,261 (51.6)		10,695 (51.3)
Right	8,989 (48.8)	1,183 (48.4)		10,172 (48.7)
Clinical stage			<0.001[b]	
N observed	18,280 (99.2)	2,426 (99.2)		20,706 (99.2)
Stage I	10,302 (56.4)	1,520 (62.7)		11,822 (57.1)
Stage IIA	5,929 (32.4)	716 (29.5)		6,645 (32.1)
Stage IIB	1,871 (10.2)	187 (7.7)		2,058 (9.9)
Stage IIIA	55 (0.3)	1 (0.0)		56 (0.3)
Stage IIIB	98 (0.5)	2 (0.1)		100 (0.5)
Stage IV	25 (0.1)	-		25 (0.1)
Grade			<0.001[b]	
N observed	8,211 (44.5)	803 (32.8)		9,014 (43.2)
Grade 1	1,042 (12.7)	147 (18.3)		1,189 (13.2)
Grade 2	3,203 (39.0)	299 (37.2)		3,502 (38.9)
Grade 3	3,897 (47.5)	351 (43.7)		4,248 (47.1)
Grade 4	69 (0.8)	6 (0.7)		75 (0.8)
Adjuvant systemic treatment			<0.001[a]	
N observed	18,435 (100.0)	2,445 (100.0)		20,880 (100.0)
No	13,548 (73.5)	1,886 (77.1)		15,434 (73.9)
Yes	4,887 (26.5)	559 (22.9)		5,446 (26.1)

(*Continued*)

Table 2. Continued

	Patients			
	Non-trial population	Trial population		Total number of patients
	(*N* = 18,435)	(*N* = 2,445)		(*N* = 20,880)
Baseline characteristics	*N* (%)	*N* (%)	*P*-value	*N* (%)
Socioeconomic status			0.0176[b]	
N observed	18,296 (99.2)	2,437 (99.7)		20,733 (99.3)
Low	4,979 (27.2)	546 (22.4)		5,525 (26.6)
Medium	7,235 (39.5)	1,102 (45.2)		8,337 (40.2)
High	6,082 (33.2)	789 (32.4)		6,871 (33.1)

Patients with missing/unknown values are excluded from the computation of the *P*-value. A nominal significance level of 0.001 has been chosen due to the large dataset.

[a] = Chi-squared test.

[b] = Cochran-Armitage trend test (two-tailed).

[c] = two-tailed *t*-test.

complete for over 97% of the patients, except for grade which was available in less than 50% of the patients. Overall, trial participants were of similar age, had a similar distribution of morphology, topography, laterality and SES compared to non-trial participants. However, trial participants more often presented with stage I or grade one disease than non-trial participants (62.7% vs. 56.4% and 18.3% vs. 12.7%, respectively). Trial participants less often received adjuvant systemic treatment compared to non-trial participants (22.9% vs. 26.5%). Sensitivity analyses restricted to patients diagnosed between 1991 and 1995 led to the same conclusions (data not shown).

3.3. Representativeness of trial participants to the Dutch population treated in any of the seven Dutch EORTC RT institutes

In total, 9,675 (46.3%) patients in the NCR were identified to be treated in any of the seven Dutch EORTC RT institutes (Table 1). Out of 2,445 patients linked with the EORTC database, 292 (11.9%) could not be linked to a specific RT institute, leaving 2,153 Dutch trial patients linked to the treating institute. Baseline characteristics in this subset, similar to the entire population, are shown in Supplementary Table 1. Patients had a median age of 53 years at diagnosis. Most patients had invasive ductal carcinoma (77.3%). More than half of the patients (58.8%) had a stage I disease and 24.2% received adjuvant systemic therapy. Distributions among institutes are displayed in Figure 1. Overall, the baseline characteristics were similar among the RT institutes, except for institute 7 in which lesser patients had invasive ductal carcinoma (62%), less patients had a stage I disease (49%) and more patients were treated with adjuvant systemic treatment (32%). At institute 4, the median age at diagnosis was slightly higher (55 years). Most notable differences among the RT institutes were observed for SES: whereas at institutes 4 and 7, less than 20% of the patients had a high SES, a high SES was reported in around half of the patients at institutes 5 and 6.

