Journal of Cancer Research and Clinical Oncology (2019) 145:1485–1493
https://doi.org/10.1007/s00432-019-02919-x

ORIGINAL ARTICLE – CANCER RESEARCH

Different statistical techniques dealing with confounding in observational research: measuring the effect of breast-conserving therapy and mastectomy on survival

Marissa C. van Maaren[1,2] · Saskia le Cessie[3] · Luc J. A. Strobbe[4] · Catharina G. M. Groothuis-Oudshoorn[2] · Philip M. P. Poortmans[5] · Sabine Siesling[1,2]

Received: 4 April 2019 / Accepted: 15 April 2019 / Published online: 24 April 2019

Abstract

Purpose Propensity trimming, hierarchical modelling and instrumental variable (IV) analysis are statistical techniques dealing with confounding, cluster-related variation or confounding by severity. This study aimed to explain (dis)advantages of these techniques in estimating the effect of breast-conserving therapy (BCT) and mastectomy on 10-year distant metastasis-free survival (DMFS).

Methods All women diagnosed in 2005 with primary T1-2N0-1 breast cancer treated with BCT or mastectomy were selected from the Netherlands Cancer Registry. We used multivariable Cox regression to correct for confounding. Propensity trimming was used to create a more homogeneous population for which the treatment choice was not self-evident. Hospital of surgery was used as hierarchical level to handle hospital-related variation, and as IV to deal with unmeasured confounding.

Results Multivariable Cox regression showed higher 10-year DMFS for BCT than mastectomy [HR 0.70 (95% CI 0.60–82)]. Propensity trimming on the 10–90th and the 20–80th percentile of the propensity score distribution and hierarchical modelling showed similar HRs. IV analysis showed no significant difference between BCT and mastectomy.

Conclusion Unmeasured confounding is very difficult to eliminate in observational research. We cannot conclude that BCT or mastectomy has a causal relationship with 10-year DMFS. It is crucial to critically evaluate all model's assumptions, and to be careful in drawing firm conclusions.

Keywords Instrumental variable · Propensity scores · Hierarchical modelling · Breast-conserving therapy · Mastectomy · Breast cancer

Background

In the 80s, randomised controlled trials (RCTs) have shown equal survival for breast-conserving surgery with radiation therapy (BCT) and mastectomy (Fisher et al. 2002; Veronesi et al. 2002), thereby putting BCT for early breast cancer on the map. RCTs are conducted under very strict conditions, in which randomisation is performed to achieve equality between the studied treatment groups regarding the distribution of factors that may affect outcomes. This procedure ensures that measured effects are real treatment effects and not a result of residual confounding or confounding by severity, resulting in perfect internal validity (Giordano 2015). However, trial populations are, due to their strictly controlled environment, not always representative for the target population (Antman et al. 1985). Older patients are largely underrepresented in RCTs (Hutchins et al. 1999),

✉ Marissa C. van Maaren
m.vanmaaren@iknl.nl

[1] Department of Research, Netherlands Comprehensive Cancer Organisation, P.O. Box 19079, 3501 DB Utrecht, The Netherlands

[2] Department of Health Technology and Services Research, Faculty of Behavioural, Management and Social Sciences, Technical Medical Centre, University of Twente, Enschede, The Netherlands

[3] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

[4] Department of Surgical Oncology, Canisius Wilhelmina Hospital, Nijmegen, The Netherlands

[5] Department of Radiation Oncology, Institut Curie, Paris, France

trial participants have fewer comorbid conditions than the general population and tend to be more motivated (Chavez-MacGregor and Giordano 2016), thereby affecting the generalisability of the results. Due to limited external validity of RCTs, and taking account of the fact that the above-mentioned trials have been conducted more than 30 years ago, comparing BCT with mastectomy may lead to different results in the contemporary real-world population. Consequently, multiple observational studies using more recent data have been performed comparing outcomes of both types of surgery in daily practice (Hwang et al. 2013; van Maaren et al. 2016a, b; Chen et al. 2015; Hofvind et al. 2015; Onitilo et al. 2015; Hartmann-Johnsen et al. 2012; Agarwal et al. 2014; Fisher et al. 2015). Most of these studies showed that BCT led to increased survival outcomes compared to mastectomy. These studies have been subjected to a lot of criticism because of their observational nature. Although observational studies are more representative for the entire population, the choice for a specific surgery type is largely related not only to patient and tumour characteristics, such as age and tumour size, but also to patient preferences and surgeon recommendations (Morrow et al. 2009). Results of observational studies may, therefore, suffer from residual confounding or confounding by severity (Hershman and Wright 2012). In the past years, several statistical methods have been developed dealing with confounding or confounding by severity, including propensity scores (PS) (Brookhart et al. 2013; Sturmer et al. 2014) and instrumental variable (IV) analysis (Greenland 2000). Furthermore, there is a lot of variation in initial type of surgery between hospitals (Siesling et al. 2005; van Maaren et al. 2018), which may also partly explain survival differences following BCT and mastectomy, and can be solved by using hierarchical modelling (Aarts et al. 2015; Austin 2017).

