Real-world evidence was feasible for estimating effectiveness of chemotherapy in breast cancer; a cohort study

Real-world evidence was feasible for estimating effectiveness of chemotherapy in breast cancer; a cohort study

Authors: Ewan Gray[1], Joachim Marti[2], David H Brewster[1], Jeremy C Wyatt[3], Romain Piaget-Rossel[2] and Peter S Hall[1]

[1]University of Edinburgh [2]University of Lausanne [3]University of Southampton

Abstract

Objective: Evidence-based guidelines recommend adjuvant chemotherapy in early stage breast cancer whenever treatment benefit is considered sufficient to outweigh the associated risks. However, many groups of patients were either excluded from or underrepresented in the clinical trials that form the evidence base for this recommendation. This study aims to determine whether using administrative healthcare data – Real World Data (RWD) - and econometric methods for causal analysis to provide 'Real World Evidence' (RWE) are feasible methods for addressing this gap.

Methods: Cases of primary breast cancer in women from 2001 to 2015 were extracted from the Scottish cancer registry (SMR06) and linked to other routine health records (inpatient and outpatient visits). Four methods were used to estimate the effect of adjuvant chemotherapy on disease-specific and overall mortality: (1) regression with adjustment for covariates (2) propensity score matching (3) instrumental variables analysis and (4) regression discontinuity design. Hazard ratios for breast cancer mortality and all-cause mortality were compared to those from a meta-analysis of randomised trials.

Results: 39,805 cases included in the analyses. Regression adjustment, propensity score matching and instrumental variables were feasible while regression discontinuity was not. Effectiveness estimates were similar between RWE and randomised trials for breast cancer mortality but not for all-cause mortality.

Conclusions: RWE methods are a feasible means to generate estimates of effectiveness of adjuvant chemotherapy in early stage breast cancer. However, such estimates must be interpreted in the context of the available randomised evidence and the potential biases of the observational methods.

Introduction

Adjuvant chemotherapy is indicated in cases of early stage breast cancer whenever treatment benefit is considered sufficient to outweigh the associated risks (1). The benefits of chemotherapy, in terms of improved survival and disease free survival, are known to vary on a case by case basis depending on a number of prognostic markers. Likewise, the risks of chemotherapy vary depending on the characteristics of the patient (2). Patients, with advice from their clinicians, must choose whether or not to undergo adjuvant chemotherapy taking account of these individual factors and their own beliefs and preferences. The shared decision making process requires a patient specific risk assessment and the effective communication of patient preference, risks, and benefits information between the clinician and the patient. To facilitate this process and help improve decisions a number of tools have been developed that quantify the risks and benefits of the available treatments (3-5). Data from randomised controlled trials (RCTs) are the primary source for decision making tools however these are limited by strict patient inclusion criteria, leading to concerns about whether treatment benefit estimates are accurate for all patients (generalisability). It is proposed that real world evidence (RWE) can contribute to informing these decisions by providing accurate treatment benefit estimates from more representative real world data (6).

Real world data (RWD) refers to data used for decision making that are collected outside of randomised controlled trials (RCTs) (7). RCTs are the gold standard for reliably measuring treatment efficacy and constitute the primary source of evidence to inform decision-making in health care. Randomisation, correctly implemented (8), guarantees unbiased estimates of the average treatment effect (in expectation) by ensuring balance of observed and unobserved covariates in treatment and control groups. Real world data in contrast are observational in nature and therefore subject to additional sources of bias (9). Methods of analysis, ‘designs’, have been developed to allow less biased estimates of treatment effects under reasonable assumptions. These methods have contributed to the ‘credibility revolution’ in economics (10) and have led some to re-evaluate existing hierarchies of evidence in medicine (11).

The “quasi-experimental” methods available to researchers differ in the mechanisms used to mirror random assignment; historically the greatest limitations to their application have been data availability and quality, and concerns about the feasibility of more advanced methods.

This study makes use of high quality routine data to implement alternative candidate methods and compare estimates both between methods and with the available RCT evidence. The randomised evidence comes from a series of progressively updated meta-analyses published by the Early Breast Cancer Trialists Collaborating Group (EBCTCG) (2,

12, 13). Results from the most recent meta-analysis are used make comparisons with the RWE estimates from this study.