Depending on the RT institute, the percentage of the trial population varied between 1.9% (institute 3) and 47.7% (institute 2). Within the RT institutes, baseline characteristics were similar for the trial and non-trial population with some exceptions: at institute 2, trial participants were slightly older than non-trial participants (median age 55 vs. 52 years). At institutes 4 and 7, less trial participants were treated with adjuvant systemic therapy than non-trial participants (17% vs. 27% at institute 4 and 23% vs. 36% at institute 7). At all institutes, except for institute 3, stage I disease was more frequent with institute 7 showing the largest difference (61% vs. 42% in the trial vs. non-trial population). Finally, at site three more patients in the trial population were of low SES compared to the non-trial population, who more often had medium SES. However, any difference observed at institute 3 should be cautiously interpreted due to few trial participants (Fig. 1).

3.4. Volume analysis

In total, 5,431 (56.1%) patients treated in one of the participating RT institutes had deceased. Median OS was 22.6 years and was similar for both the trial and non-trial population (Supplementary Fig. 1). Table 3 displays the classification of the RT institutes by volume definition. In the volume analysis based on total number of patients treated, no significant volume effect was observed in univariable (*P*-value = 0.26) or multivariable analyses (*P*-value = 0.45) (Table 4A, Fig. 2A). The crude 20-year OS rate was 56.1% (95% CI: 54.7%–57.4%) in high, 54.7% (95% CI: 52.9%–56.6%) in medium and 57.4% (95% CI: 54.9%–59.9%) in low volume institutes.

In the univariable volume analysis based on number of included trial patients, patients treated in medium volume institutes had lower OS (crude 20-year OS 52.5% [95% CI: 50.3%–54.6%]) as compared to patients treated in low or in high-volume institutes (crude 20-year OS 57.3% (95% CI: 55.3%–59.2%) and 56.6% (95% CI: 55.3%–58.0%), respectively (P-value = 0.0002)). However, the difference disappeared after adjustment for confounding in multivariable analysis (P-value = 0.20). Exploratory heterogeneity analyses indicated no statistically significant interaction between RT institute and trial participation (Fig. 3).

4. Discussion

In this linkage case study, patients included in the EORTC boost-no-boost trial were largely representative for the contemporaneous general Dutch breast cancer population diagnosed between 1989 and 1996 and treated with BCS and RT. Among the baseline differences observed overall and in most RT institutes, early-stage disease was more frequent in the trial population than the non-trial population. This finding has also been described in literature [1]. This contributes to only 23% of the Dutch trial participants receiving adjuvant systemic treatment compared to 27% of the patients treated in daily practice.

The literature frequently reports underrepresentation of elderly patients in clinical trials [1,13]. Interestingly, we did not find that the trial population consisted of younger patients as compared to the general population—within the inclusion age limits of 18–70 years. The representativeness of trial participants in this specific study could be related to the nature of the intervention, which is in this case study treatment optimization. In drug development studies, selection criteria would be generally even stricter, probably leading to lower representativeness.

In multivariable analyses, no differences in long-term OS were observed in relation to the volume of patients treated in the institutions, independent of the volume definition. In our study, the 20-year crude OS rate was around 56%, which is similar to the 55% found in a Spanish registry study [14], but slightly lower than the approximately 60% and 59% reported in long-term follow-up clinical trials analyses [8,15]. This is likely to be explained by the slightly more complete follow-up in registries. Indeed, as reported in literature [16], registries often contain more up-to-date survival data compared to clinical trials in which obtaining long-term follow-up is challenging.