Propensity trimming

Propensity trimming is a form of PS analysis (Sturmer et al. 2014). A PS is the conditional probability of assignment to a particular treatment given a patient's characteristics. It is a function of all measured confounding variables such that conditional on the PS, the distribution of confounding variables is equal in both treatment groups (Sturmer et al. 2014). PS analyses do not cover the problem of unmeasured confounding. However, by excluding patients in the nonoverlapping parts of the PS distribution, a more homogeneous group will be created allowing us to study a group that, based on measured confounders, is eligible for both treatment types. Further propensity trimming excludes patients who are present in the tails of the PS distribution (Sturmer et al. 2014). In this manner, patients with the highest likelihood of a certain treatment are excluded, resulting in an even more homogeneous population.

Hierarchical modelling

Hierarchically structured data are frequently present in several research disciplines. For instance, when analysing survival in patients treated with either BCT or mastectomy, all patients are clustered within hospitals. Hierarchical modelling allows us to take dependency of patients within hospitals into account, and corrects in this manner for cluster-related variation (Austin 2017).

Instrumental variable analysis

An IV is a variable that influences the probability of receiving a particular treatment, but is unrelated to patient characteristics or prognosis. In this way, an IV behaves similar to randomisation and may overcome the problem of unmeasured confounding in observational studies. Rather than comparing two treatment groups, IV analysis compares groups of patients on the level of the IV. For example, if hospital of surgery largely determines the type of treatment and is unrelated to patient characteristics or prognosis, the analysis compares patients on the level of their hospital (Boef et al. 2013). To imitate randomisation, the IV should meet the following three assumptions: (1) the IV is associated with treatment. (2) The IV has no effect on the outcome, except through its effect on the treatment itself. (3) The IV is independent of unmeasured confounding. These assumptions require very careful justification, as it is unverifiable whether these assumptions are really met (Dekkers 2011; Baiocchi et al. 2014).

Aim of the study

This study aimed to describe outcomes of different statistical techniques dealing with measured confounding, cluster-related variation and confounding by severity to estimate the relationship of BCT and mastectomy with 10-year distant metastasis-free survival (DMFS).

Methods

Patients and study design

For this population-based study, all women diagnosed in 2005 with primary invasive T1-2N0-1 breast cancer, treated with BCT or mastectomy in the Netherlands, were selected from the nationwide Netherlands Cancer Registry (NCR). Patient-, tumour-, and treatment-related characteristics were prospectively registered by trained data managers. Tumour topography, morphology and grade were coded according to the International Classification of Diseases for Oncology, seventh edition (Fritz et al. 2000). Stage was coded according to the tumour, node and metastasis (TNM) system of the International Union Against Cancer, sixth edition (Sobin 2002). Data on vital status and date of death were derived from the Municipal Personal Records database, completed until February 2017. Data on distant metastases were retrospectively gathered from patient files, and defined according to existing consensus-based definitions for recurrence classification (Moossdorff et al. 2014).

Statistical analysis

Patient-, tumour-, and treatment-related characteristics were summarized and compared between the treatment groups using the Chi-squared or Wilcoxon rank sum tests. To assess the effect of BCT compared to mastectomy on 10-year DMFS we first used the conventional multivariable Cox proportional hazard regression to estimate hazard ratios (HRs) with 95% confidence intervals (CIs), in which we corrected for potential confounding variables differing significantly between the treatment groups ($p < 0.2$). The variables ultimately corrected for were age (< 40, 40–49, 50–59, 60–69, 70–79 and ≥ 80 years), geographical region, sublocalisation of the tumour within the breast, histological subtype, differentiation grade, tumour size, number of positive lymph nodes, multifocality, HER2 status, use and type of adjuvant systemic therapy and axillary lymph node dissection. Results of the three statistical techniques were compared with the results of the Cox model.