We consider four candidate methods:

1. Regression with adjustment for covariates (RA): Uses multiple regression based methods to adjust for the imbalance in observed covariates between treated and untreated cases.
2. Propensity score matching (PSM) (14): Uses rich prognostic data to create propensity scores and match treated and untreated cases, reducing confounding by indication
3. Instrumental variables (IV) (15): Makes use of variables that are assumed to causally affect the treatment decision but have no effect on outcomes other than indirectly, via changing the probability of treatment.
4. Regression discontinuity design (RDD) (16): Exploits the variation in treatment use created by a treatment guideline based on a threshold level of estimated treatment benefit provided by an online tool.

This study aims to explore the feasibility, and compare the results of RWE methods for estimating the effectiveness of adjuvant chemotherapy for early stage breast cancer. Due to the reasonable concerns about bias in RWE methods the approach we have taken is exploratory in nature and seeks to provide extensive contextual information to inform the judgements of readers on the interpretation of RWE estimates. We believe such an open approach, which differs from the typical ‘stand alone’ inference of a randomised trial, is necessary if RWE is to be useful for informing patient and clinician decision making in this setting.

A key feature of this study is that all methods make use of prognostic and treatment benefit predictions about individual women provided by an online prognostication and treatment benefit tool for patients with early stage breast cancer - PREDICT (5).

Some RWE methods have previously been employed in this setting. PSM was used with a large observational study conducted in the USA (17) comparing mastectomy and breast conserving surgery in node negative patients using a registry data set. The results corresponded closely with previously reported trials and provided evidence that the estimated hazard ratios could be generalised beyond trial populations, and this information was influential for clinical practice. The success of PSM for comparing surgical strategies in the same patient group is one reason to believe PSM might also be appropriate for

addressing questions relating to the effectiveness of adjuvant therapies. Other PSM studies focusing on adjuvant therapies have not made comparisons with randomized data.[1]

Application of RDD for the evaluation of healthcare interventions has recently received increased attention and there are some successful examples of this method (18). However, the application of RDD to this clinical area is, to the best of our knowledge, completely novel.

Methods

Patient Data

Patient level data were transferred into the National Services Scotland Safe Haven as an extract from the Scottish Cancer Registry (SCR). All records in the registry with a diagnosis of primary invasive breast cancer (ICD-10 C50) diagnosed in the period between January 2001 and December 2015 were retrieved for analysis. SCR is a population-based registry that covers all residents of Scotland (population approximately 5.5 million). National Records of Scotland provides notification of deaths for registry records. Vital status was recorded up to 1st February 2017 in the analysis extract[2]. Deaths due to breast cancer were defined in accordance with the ICD-10 coding system for causes of death, recorded either as the primary cause of death or as one of three contributing causes of death. Data were restricted to the first occurrence of a primary breast cancer for each patient, subsequent primary breast cancers were excluded. Data linkage was provided by Information Services Division (ISD) to Scottish hospital inpatient and day case records (SMR01) and outpatient records (SMR00). Deterministic linkage was achieved using the Community Health Index (CHI) number unique individual identifiers, which includes a check digit. The linked data sets included all records linked to an included registry case from the period up to 5 years prior to the date of diagnosis. Prognostic factors available for use in the analysis, including the derived PREDICT scores, are described in detail in the supplementary appendix (SA1).

Details of the use of PREDICT scores in each method are displayed in table 1. In the RA, PSM and IV methods PREDICT version 2 scores were used. ‘Prognostic score’ was the PREDICT probability of death from any cause over 10 years. In models with breast cancer specific mortality as the outcome this was the probability of death due to breast cancer over 10 years. PREDICT benefit score was the difference in the probability of survival at 10 years following diagnosis with and without adjuvant chemotherapy.

[1] PSM has also previously been applied to estimate adjuvant chemotherapy effectiveness in a large case series from a single institution in France (18). PSM has also been used for estimating the effects of adjuvant chemotherapy for older women (19, 20) (a specific trial ineligible group) in data from the USA .

[2] For those emigrating from Scotland a date of embarkation is available allowing censoring these observations at the appropriate time.

Cases were excluded if the patient was male, had advanced cancer (clinical M stage = 1), did not receive surgery or received neoadjuvant therapy (chemotherapy or hormone therapy recorded prior to surgery).