Linkage studies have been shown to be extremely useful in research [17,18], but may come with several challenges related to technical aspects and privacy issues [5]. Moreover, the utility of the linkage may be compromised by bias from linkage errors where records cannot be linked or are incorrectly linked together [3]. It has long been recognized that even small amounts of missed-matches and false-matches can lead to considerably biased results [19]. In our case study, the first linkage attempt using dates of birth and start of RT yielded coverage of only 28%. Date of death could not be used as a linkage variable, as discrepancies in survival data between the two databases have been noted. This is due to the EORTC dataset being frozen at the end of the data collection, whereas the NCR is a live database, and the challenges related to long-term follow-up data collection in clinical trials [17]. Given that incomplete linkage can result in biased estimates [3], the one-to-one matching of the de-identified clinical trial records had to be improved to get useful and reliable results. The linkage based on date of birth and patient code (first letter of first name followed by first two letters of last name given at birth) yielded 94% of complete matches.

While initially aiming for a 100% linkage coverage, challenges during the linkage procedure came to light and led us to consider >90% of complete matches as satisfying. First, before 1990 any RT not delivered within three months after diagnosis was not always recorded in the NCR and therefore incomplete. Second, patient codes in the EORTC database were not standardized and letters did not always match with initials of name and surname. Third, in the EORTC trial database, a few patients aged>70 years at randomization were identified. Since age at diagnosis was not available in the EORTC database, we could not identify whether randomization took place more than 1 year after diagnosis or if an exception for enrolment had been made. Fourth, in the EORTC trial, eligibility deviations were strictly based on baseline data, as per good standard practice in clinical trials. However, in case a patient had no metastases at baseline (no so eligibility deviation), but after randomization metastases were detected, in the NCR registry such a patient is coded as having metastases and thus excluded from the selected dataset. Consequently, this patient could not be linked. Last, but not least, with breast cancer being a frequent cancer there were a few instances of one-to-many matches.

It should be noted that the availability of the patient codes in the EORTC trial made the present study feasible. Nowadays, with increasingly strict data protection regulations, trials are no longer allowed to collect directly identifying variables. Researchers may thus face challenges when trying to link more recent trial data to registry data. In recent years, several guidelines and recommendations for data linkage—in case direct linkage variables are not available—have been published [20–22] which aid researchers to execute linkages with high accuracy and efficiency. Linkage studies of trial data to registry data not only can—as described in this study—close the gap between research and clinical practice, but they also have the potential to largely reduce costs associated with clinical trials [23].

It is worth noting the limitation in our project related with incomplete information on the treating RT institute in the NCR. We managed to overcome this challenge by using a proxy variable in which the treating RT institute was estimated based on referral patterns. Another limitation is the inclusion of patients diagnosed between 1989 and 1996. There have been drastic improvements in detection