For propensity trimming, we first calculated the PS. This was achieved by performing logistic regression with type of treatment as outcome (0 = mastectomy,1 = BCT) and the same characteristics as used in the conventional Cox regression as covariates. To verify its accuracy, we assessed the balance of all measured covariates between the treatment groups among quintiles of patients based on the PS distribution. The PS was considered to be accurate when no significant differences were observed between covariates and the treatment variable in each quintile. Trimming was performed by, first, excluding all patients in the nonoverlapping parts of the PS distribution. Like this, we restricted the analysis to observations within a PS that was common to both patients treated with BCT and mastectomy. Subsequently, we applied further trimming at different cut-off points ($0.1 < PS < 0.9$ and $0.2 < PS < 0.8$) to exclude patients present at the tails of the distribution, meaning that patients with the highest chance to be treated with either BCT or mastectomy were excluded. For the remaining patients, multivariable Cox regression was performed in which we applied inverse probability weighting (IPW) (Mansournia and Altman 2016), in which weights are based on each individual's probability of receiving a specific treatment given the PS. For continuous variables, the proportional hazards assumption was tested by plotting the scaled Schoenfeld residuals over time and checking these for consistency. For categorical variables, this assumption was tested by plotting the log of the − log survival function {log[− log($S(t)$)]} against the log of the survival time [log(t)]. When this plot showed linear graphs, the proportional hazards assumption was considered to be met. No deviations were found.

Hierarchical modelling was performed using a multilevel mixed-effects parametric survival model (-mestreg- command in STATA) assuming a Weibull distribution, in which we used hospital of surgery ($n = 87$) as hierarchical level. This model uses a likelihood ratio test to determine whether the hierarchical level significantly contributes to the model. As the Weibull distribution assumes a parametric form for the distribution of survival time, we tested this assumption by plotting the log of the − log survival function {log[− log(S(t))]} against the log of the survival time [log(t)]. No violations were found. In this model, we corrected for all other potential confounding factors as described above.

For IV analyses, pseudo-observations at 10-years for the survival function were generated using the -stpsurv- command in STATA (Kjaersgaard and Parner 2016), accounting for censored data. Subsequently, these pseudo-observations were used in a linear regression analysis (-ivregress-) using the generalised method of moments (GMM) estimator (Kjaersgaard and Parner 2016) and hospital of surgery as IV. This resulted in a coefficient indicating the difference in 10-year DFMS between the two types of surgery. Covariates significantly differing among the hospitals of surgery were added to the model. As there is no statistical test that judges the IV's validity, we evaluated it as follows. First, to get insight in the predictive ability of the chosen IV, we verified whether hospital of surgery was related to type of surgery by comparing the percentage of BCTs among the different hospitals (assumption 1). Assumption 2 and 3 could not be tested, but were carefully thought of. Therefore, we observed the distribution of all measured potential confounders in all hospitals, to get insight in the random assignment of the IV, and we evaluated the effect of hospital of surgery on 10-year DMFS.

For all different methodologies, we performed the analysis in the entire cohort and stratified for T and N stages, as both T and N stages are important prognostic factors for survival. A p value < 0.05 was considered as statistically significant. All analyses were executed in STATA version 14.1.

Ethics and consent to participate

This study was approved by the privacy committee of the Netherlands Cancer Registry.

Results

A total of 10,327 primary nonmetastatic invasive breast cancer patients diagnosed in 2005 and treated with surgery in the Netherlands were identified. Patients with T3-4N2-3 stage breast cancer ($n = 1950$), patients who received radiation therapy following mastectomy ($n = 258$), were treated with BCT without radiation therapy ($n = 74$), patients with unknown type of surgery ($n = 5$), with macroscopic residual disease ($n = 3$) or treated in an unknown hospital ($n = 1$) were excluded. The final population consisted of 8036 patients. Baseline characteristics, specified for type of surgery, are shown in Table 1.

Multivariable Cox regression

First, we used multivariable Cox regression to estimate 10-year DMFS. As shown in Table 2, 10-year DMFS was higher for BCT than mastectomy [HR 0.70 (95% CI 0.60–82)]. This result remained in T1N0 [HR 0.73 (95% CI 0.56–0.94)], T2N0 [HR 0.52 (95% CI 0.37–0.73)] and T2N1 stage [HR 0.71 (95% CI 0.50–0.99)]. In T1N1 stage, BCT was not statistically different from mastectomy [HR 0.84 (95% CI 0.53–1.34)].

Propensity trimming

The PS distribution for both types of surgery is shown in Fig. 1. The median PS (reflecting the probability of receiving a BCT) was 0.72 (interquartile range 0.48–0.84). The figure shows the different cut-offs on which propensity trimming was performed. No significant differences between the measured confounders and the treatment variable were found among the quintiles of the PS distribution (data not shown).