Because some groups of patients such as the over 70s and those with comorbidities, were either excluded from, or underrepresented in, the randomised trials, we made a comparison with a subgroup of RWE patients who would meet trial inclusion criteria as well as with the full cohort. Cases were defined as 'trial represented' (TR) if they were under 70 years of age with no recorded Charlson comorbidities, and met criteria related to prognostic factors; either node positive or with 2 or more of: [1] >30mm tumour size, [2] grade 3, [3] ER-, and [4] Her2+ status. The definition of trial represented was based on assessment of the protocols of a number of trials in the meta-analysis (2) and clinical expert opinion. It should be noted that trial inclusion/exclusion protocols varied to some degree and often included elements of clinical judgement.

Econometric methods

All analyses were repeated for two outcomes; breast cancer specific mortality and all-cause mortality as recorded on death certificates. Furthermore, the analyses were repeated in the full cohort and the TR subgroup. Each method of analysis therefore produced four estimated hazard ratios. The details of the implementation are reported in the supplementary appendix (SA2).

Table 1 - Use of person-specific PREDICT scores in the four RWE methods

Method	**How PREDICT outputs were used:**
Regression Adjustment	PREDICT Prognostic score and benefit score were used as explanatory variables. This effectively includes prognostic variable interactions and transforms informed by external evidence (from derivation cohort and previous studies that informed the specification used in PREDICT modelling).
Propensity Score Matching	Prognostic score and benefit scores were used as explanatory variables in the propensity score model.
Instrumental Variables	PREDICT benefit score was used as an instrument. Benefit score interacted with post-2010 dummy variable used as an alternative instrument. Prognostic score is an (independent) explanatory variable.

Regression Discontinuity	Benefit score used as the continuous assignment variable

Comparison with randomised studies

Clinical expert opinion suggests that the predominant chemotherapy regimens in use in Scotland during the 2001-2015 period were newer anthracycline-containing regimens (CAF/CEF), therefore estimates for these types of regimen were used as the best comparison with the RWE estimates. Both direct evidence, from trials comparing newer anthracycline-containing regimens vs. placebo, and indirect evidence via trials of newer anthracycline-containing regimens vs. other active regimens were considered. This required some re-analysis using the trial level summary statistics presented in (2) as a network meta-analysis (19). In the primary analysis versus RWE methods the comparison was made using only direct trial evidence.

To compare RWE with randomised evidence two approaches were taken. First, a statistical test of the difference in the estimated treatment effect between trial and observational sources based on z scores, as suggested by Ioannidis et al (20). z scores above 1.96 or below -1.96 are considered as sufficient evidence that the difference in estimates is beyond that expected by chance. Second, meta-analysis estimates were calculated with and without inclusion of RWE estimates. Results were presented in forest plots to allow visual assessment of heterogeneity. A random effects model was assumed because of known between-trial and trial-observational differences in study populations. Cochrane's Q was used as a statistical test of heterogeneity.

Results

Sample selection

63,116 records were retrieved from the registry. Following removal of duplicate records (non-first cancers and multiple synchronous tumours) and application of the exclusion criteria a total of 39,805 cases remained in the primary analysis. The process is detailed in figure 1. 13% of otherwise eligible cases contained missing prognostic variable data but in most cases only a single variable was missing. A summary of sample characteristics can be found in Table 2.