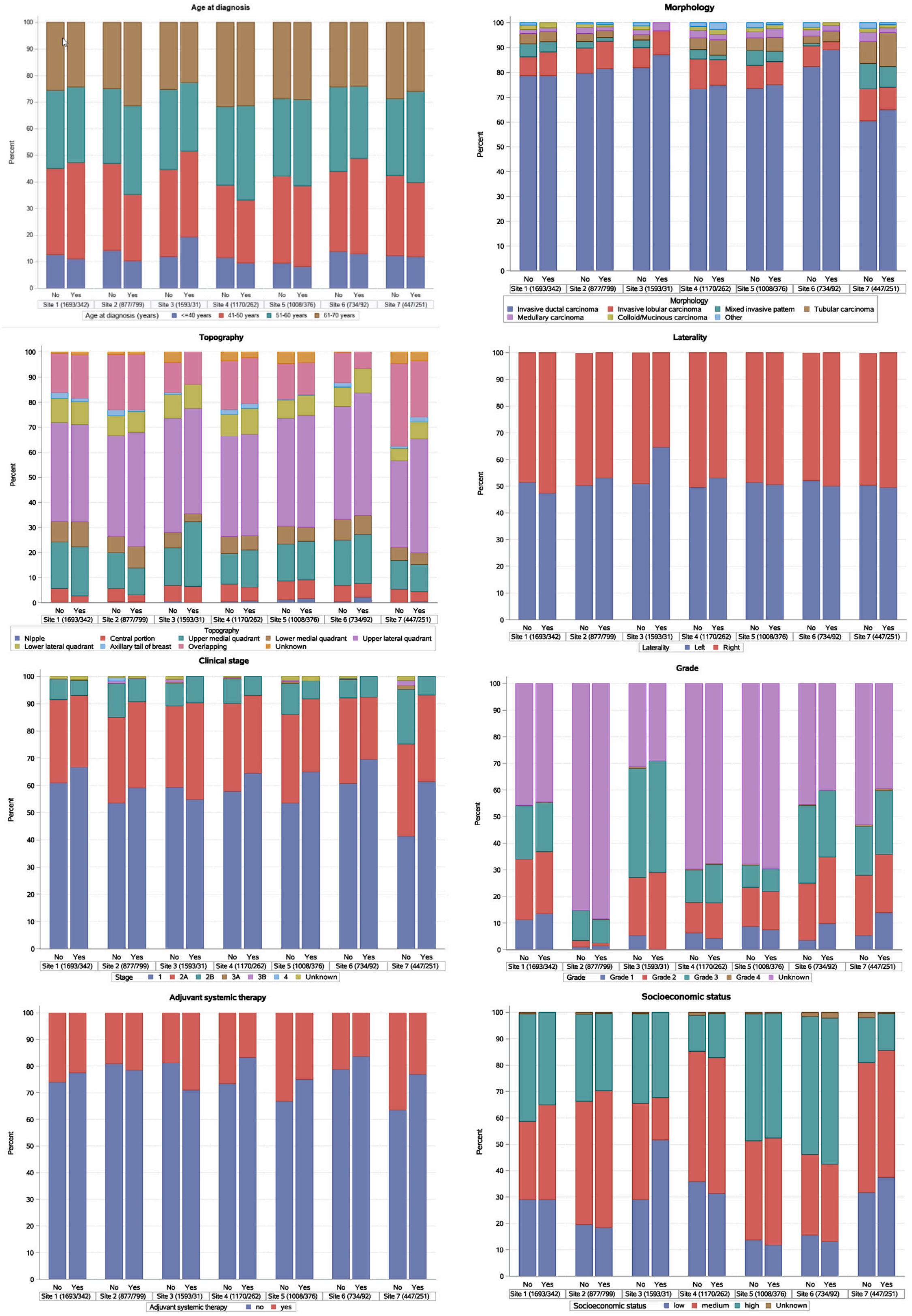

Fig. 1. Patient-, tumor- and treatment characteristics at baseline by Dutch radiation therapy institute (Site 1 – Site 7) and trial participation (No/Yes). For each site, two numbers are given in brackets: the first number displays the number of patients treated in the general clinical setting and the second the number of trial patients at the given site.

Table 3. Volume of each Dutch RT institute participating in the EORTC boost no-boost trial

EORTC Dutch RT institute	Total number of patients treated *N*	Total number of included trial patients *N*	Volume based on total number of patients treated[a]	Volume based on number of included trial patients[b]
Site 1	2,035	342	High	High
Site 2	1,676	799	High	High
Site 3	1,624	31	High	Low
Site 4	1,432	262	Medium	Medium
Site 5	1,384	376	Medium	High
Site 6	826	92	Low	Low
Site 7	698	251	Low	Medium

Abbreviations: EORTC, European Organisation for Research and Treatment of Cancer; RT, radiation therapy.

[a] low: ≤1,000 patients, medium: 1,001–1,500 patients, high: >1,500 patients.

[b] low: <100 trial patients, medium: 101–300 trial patients, high: >300 trial patients.

Table 4. Univariable and multivariable volume effect (A based on number of patients treated and B based on number of included trial participants) on overall survival

A) Volume based on total number of patients treated

			Univariable analysis		Multivariable analysis	
	Event/Total	OS estimates (95% CI)	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value
				0.2649[a]		0.4531[b]
High	2,982/5,326	10y: 76.5 (75.3–77.6%) 15y: 66.2 (64.9–67.5%) 20y: 56.1 (54.7–57.4%) 25y: 44.6 (43.2–46.0%)	Reference		Reference	
Medium	1,618/2,805	10y: 76.9 (75.3–78.5%) 15y: 66.1 (64.3–67.8%) 20y: 54.7 (52.9–56.6%) 25y: 43.5 (41.6–45.4%)	1.04 (0.97–1.10)		0.98 (0.92–1.04)	
Low	831/1,509	10y: 76.2 (73.9–78.2%) 15y: 66.7 (64.3–69.1%) 20y: 57.4 (54.9–59.9%) 25y: 46.4 (43.8–49.0%)	0.97 (0.90–1.05)		0.95 (0.88–1.03)	