When using propensity trimming on the nonoverlapping parts of the PS distribution (> 0.009597 and < 0.9507796), ten patients were excluded from the analysis. After IPW weighting, similar results were found as in the multivariable Cox regression [HR 0.70 (95% CI 0.59–0.83)]. In T2N1 stage, the HR was similar to the HR using multivariable Cox regression, although it was nonsignificant [HR 0.71 (95% CI 0.49–1.02)]. When trimming on the 10th and 90th percentile, IPW weighting resulted in similar outcomes for each T and N stage as in the conventional Cox regression, except for T1N0 in which no significant difference between BCT and mastectomy was observed anymore [HR 0.86 (95% CI 0.65–1.13)]. Trimming on the 20th and 80th percentile gave comparable results (Table 2).

Hierarchical modelling

Using hospital of surgery as hierarchical level, BCT was associated with higher 10-year DMFS compared to mastectomy in the entire cohort [HR 0.71 (95% CI 0.61–0.83)]. This result was similar in T1N0 stage. Here, hospital of surgery as hierarchical level did not significantly contribute to the model ($p = 0.07$), but was included as the p value may be sensitive to sample size. In T1N1 stage, BCT was not significantly different from mastectomy (HR 0.88 (95% CI 0.56–1.39)]. Of note, hospital of surgery as level did not contribute significantly to the model ($p = 1.0$). In T2N0 and T2N1, BCT was significantly related to higher 10-year DMFS than mastectomy. Here, hospital of surgery as hierarchical level significantly contributed to the model (Table 2).

Instrumental variable analysis

We verified that hospital of surgery ($n = 87$) was related to treatment (assumption 1): percentages of BCTs varied from 11.8% to 90.6%. We observed lots of variation in patient-, tumour- and treatment-related characteristics, indicating a non-random assignment of the IV, and we observed significant differences among the hospitals in 10-year DMFS (data not shown). An IV analysis with hospital of surgery as IV and type of surgery as independent variable provided a R2 of almost zero, implying that hospital of surgery explains very little of the variability in type of surgery. To correct as good as possible for confounding, we additionally corrected for all measured confounders in subsequent IV analyses.

The overall analysis showed no significant difference between BCT and mastectomy on 10-year DMFS [− 0.05 95% CI (− 0.13 to 0.04]. However, after stratifying for T and N stage, a significant difference in favour of BCT on 10-year DMFS was found in T1N1 stage [− 0.13 (95% CI − 0.20 to − 0.07] (Table 2).

Discussion

This study evaluated several methods dealing with confounding, cluster-related variation and confounding by severity. We showed that the effect estimates obtained using multivariable Cox regression, propensity trimming and hierarchical modelling were quite similar. The 95% CIs were wider for propensity trimming, as this method excluded a number of patients, thereby lowering the power. The results of the IV analysis were not directly comparable with those from the other methods. No significant difference between BCT and mastectomy was observed in the entire cohort using this method. Since multivariable Cox regression, propensity trimming and hierarchical modelling do not solve the problem of unmeasured confounding, we cannot conclude that BCT has a causal relationship with 10-year DMFS. IV analysis may deal with unmeasured confounding, but in this study it is questionable whether the IV fulfills assumptions 2 and 3 and, therefore, no firm conclusions from this method can be drawn either.

Table 1 Patient-, tumour-, and treatment-related characteristics specified for type of surgery