Figure 1 - Sample selection flow diagram

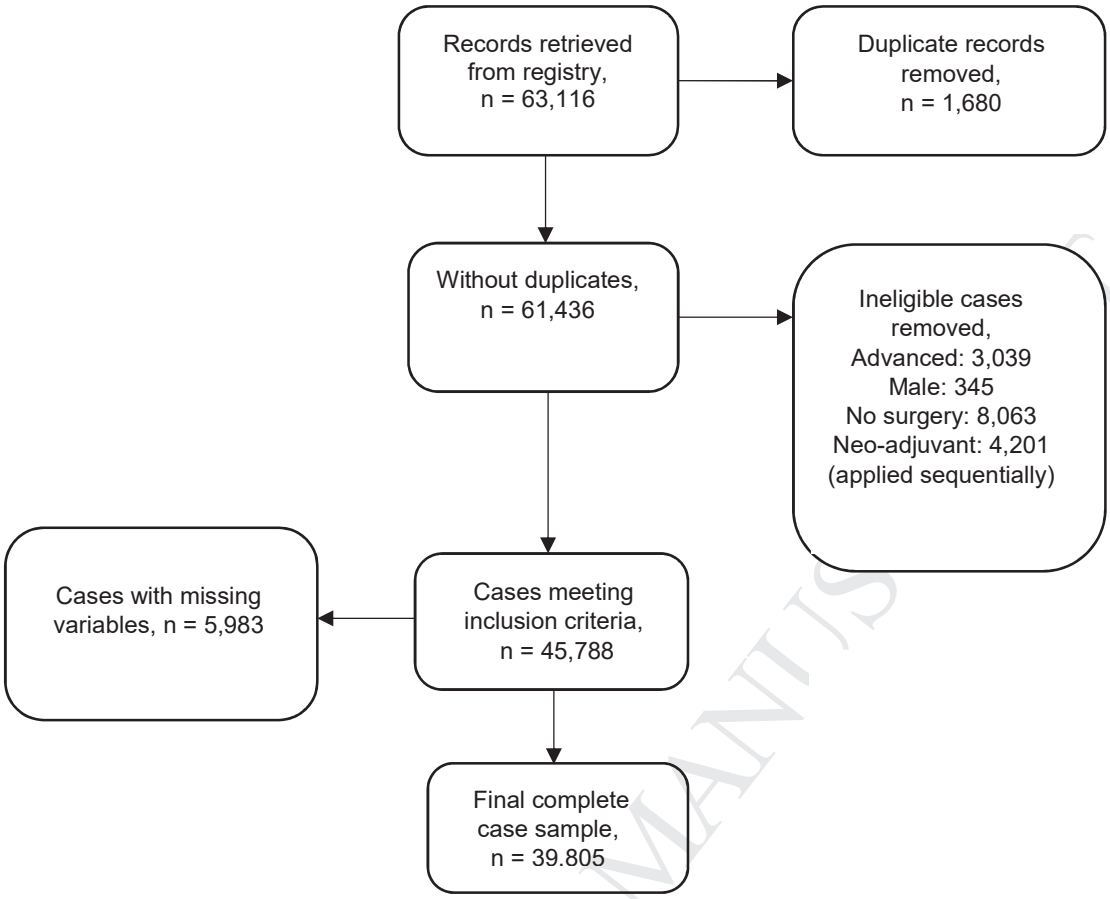


Table 2 - Sample Characteristics

	Full cohort		Trial representative cohort	
Total number of subjects	39,805		12,870	
Total time at risk (years)	278,984		93,222	
Median follow-up (years)	6.35		6.64	
Number of breast cancer deaths	4,977		2,126	
Number of other deaths	3,624		511	
Five-year breast cancer survival rate	87.1%		86.8%	
Median age at diagnosis, years	61		54	
	Number	(%)	Number	(%)
Age <35	503	1.3	321	2.5
35-49	7015	17.6	3879	30.1
50-64	16792	42.2	6655	51.7
65-74	9797	24.6	2015	15.7

>=75	5698	14.3	0	0
Nodes 0	26275	66	3294	25.6
Nodes 1	5747	14.4	4154	32.3
Nodes 2-4	4456	11.2	3202	24.9
Nodes 5-9	1635	4.1	1079	8.4
Nodes 10+	1453	3.7	978	7.6
Tumour size <10mm	5453	13.7	717	5.6
10-19mm	15774	39.6	3907	30.4
20-29mm	10727	26.9	4029	31.3
30-49mm	5967	15	3146	24.4
>=50mm	1884	4.7	1071	8.3
Grade I	5866	14.7	758	5.9
II	19130	48.1	4473	34.8
III	14606	36.7	7564	58.8
ER-	6208	15.6	3900	30.3
ER+	33597	84.4	8970	69.7
Chemotherapy	14589	36.7	10439	81.1
Hormone therapy	29991	75.3	7822	60.8
Chemo + hormone therapy	8875	22.3	6060	47.1
Screen detected	14887	37.4	3602	28
Symptomatic	24827	62.4	9205	71.5
Charlson Index >= 1	2455	6.2	0	0
Charlson Index = 0	37350	93.8	12870	100
SIMD deprivation quintile 1 (most deprived	6893	17.3	2377	18.5
SIMD 2	7823	19.7	2480	19.3
SIMD 3	8478	21.3	2705	21
SIMD 4	8300	20.9	2719	21.1
SIMD 5 (least deprived)	8310	20.9	2588	20.1
<3% Chemo benefit PREDICT	24475	61.5	3357	26.1
3-5%	8382	21.1	4570	35.5
>5%	6948	17.5	4943	38.4
	Mean	s.d.	Mean	s.d.
Age	60.72	0.067	53.47	0.089

Tumour size	22.03	0.076	27.02	0.149
Total inpatient days	2.88	0.065	1.35	0.061
Total outpatient visits	6.2	0.048	5.33	0.07
PREDICT benefit score	2.93	0.012	4.66	0.02

s.d.: standard deviation.