B) Volume based on number of included trial patients

	Event/Total	OS estimates (95% CI)	Univariable analysis		Multivariable analysis	
	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value		
				0.0002[a]		0.2042[b]
High	2,846/5,091	10y: 77.1 (75.9–78.2%) 15y: 66.9 (65.6–68.2%) 20y: 56.6 (55.3–58.0%) 25y: 45.4 (44.0–46.8%)	Reference		Reference	
Medium	1,268/2,113	10y: 74.5 (72.6–76.3%) 15y: 63.6 (61.5–65.7%) 20y: 52.5 (50.3–54.6%) 25y: 40.9 (38.7–43.1%)	1.13 (1.06–1.20)		1.06 (0.99–1.13)	
Low	1,317/2,436	10y: 77.2 (75.5–78.9%) 15y: 67.1 (65.2–69.0%) 20y: 57.3 (55.3–59.2%) 25y: 46.1 (44.0–48.2%)	0.97 (0.91–1.04)		1.00 (0.93–1.06)	

Abbreviations: OS, overall survival; HR, hazard ratio; CI, confidence interval; y, year.

Patients with missing follow-up time have been removed from the analysis ($N = 35$, 0.4%). The multivariable Cox model was stratified for age and clinical stage and adjusted for adjuvant systemic treatment, socioeconomic status and morphology assuming a piece-wise constant hazard rate in time intervals of 5 years for the latter covariate.

[a] Logrank test.

[b] Wald test.