Characteristics	Mastectomy (*n* = 2968)	Breast-conserving therapy (*n* = 5068)	*p* value
Age (years)			
< 40	162 (5.5)	264 (5.2)	**< 0.001**
40–49	498 (16.8)	994 (19.6)	
50–59	692 (23.3)	1593 (31.4)	
60–69	574 (19.3)	1385 (27.3)	
70–79	585 (19.7)	719 (14.2)	
≥ 80	457 (15.4)	113 (2.2)	
Hospital volume			
0–49	426 (14.4)	705 (13.9)	0.132
50–99	1747 (58.9)	2881 (56.9)	
100–149	499 (16.8)	937 (18.5)	
150 +	296 (10.0)	545 (10.8)	
Region			
A	366 (12.3)	736 (14.5)	**< 0.001**
B	309 (10.4)	453 (8.9)	
C	169 (5.7)	387 (7.6)	
D	497 (16.8)	1003 (19.8)	
E	308 (10.4)	524 (10.3)	
F	547 (18.4)	586 (11.6)	
G	366 (12.3)	743 (14.7)	
H	187 (6.3)	300 (5.9)	
I	219 (7.4)	336 (6.6)	
T stage			
T1	1594 (53.7)	3840 (75.8)	**< 0.001**
T2	1374 (46.3)	1228 (24.2)	
Mean tumour size in mm (IQR)	19 (13–26)	15 (11–20)	**< 0.001**
N stage			
N0	1936 (65.2)	3852 (76.0)	**< 0.001**
N1	1032 (34.8)	1216 (24.0)	
Mean number of positive lymph nodes (IQR)	0 (0–1)	0 (0–0)	**< 0.001**
Lateralisation			
Left	1537 (51.8)	2651 (52.3)	**< 0.001**
Right	1431 (48.2)	2417 (47.7)	
Sublocalisation			
Outer quadrants	1308 (44.1)	2562 (50.6)	**< 0.001**
Inner quadrants	496 (16.7)	1089 (21.5)	
Central parts	293 (9.9)	288 (5.7)	
Overlapping lesions	803 (27.1)	1053 (20.8)	
Unknown	68 (2.3)	76 (1.5)	
Histological tumour type			
Ductal	2231 (75.2)	4157 (82.0)	**< 0.001**
Lobular	400 (13.5)	407 (8.0)	
Mixed	136 (4.5)	157 (3.1)	
Other	193 (6.5)	333 (6.6)	
Grade			
1	541 (18.2)	1370 (27.0)	**< 0.001**
2	1340 (45.2)	2179 (43.0)	
3	925 (31.2)	1331 (26.3)	
Unknown	162 (5.5)	188 (3.7)	

Table 1 (continued)

Characteristics	Mastectomy ($n = 2968$)	Breast-conserving therapy ($n = 5068$)	p value
Multifocality			
No	2122 (71.5)	4653 (91.8)	**< 0.001**
Yes	767 (25.8)	317 (6.3)	
Unknown	79 (2.7)	98 (1.9)	
Hormonal receptor status			
Positive	1872 (66.4)	3442 (67.9)	0.230
Mixed	495 (16.7)	775 (15.3)	
Negative	460 (15.5)	774 (15.3)	
Unknown	41 (1.4)	77 (1.5)	
HER2 status			
Negative	1977 (66.6)	3618 (71.4)	**< 0.001**
Unclear	303 (10.2)	443 (8.7)	
Positive	400 (13.5)	560 (11.1)	
Unknown	288 (9.7)	447 (8.8)	
Adjuvant systemic therapy			
None	1222 (41.2)	2780 (54.9)	**< 0.001**
Endocrine therapy	832 (28.0)	801 (15.8)	
Chemotherapy	266 (9.0)	516 (10.2)	
Both	648 (21.8)	969 (19.1)	
Targeted therapy (trastuzumab)			
Yes	142 (4.8)	244 (4.8)	0.951
No	2826 (95.2)	4824 (95.2)	
Axillary lymph node dissection			
Yes	1703 (57.4)	1503 (29.7)	**< 0.001**
No	1265 (42.6)	3565 (70.3)	

p values indicated in bold are considered as statistically significant ($p < 0.05$)

IQR interquartile range, *HER2* human epidermal growth factor receptor 2

Here, we will summarizse results of the three statistical techniques. First, we used propensity trimming with IPW weighting, which deals in a different way with confounding when compared to conventional multivariable analysis. PS analysis corrects for all measured (potential) confounders and is a more appropriate method in case of few events (Braitman and Rosenbaum 2002). The additional propensity trimming resulted in a more homogeneous group of patients, by excluding patients with the highest likelihood of either one or the other treatment (Sturmer et al. 2014), ending up with a group in which both treatments are likely. Our results show that trimming results in slightly different estimates. However, the interpretation of the outcomes remained largely similar. We additionally performed all these analyses using propensity matching (Brookhart et al. 2013) instead of IPW, which gave similar results (data not shown). However, both multivariable Cox regression and propensity methods are not able to deal with unmeasured confounding. As we lack data on family history, comorbidities, contraindications for RT or personal circumstances that make one or the other treatment unlikely, residual confounding and confounding by severity is likely to be present, and the treatment effect estimates may still cause biased treatment effects.

When using hierarchical modelling of hospital of surgery to account for cluster-related variation, similar results as conventional multivariable regression were obtained.