Feasibility

RA using Cox regression was feasible in both TR and full cohort groups. PSM was also determined to be feasible. The distribution of propensity scores are displayed in supplementary appendix (SA3). The region of common support was sufficient in both the trial represented and full cohort groups. The balance of baseline covariates between matched treated and non-treated units in both groups was compared for each matching method (SA3). PSM 1 showed some imbalances for the TR group and more severe imbalance in the full cohort. PSM 2 achieved good balance in the TR but not in the full cohort. PSM 3 achieved good balance for both TR and full cohorts. Consequently, PSM 3 was selected as the primary analysis. Results for all PSM methods are available in appendix SA3.

Both IV approaches demonstrated feasibility. The first-stage regression results are displayed in appendix SA4. Both instruments showed promise through statistically significant associations with chemotherapy use in the first stage regressions. Note that much of the variation in chemotherapy use caused by the instruments may be in node negative patients who would not meet the defined trial represented criteria, therefore these instruments may be more powerful in the full cohort.

RDD was not feasible. Inspection of histograms confirmed that the requirement of continuity in the region of the 3% and 5% thresholds was met for PREDICT v2 but not v1.2 (figures SA5.1 and SA5.2), eliminating PREDICT 1.2 as a candidate assignment variable.

The PREDICT version 2 benefit score with 3% and 5% thresholds had potential for an RDD. To determine whether or not treatment guidelines and norms actually create such a discontinuity we inspected the binned scatterplots displayed in Figures SA5.3 and S5.4, to visualise the relationship between the assignment variable and the probability of chemotherapy use.

The binned scatterplots show no clear discontinuity in the probability of using chemotherapy at the 3% or 5% thresholds in the trial represented group. The plots suggest that the probability of chemotherapy use is already high by 3% chemotherapy benefit and increases

only gradually past this threshold. Based on these core assumptions not being met, estimation of the treatment effect using RDD was halted.

Comparison with trial meta-analysis estimates

Full details of the randomised trial evidence and network meta-analysis (NMA) results are in the Supplementary Appendix (SA6). Comparisons are made with only the direct randomised evidence to simplify the analysis and presentation, and because the NMA results indicate little difference from including the indirect evidence. All the preceding RWE estimates of the effectiveness of adjuvant chemotherapy are summarised together in table 6, together with the combined estimates from a random effects meta-analysis and results of a test of the difference in estimated HR between RWE and trial meta-analysis. Forest plots displaying the estimated HRs for individual trials and RWE methods are displayed in figure 2. Forest plots displaying the individual and pooled estimates are available in SA6.

Breast cancer mortality estimates from each RWE method lie relatively close to the pooled randomised trial estimates, except for IV1. In contrast, there is some evidence of differences between RWE and trial evidence in relation to all-cause mortality, with the PSM estimating lower hazard ratios than trials.

For all RWE methods except PSM Cox, estimates of the hazard ratios for all-cause mortality are lower than for BC-specific mortality, in contrast to the trial estimates. This suggests a downward bias in the all-cause mortality estimates, as the effect of chemotherapy on all-cause mortality is expected to be smaller and proportional to its effect on BC mortality. This is based on the mechanism of action and was demonstrated in the randomised trial meta-analysis results.