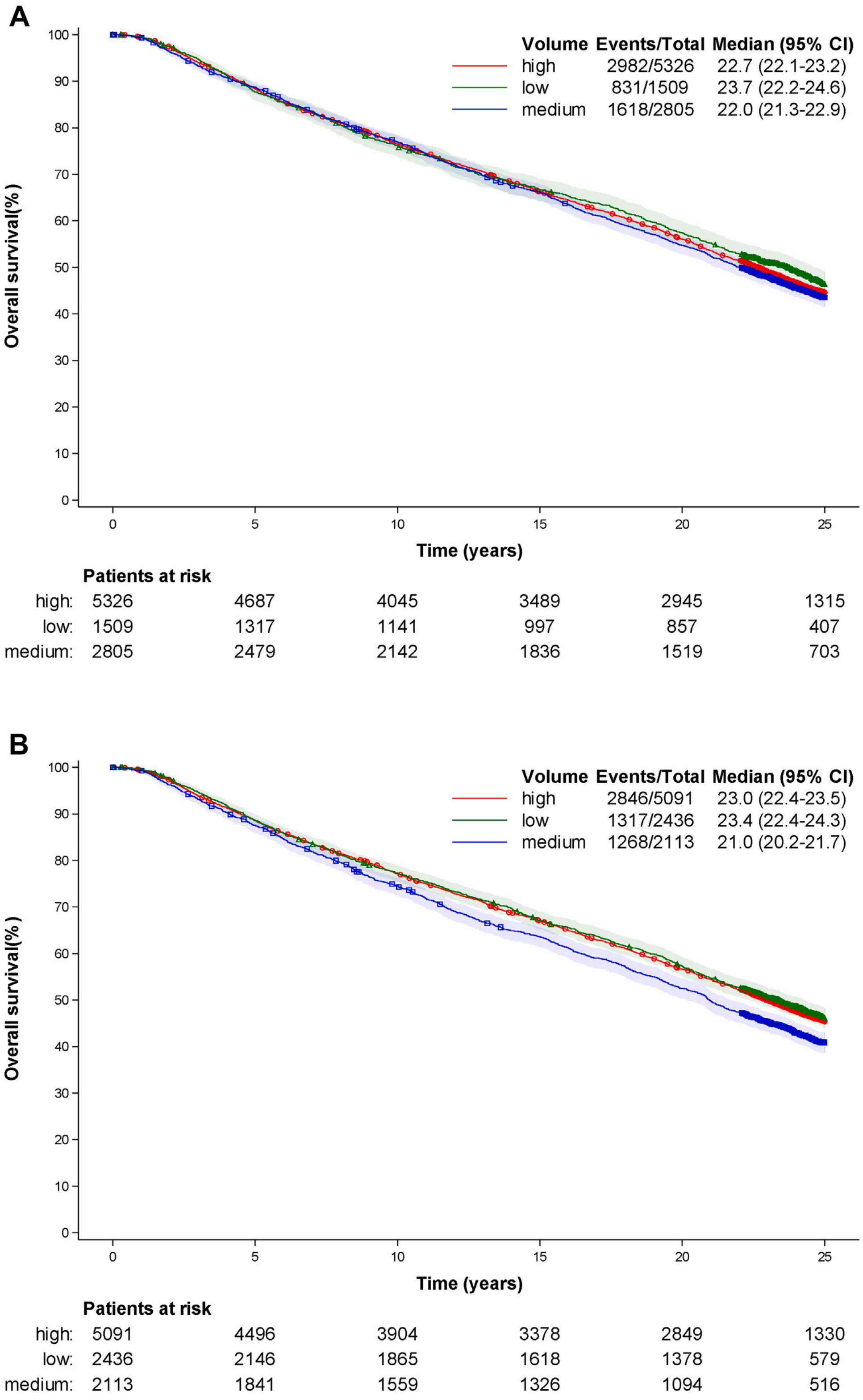

Fig. 2. (A) Overall survival by volume based on total number of patients treated (B) Overall survival by volume based on number of included trial patients.