The IV analyses were the most challenging. Finding a good IV is very difficult (Hernan and Robins 2006), as was also witnessed in our study. It cannot be verified whether all assumptions of the IV are met (Dekkers 2011). However, the best IV we could find in our dataset, hospital of surgery, was evaluated thoroughly. We found that a lot of measured potential confounders in our dataset were significantly different among the hospitals. Only after correcting for all casemix variables, the IV estimate may be valid (no association between unmeasured confounding and the IV). Our IV analyses showed no difference in 10-year DMFS in the entire cohort and in most subgroups, but in T1N1 stage BCT seemed to be related to better DMFS. However, since we cannot verify whether all assumptions are met, we are not able to draw definite conclusions. The surgeon's preference

Table 2 Hazard ratios for BCT versus mastectomy on 10-year OS and DMFS using different statistical techniques

Statistical technique	Subgroup analyses	Treatment	10-year DMFS		
			n	HR (95% CI)	*p* value
Multivariable Cox regression	Entire cohort	Mastectomy	2968	1	**< 0.001**
		BCT	5068	0.70 (0.60–0.82)	
	T1N0	Mastectomy	1137	1	**0.016**
		BCT	3052	0.73 (0.56–0.94)	
	T1N1	Mastectomy	457	1	0.464
		BCT	788	0.84 (0.53–1.34)	
	T2N0	Mastectomy	799	1	**< 0.001**
		BCT	800	0.52 (0.37–0.73)	
	T2N1	Mastectomy	575	1	**0.046**
		BCT	428	0.71 (0.50–0.99)	
Propensity trimming nonoverlapping parts excluded	Entire cohort	Mastectomy	2373	1	**< 0.001**
		BCT	4238	0.70 (0.59–0.83)	
	T1N0	Mastectomy	872	1	**0.023**
		BCT	2508	0.73 (0.56–0.96)	
	T1N1	Mastectomy	375	1	0.410
		BCT	672	0.81 (0.49–1.34)	
	T2N0	Mastectomy	656	1	**0.001**
		BCT	691	0.58 (0.41–0.81)	
	T2N1	Mastectomy	470	1	0.066
		BCT	367	0.71 (0.49–1.02)	
Propensity trimming 10–90th percentile	Entire cohort	Mastectomy	1728	1	**0.001**
		BCT	3559	0.75 (0.63–0.89)	
	T1N0	Mastectomy	690	1	0.279
		BCT	1931	0.86 (0.65–1.13)	
	T1N1	Mastectomy	309	1	0.726
		BCT	627	1.10 (0.66–1.83)	
	T2N0	Mastectomy	428	1	**0.005**
		BCT	660	0.61 (0.43–0.86)	
	T2N1	Mastectomy	301	1	**0.007**
		BCT	341	0.61 (0.43–0.88)	
Propensity trimming 20–80th percentile	Entire cohort	Mastectomy	1194	1	**0.003**
		BCT	2771	0.75 (0.62–0.91)	
	T1N0	Mastectomy	494	1	0.531
		BCT	1349	0.60 (0.66–1.24)	
	T1N1	Mastectomy	209	1	0.772
		BCT	544	0.92 (0.51–1.64)	
	T2N0	Mastectomy	296	1	**0.021**
		BCT	586	0.63 (0.43–0.93)	
	T2N1	Mastectomy	195	1	**0.007**
			292	0.58 (0.39–0.86)	
Hierarchical modelling using -mestreg-	Entire cohort	Mastectomy	21,968	1	**< 0.001**
		BCT	5068	0.71 (0.61–0.83)	
	T1N0	Mastectomy	1137	1	**0.009**
		BCT	3052	0.71 (0.55–0.92)*	
	T1N1	Mastectomy	457	1	0.587
		BCT	788	0.88 (0.56–1.39)*	
	T2N0	Mastectomy	799	1	**0.001**
		BCT	800	0.57 (0.41–0.81)	
	T2N1	Mastectomy	575	1	**0.047**
			428	0.71 (0.51–1.00)	

A *p* value indicated in bold is considered as statistically significant ($p < 0.05$)

OS overall survival, *DMFS* distant metastasis-free survival, *HR* hazard ratio, *CI* confidence interval, *BCT* breast-conserving surgery

Table 2 (continued)

*Hospital of surgery ($n = 87$) as level did not contribute significantly to the models

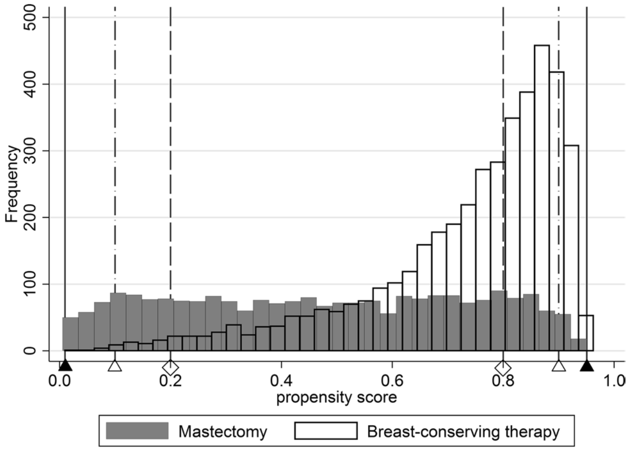

Fig. 1 Propensity score distribution per type of surgery. The dark triangle (and solid line) indicates exclusion of patients present in the nonoverlapping parts of the distribution. The open triangle (and dashed/dotted line) indicates trimming on 0.1 and 0.9 of the propensity score distribution. The open diamond (and dashed line) indicates trimming on 0.2 and 0.8 of the propensity score distribution

might have been a good alternative as IV. However, as we lacked this information, we could not execute these analyses.