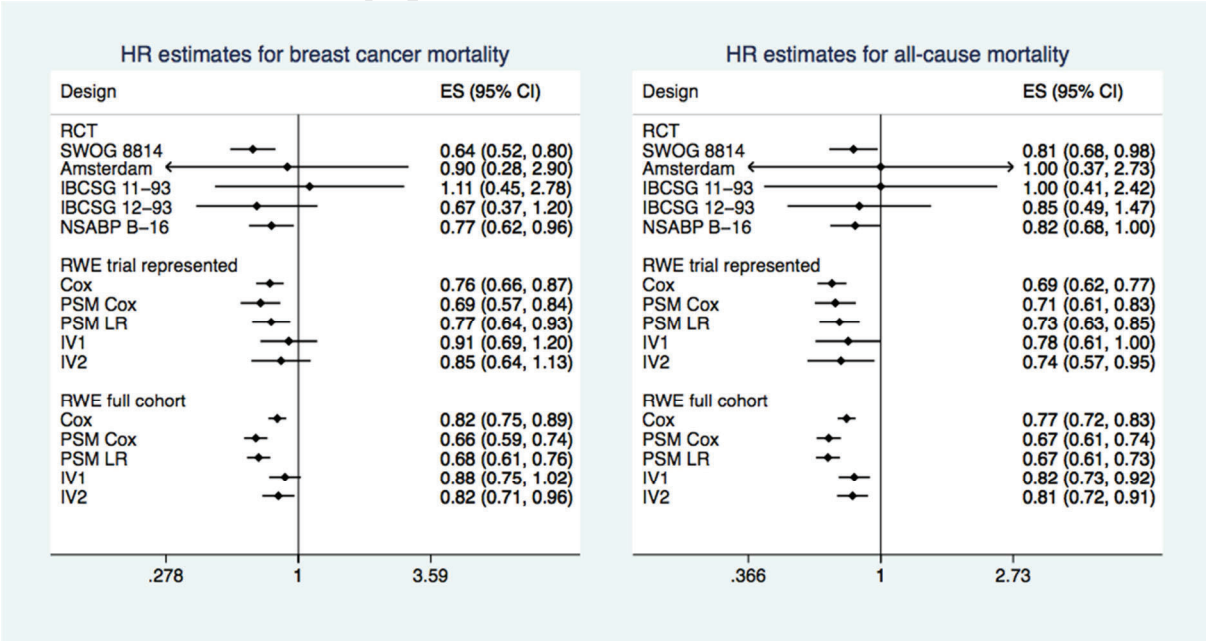
Some differences between RWE estimates are large enough to produce differences in the combined HRs which may be of clinical importance. Estimates from the IV analyses are relatively imprecise in the TR group, which may be due to the relatively small proportion of variation in the use of chemotherapy explained by the instruments. Despite our relatively large sample sizes, the synthesis of RWE and trial evidence in a random effects meta-analysis reduces the uncertainty about the HR by only a small amount compared to using the randomised trial data alone, irrespective of using the TR or full cohort samples.

Table 3 - Comparison of effect sizes (HR, 95% confidence intervals) and pooled estimates from randomised trials, RWE (trial represented and RWE full cohort

	Breast cancer mortality		All-cause mortality	
	HR	**Combined HR**	**HR**	**Combined HR**
EBCTCG meta-analysis: Direct evidence newer anthracycline regimens vs placebo				
	0.71 (0.62; 0.83)	-	0.83 (0.73; 0.94)	-
RWE estimates (trial represented)				
RA	0.76 (0.67; 0.87)	0.74 (0.67; 0.82)	**0.69** (0.62; 0.77)**	0.75 (0.69; 0.81)
PSM Cox	0.69 (0.57; 0.84)	0.71 (0.63; 0.79)	0.71 (0.61; 0.83)	0.78 (0.71; 0.86)
PSM LR	0.77 (0.64; 0.93)	0.74 (0.66; 0.82)	0.73 (0.63; 0.85)	0.79 (0.71; 0.87)
IV1	0.91 (0.70; 1.20)	0.75 (0.66; 0.86)	0.78 (0.61; 1.00)	0.82 (0.73; 0.91)
IV2	0.85 (0.64; 1.13)	0.74 (0.65; 0.84)	0.74 (0.57; 0.95)	0.81 (0.72; 0.91)
RWE estimates (full cohort)				
RA	0.82 (0.75; 0.89)	0.79 (0.73; 0.85)	0.77 (0.72; 0.83)	0.78 (0.74; 0.83)
PSM Cox	0.66 (0.59; 0.74)	0.68 (0.62; 0.75)	**0.67** (0.61; 0.74)**	0.75 (0.67; 0.85)
PSM LR	0.68 (0.61; 0.76)	0.69 (0.63; 0.76)	**0.67** (0.61; 0.73)**	0.76 (0.67; 0.85)
IV1	**0.88* (0.75; 1.02)**	0.78 (0.68; 0.89)	0.82 (0.73; 0.92)	0.82 (0.75; 0.90)
IV2	0.82 (0.71; 0.96)	0.76 (0.69; 0.85)	0.81 (0.72; 0.91)	0.82 (0.75; 0.89)

Note: * p-value < 10%, ** p-value < 5%. P-values obtained from a student t-test with H0: RWE hazard ratio = EBCTCG hazard ratio. Forest plot of EBCTCG estimates and each RWE estimate obtained from trial represented sample is provided in SA6 figures SA6.4-SA6.23.