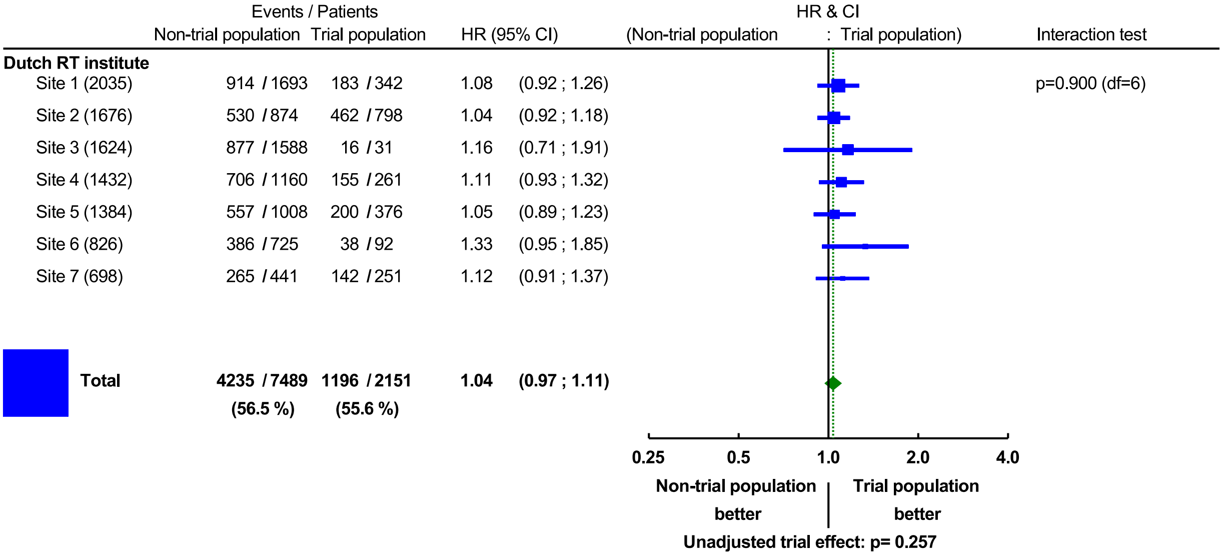

Fig. 3. Forest plot for overall survival. An unstratified univariable Cox model was used to estimate the hazard ratios in the trial population compared with the non-trial population. An unstratified Cox model including the trial group, the EORTC Dutch RT institute and the interaction term (e.g., trial participation × RT institute) was used to perform the interaction test and estimate the hazard ratios for the subgroups. *P*-values were yielded by the test of trial participation or by the test of interaction; for each, the Wald test was used. The sizes of the blue boxes are nonlinearly proportional to the numbers of events. The green diamond is centered on the overall hazard ratio (dashed line) and covers its 95% confidence interval. In the subgroup analyses, 95% confidence intervals (blue lines) are presented. *Abbreviations*: RT, radiation therapy; HR, hazard ratio; CI, confidence interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

methods and treatment strategies over time [24] and therefore trial results cannot be directly extrapolated to the contemporary breast cancer population. However, as the boost-no-boost trial is the most important existing evidence regarding boost treatment and there is still a variation in its application [11], our analyses are useful to clinical practice by showing that the trial population was largely representative to the general population at time of the trial. Taking into account the changing landscape since the trial started [25,26], results of the boost-no-boost trial can safely be applied to the general population. Nevertheless, benefits and harms of boost irradiation should be carefully considered and discussed with the patient in the light of contemporary outcomes.

In conclusion, the present case study, linking the EORTC boost-no-boost trial database with the NCR, illustrates the opportunities for research into representativeness of trial results to the target population. The described linkage procedure does not only give insight in challenges of linking data from different sources, but also highlights research possibilities for collaborations between clinical research organization and cancer registries.

CRediT authorship contribution statement

Anouk Neven: Formal analysis, Methodology, Software, Validation, Visualization, Writing – review & editing. **Marissa C. van Maaren:** Writing – original draft, Writing – review & editing. **Kay Schreuder:** Conceptualization, Resources, Writing – review & editing. **Ries Kranse:** Formal analysis, Resources, Validation, Writing – review & editing. **Henk Struikmans:** Writing – review & editing. **Philip M. Poortmans:** Writing – review & editing. **Harry Bartelink:** Writing – review & editing. **Laurence Collette:** Resources, Supervision, Writing – review & editing. **Lifang Liu:** Conceptualization, Resources, Supervision, Writing – review & editing. **Sabine Siesling:** Conceptualization, Resources, Supervision, Writing – review & editing.

Acknowledgments

We thank the Netherlands Cancer Registry for providing the data, as well as the registration clerks for their effort in gathering the data in the Netherlands Cancer Registry.

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.04.014.

References

[1] Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 2015;16:495.

[2] Dirix P, Wyld L, Paluch-Shimon S, Poortmans P. Time for more inclusive cancer trials. J Natl Compr Canc Netw 2020;18(10):1431–4.

[3] Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. Big Data Soc 2017;4(2). 2053951717745678.

[4] Playford CJ, Gayle V, Connelly R, Gray AJ. Administrative social science data: the challenge of reproducible research. Big Data Soc 2016;3(2). 2053951716684143.

[5] Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res 2010;10:346.

[6] Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. Hoboken, New Jersey, Verenigde Staten: John Wiley & Sons; 2015.

[7] Immink JM, Putter H, Bartelink H, Cardoso JS, Cardoso MJ, van der Hulst-Vijgen MHV, et al. Long-term cosmetic changes after breast-conserving treatment of patients with stage I-II breast cancer and included in the EORTC 'boost versus no boost' trial. Ann Oncol 2012;23:2591–8.

[8] Bartelink H, Maingon P, Poortmans P, Weltens C, Fourquet A, Jager J, et al. Whole-breast irradiation with or without a boost for patients treated with breast-conserving surgery for early breast cancer: 20-year follow-up of a randomised phase 3 trial. Lancet Oncol 2015; 16(1):47–56.