Importantly, all methods used in this study depend on untestable assumptions [no unmeasured confounding in case of PS analysis and multivariable Cox regression, a direct effect of the IV on the outcome (assumption 2) and no association between unmeasured confounding and the IV in IV analysis (assumption 3)]. Therefore, we have to remain careful in making statements in comparing BCT to mastectomy.

Several other studies have compared statistical techniques and its effect on treatment effect estimates (Tsuchiya et al. 2016; Federspiel et al. 2016; Stukel et al. 2007). They all showed approximately similar effect estimates using all methods and conclude that results have to be interpreted with care. A strong IV might closely resemble randomisation, but the unverifiable assumptions need very careful justification. It is of crucial importance to realise that confounding by severity is present in many observational studies and that statistical methods may not be able to account for this. This does, however, not mean that results of observational studies are not informative. As RCTs do not provide all answers as well, as they are often carried out in selected populations in a given time frame and in a limited number of patients from selected centres. Moreover, the Hawthorne effect—changes in behaviour by study participants due to the awareness of being observed—may influence outcomes (Chavez-MacGregor and Giordano 2016). Therefore, we continue encouraging the use of observational studies to estimate treatment effects in the real-world population, provided that results are interpreted carefully.

Conclusions

Multivariable Cox regression, propensity trimming and hierarchical modelling do not solve the problem of unmeasured confounding. IV analysis may deal with unmeasured confounding, but in this study it is questionable whether the IV fulfills all assumptions and, therefore, we cannot conclude that treatment has a causal relationship with 10-year DMFS. We would like to stress that all of these statistical techniques have some very strong untestable assumptions. It is, therefore, of crucial importance to critically evaluate these assumptions and to be very careful in drawing definite conclusions.

Acknowledgements The authors thank the registration team of the Netherlands Comprehensive Cancer Organisation (IKNL) for the collection of data for the Netherlands Cancer Registry.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

References

Aarts E, Dolan CV, Verhage M, van der Sluis S (2015) Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. BMC Neurosci 16:94

Agarwal S, Pappas L, Neumayer L, Kokeny K, Agarwal J (2014) Effect of breast conservation therapy vs mastectomy on disease-specific survival for early-stage breast cancer. JAMA Surg 149(3):267–274

Antman K, Amato D, Wood W, Carson J, Suit H, Proppe K et al (1985) Selection bias in clinical trials. J Clin Oncol 3(8):1142–1147

Austin PC (2017) A tutorial on multilevel survival analysis: methods: models and applications. Int Stat Rev. 85(2):185–203

Baiocchi M, Cheng J, Small DS (2014) Instrumental variable methods for causal inference. Stat Med 33(13):2297–2340

Boef AG, le Cessie S, Dekkers OM (2013) Instrumental variable analysis. Ned Tijdschr Geneesk. 157(4):A5481

Braitman LE, Rosenbaum PR (2002) Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Int Med. 137(8):693–695

Brookhart MA, Wyss R, Layton JB, Sturmer T (2013) Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes. 6(5):604–611

Chavez-MacGregor M, Giordano SH (2016) Randomized clinical trials and observational studies: is there a battle? J Clin Oncol 34(8):772–783

Chen K, Liu J, Zhu L, Su F, Song E, Jacobs LK (2015) Comparative effectiveness study of breast-conserving surgery and mastectomy in the general population: a NCDB analysis. Oncotarget 6(37):40127–40140

Dekkers OM (2011) On causation in therapeutic research: observational studies, randomised experiments and instrumental variable analysis. Prev Med 53(4–5):239–241

Federspiel JJ, Anstrom KJ, Xian Y, McCoy LA, Effron MB, Faries DE et al (2016) Comparing inverse probability of treatment weighting and instrumental variable methods for the evaluation of adenosine diphosphate receptor inhibitors after percutaneous coronary intervention. JAMA Cardiol. 1(6):655–665

Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER et al (2002) Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. N Engl J Med 347(16):1233–1241

Fisher S, Gao H, Yasui Y, Dabbs K, Winget M (2015) Survival in stage I-III breast cancer patients by surgical treatment in a publicly-funded healthcare system. Ann Oncol 26(6):1161–1169

Fritz APC, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, Whelan S (2000) International classification of diseases for oncology, 3rd edn. World Health Organization, Geneva

Giordano SH (2015) Comparative effectiveness research in cancer with observational data. Am Soc Clin Oncol Educ Book. pp e330–e335

Greenland S (2000) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29(4):722–729

Hartmann-Johnsen OJ, Karesen R, Schlichting E, Nygard JF (2012) Survival is better after breast conserving therapy than mastectomy for early stage breast cancer: a registry-based follow-up study of Norwegian women primary operated between 1998 and 2008. Ann Surg Oncol 22(12):3836–3845

Hernan MA, Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? Epidemiology 17(4):360–372

Hershman DL, Wright JD (2012) Comparative effectiveness research in oncology methodology: observational data. J Clin Oncol 30(34):4215–4222

Hofvind S, Holen A, Aas T, Roman M, Sebuodegard S, Akslen LA (2015) Women treated with breast conserving surgery do better than those with mastectomy independent of detection mode, prognostic and predictive tumor characteristics. Eur J Surg Oncol 41(10):1417–1422

Hutchins LF, Unger JM, Crowley JJ, Coltman CA Jr, Albain KS (1999) Underrepresentation of patients 65 years of age or older in cancer-treatment trials. New Engl J Med. 341(27):2061–2067

Hwang ES, Lichtensztajn DY, Gomez SL, Fowble B, Clarke CA (2013) Survival after lumpectomy and mastectomy for early stage invasive breast cancer: the effect of age and hormone receptor status. Cancer 119(7):1402–1411

Kjaersgaard MI, Parner ET (2016) Instrumental variable method for time-to-event data using a pseudo-observation approach. Biometrics 72(2):463–472

Mansournia MA, Altman DG (2016) Inverse probability weighting. Bmj 352:i189

Moossdorff M, van Roozendaal LM, Strobbe LJ, Aebi S, Cameron DA, Dixon JM et al (2014) Maastricht Delphi consensus on event definitions for classification of recurrence in breast cancer research. J Natl Cancer Inst. https://doi.org/10.1093/jnci/dju288

Morrow M, Jagsi R, Alderman AK, Griggs JJ, Hawley ST, Hamilton AS et al (2009) Surgeon recommendations and receipt of mastectomy for treatment of breast cancer. JAMA 302(14):1551–1556

Onitilo AA, Engel JM, Stankowski RV, Doi SA (2015) Survival comparisons for breast conserving surgery and mastectomy revisited: community experience and the role of radiation therapy. Clin Med Res 13(2):65–73

Siesling S, van de Poll-Franse LV, Jobsen JJ, Repelaer van Driel OJ, Voogd AC (2005) Trends and variation in breast conserving surgery in the southeast and east of the Netherlands over the period 1990-2002. Ned Tijdschr Genesk 149(35):1941–1946

Sobin LHWC (2002) International union against cancer, TNM classification of malignant tumours, 6th edn. Wiley, New York

Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ (2007) Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 297(3):278–285

Sturmer T, Wyss R, Glynn RJ, Brookhart MA (2014) Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. J Int Med. 275(6):570–580

Tsuchiya A, Tsutsumi Y, Yasunaga H (2016) Outcomes after helicopter versus ground emergency medical services for major trauma–propensity score and instrumental variable analyses: a retrospective nationwide cohort study. Scand J Trauma Resusc Emerg Med. 24(1):140

van Maaren MC, de Munck L, Jobsen JJ, Poortmans P, de Bock GH, Siesling S et al (2016a) Breast-conserving therapy versus mastectomy in T1-2N2 stage breast cancer: a population-based study on 10-year overall, relative, and distant metastasis-free survival in 3071 patients. Breast Cancer Res Treat 160(3):511–521

van Maaren MC, de Munck L, de Bock GH, Jobsen JJ, van Dalen T, Linn SC et al (2016b) 10 year survival after breast-conserving surgery plus radiotherapy compared with mastectomy in early breast cancer in the Netherlands: a population-based study. Lancet Oncol. 17(8):1158–1170

van Maaren MC, Strobbe LJA, Koppert LB, Poortmans PMP, Siesling S (2018) Nationwide population-based study of trends and regionalvariation in breast-conserving treatment for breast cancer. Br J Surg 105(13):1768–1777

Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A et al (2002) Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. The New Engl J Med 347(16):1227–1232

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.