[9] Federatie Medisch Specialisten. Richtlijnendatabase. Borstkanker. Available at https://richtlijnendatabase.nl/richtlijn/borstkanker/invasief_carcinoom/lokale_behandeling_stadium_i_ii/borstsparende_therapie_mst.html. Accessed December 24, 2022.

[10] Buchholz TA, Somerfield MR, Griggs JJ, El-Eid S, Hammond ME, Lyman GH, et al. Margins for breast-conserving surgery with whole-breast irradiation in stage I and II invasive breast cancer: American Society of Clinical Oncology endorsement of the Society of Surgical Oncology/American Society for Radiation Oncology consensus guideline. J Clin Oncol 2014;32:1502–6.

[11] Schreuder K, Maduro JH, Spronk PER, Bijker N, Poortmans PMP, van Dalen T, et al. Variation in the use of boost irradiation in breast-conserving therapy in The Netherlands: the effect of a national guideline and cofounding factors. Clin Oncol (R Coll Radiol) 2019; 31(4):250–9.

[12] Schouten LJ, Höppener P, van den Brandt PA, Knottnerus JA, Jager JJ. Completeness of cancer registration in Limburg, The Netherlands. Int J Epidemiol 1993;22(3):369–76.

[13] Sorensen HT, Lash TL, Rothman KJ. Beyond randomized controlled trials: a critical comparison of trials with nonrandomized studies. Hepatology 2006;44(5):1075–82.

[14] Cleries R, Ameijide A, Buxo M, Martinez JM, Marcos-Gragera R, Vilardell ML, et al. Long-term crude probabilities of death among breast cancer patients by age and stage: a population-based survival study in Northeastern Spain (Girona-Tarragona 1985-2004). Clin Transl Oncol 2018;20(10):1252–60.

[15] Polgar C, Major T, Takacsi-Nagy Z, Fodor J. Breast-conserving surgery followed by partial or whole breast irradiation: twenty-year results of a phase 3 clinical study. Int J Radiat Oncol Biol Phys 2021; 109(4):998–1006.

[16] Liu L, Neven A, Giusti F, Maraldo MV, Meijnders P, Aurer I, et al. Using both clinical research and population-based cancer registry in long-term research- a case study using EORTC trials and the Dutch national cancer registry (IKNL). J Cancer Policy 2020;24: 100226.

[17] Hay AE, Mittmann N, Crump M, Cheung MC, Sleeth J, Needham J, et al. A Canadian prospective study of linkage of randomized clinical trial to cancer and mortality registry data. Curr Oncol 2021;28(2): 1153–60.

[18] Henry D, Fitzpatrick T. Liberating the data from clinical trials. BMJ 2015;351:h4601.

[19] Neter J, Maynes E, Ramanathan R. The effect of mismatching on the measurement of response error. J Am Stat Assoc 1965;60: 1005–27.

[20] Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for information about linking data sets. J Public Health (Oxf) 2018;40(1):191–8.

[21] March S, Andrich S, Drepper J, Horenkamp-Sonntag D, Icks A, Ihle P, et al. Good practice data linkage (GPD): a translation of the German version. Int J Environ Res Public Health 2020;17(21): 7852.

[22] Gliklich RE, Dreyer NA, Leavy MB, editors. Registries for evaluating patient outcomes: a user's guide. 3rd ed. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014.

[23] Ji C, Quinn T, Gavalova L, Lall R, Scomparin C, Horton J, et al. Feasibility of data linkage in the PARAMEDIC trial: a cluster randomised trial of mechanical chest compression in out-of-hospital cardiac arrest. BMJ Open 2018;8(7):e021519.

[24] Kaplan HG, Malmgren JA, Atwood MK, Calip GS. Effect of treatment and mammography detection on breast cancer survival over time: 1990-2007. Cancer 2015;121(15):2553–61.

[25] Smith IE, Okines AFC. De-escalating and escalating systemic therapy of early breast cancer. Breast 2017;34 Suppl 1:S5–9.

[26] Curigliano G, Burstein HJ, Winer EP, Gnant M, Dubsky P, Loibl S, et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen international expert consensus conference on the primary therapy of early breast cancer 2017. Ann Oncol 2017; 28:1700–12.