Figure 2 - Forest plots; Trial and RWE estimated HR for breast cancer and all-cause mortality



Discussion

The RWD used in this study is of large size and high quality. Linkage to a range of other routine data sets allows an assessment of patient comorbidities that would not be possible with registry records alone. Use of an existing, validated prognostic score efficiently uses prognostic information in a manner consistent with existing epidemiological evidence.

We have attempted to be comprehensive and transparent in the application and evaluation of RWE methods, rather than simply presenting one method with what we judge the 'best' result or justification. This guards against the potential for spurious results arising from a selective post hoc application of the methods or selective use of a particular specification within each method based on the results. Using RWD there is clearly a wider scope for creating bias in this way, as compared to using trial data. This is an important analytical problem similar to 'p-hacking' or a 'garden of forking paths' that has been described in other settings (21).

A limitation of this study is that the comparison of real world and clinical trial evidence may not directly assess the validity of the real-world evidence. Validating against clinical trials could be misleading if the average treatment effect in a trial represented RWD group and the actual trial samples differs systematically. At the same time, interpreting the results as estimates of systematic differences in treatment effects between trial and real-world populations will be unwise if the observational methods lack internal validity. In this study the use of a trial represented population in the RWD should minimise this potential problem. Other solutions to this problem could be attempted in future research if additional data are available. While we were fortunate that a comprehensive IPD meta-analysis was available, this analysis was restricted to using the summary statistics from the published report. This limited the degree to which we could match the randomised trial and RWD populations for baseline characteristics and treatment protocols or make use of other methods of adjustment (22). Future studies should consider a deeper collaboration with trialists to allow more nuanced comparison. Another area that could be explored in future research is the design of other tests of internal validity, such as falsification endpoints (23), of RWE methods such as IV in this setting.

This analysis would be enhanced if more detailed data were available for specific variables. Chemotherapy use was only available as a binary variable, rather than the details of the regimens used. Her2 status and trastuzumab use were not available for many observations (and not available in any case prior to 2009). 13% of cases were excluded from the analysis

due to other missing prognostic data. Another limitation of this analysis is that it relies on accurate coding of causes of death on death certificates to ascertain breast cancer deaths. This may be less accurate in RWD than in the trial setting, where detailed follow-up to ascertain deaths with recurrence of breast cancer has been used. RWE could also be enhanced by reporting of effects on multiple alternative causes of death (e.g. breast cancer, cardiovascular, other) where such data are available, using an appropriate methodology to account for competing risks. This would also more nuanced conclusions regarding the potential biases related to each of these outcomes.

Conclusions

Regression adjustment, PSM and IV were feasible methods for obtaining treatment effect estimates from these real-world data while RDD was not. While concordance with randomised trial evidence was demonstrated for cause-specific mortality there was some indication of bias in estimates of chemotherapy effects on all-cause mortality. We conclude that even with large, good quality datasets and careful validation of the assumptions underlying these methods, RWE should be interpreted cautiously, in the context of the available RCT evidence and with consideration of alternative methods that can be implemented using observational data.

Acknowledgements

We would like to thank the SATURNE advisory group - David Cameron, Fiona Watt, Iain MacPherson, Larry Hayward, Colin McCowen, Gianluca Baio and Paul Pharoah - for their generous advice and support. We would also like to thank David MacAllister, Rachel Meacock, Patrick Taffe, the HESG participants and two anonymous reviewers for their helpful suggestions. Finally, we would like to thank all the women who have provided the data that has been used in this study.

Conflict of interests

All authors have no conflicts of interest to declare.

Funding

This study was funded by the Chief Scientist Office (CSO), Scotland. The funder had no role in the design or reporting of the study.

References

1. National Collaborating Centre for Cancer. CG80 Early and locally advanced breast cancer: diagnosis and treatment. National Institute For Health and Clinical Excellence; 2009.
2. Early Breast Cancer Trialists' Collaborative Group. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100 000 women in 123 randomised trials. The Lancet. 2012;379(9814):432-44.
3. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in Primary Breast-Cancer. Breast Cancer Res Tr. 1992;22(3):207-19.
4. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol. 2001;19(4):980-91.
5. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res. 2010;12(1):R1.
6. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? New Engl J Med. 2016;375(23):2293-7.
7. Garrison LP, Jr., Neumann PJ, Erickson P, Marshall D, Mullins CD. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. Value Health. 2007;10(5):326-35.
8. Altman DG, Schulz KF, Moher D. Turning a blind eye - Testing the success of blinding and the CONSORT statement. Brit Med J. 2004;328(7448):1135-.
9. Rosenbaum PR. Discussing Hidden Bias in Observational Studies. Ann Intern Med. 1991;115(11):901-5.
10. Angrist JD, Pischke J-S. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. Journal of Economic Perspectives. 2010;24(2):3-30.
11. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J Epidemiol Community Health. 2012;66(12):1182-6.

12. Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomised trials. The Lancet. 1998;352(9132):930-42.
13. Early Breast Cancer Trialists' Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet. 2005;365(9472):1687-717.
14. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
15. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91(434):444-55.
16. Oldenburg CE, Moscoe E, Barnighausen T. Regression Discontinuity for Causal Effect Estimation in Epidemiology. Curr Epidemiol Rep. 2016;3:233-41.
17. General Accounting Office (GAO) United States of America. Breast Conservation versus Mastectomy; Patient Survival in Day-to-Day Medical Practice and in Randomized Studies. Washington, D.C.; 1994.
18. Moscoe E, Bor J, Barnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. J Clin Epidemiol. 2015;68(2):122-33.
19. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Methods. 2012;3(2):111-25.
20. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA. 2001;286(7):821-30.
21. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychol Sci. 2011;22(11):1359-66.
22. Signorovitch J, Sikirica V, Erder M, Xie J, Lu M, Hodgkins P, et al. Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research. Value in Health. 2012;15(6):940-7.
23. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? Jama. 2013;309(3):241-2.

Appendix

1 Prognostic variables & PREDICT further details

Prognostic factors available in the registry extract included: age at diagnosis, number of lymph nodes examined and number positive, tumour size (maximum pathological diameter in mm), tumour histological grade (categorical: 1-3), mode of detection (screen-detected or symptomatic), Estrogen receptor (ER) status, Her2 status, chemotherapy use (binary) and hormone therapy use (binary). The linked data sets include many variables related to each type of healthcare use. This analysis makes use of three specific derived variables:

1. Total number of inpatient bed days (log transformed) in the previous 5 years.
2. Any inpatient or day case attendance related to a Charlson Index (1) included comorbidity (binary variable) in the previous 5 years.
3. Total number of prescribed medications in the previous 12 months (log transformed).

PREDICT 10-year prognostic scores were calculated for each individual woman based on their recorded risk factor information using the algorithms supplied by the PREDICT authors (version 1.2 and version 2). Data were unavailable for some variables needed as inputs for the PREDICT model, these included HER2 status (prior to 2009), trastuzumab use and Ki-67 status. Ki-67 status is not recorded in these data, as it was not routinely recorded in Scotland, therefore all cases were assigned to the “unknown” category for this variable. HER2 status is only recorded from 2009. Cases with missing data for HER2 were assigned the “unknown” category. The scores include the probability of death from all causes, probability of death from breast cancer accounting for competing risk, and adjuvant therapy benefit (chemotherapy, hormone therapy and trastuzumab) as the percentage point reduction in the probability of all-cause mortality for each adjuvant therapy (‘Benefit scores’). Details of the calculation of the prognostic index are available in (2, 3).

PREDICT provides personalised prognostic information displayed as 5-year and 10-year survival estimates both with and without adjuvant therapies (chemotherapy, hormone therapy and trastuzumab). Results are presented both in textual format using a frequency based description of risk, and graphically in the form of bar charts with percentages labelled (http://www.predict.nhs.uk/). The first online version of the tool was published in 2010 (4) (v1). A series of updates made since launch have added new prognostic variables and refined the algorithm’s predictions. The first update published in 2012 (2), added HER2 status as a prognostic marker and allowed calculation of Trastuzumab treatment benefit estimates (v.1.2). In 2014, the tumour proliferative marker Ki-67 was added as an optional

prognostic variable (5) (v1.3). The most recent update, in 2016, refined the model by including age at diagnosis in the breast cancer-specific death prediction as well as recoding tumour size and nodal status as continuous rather than categorical variables (v2) (3). In this study we make use of versions 1.2 and 2.

Local treatment protocols in Scotland (and in the rest of UK) have defined thresholds, derived from international consensus, for recommending treatments based on estimated treatment benefit (personal communication – treatment protocols for NHS Lothian & NHS Greater Glasgow & Clyde)(6):

- <3% (percentage point) reduction in mortality over 10 years: Do not recommend adjuvant chemotherapy
- 3-5% reduction in mortality over 10 years: Discuss adjuvant chemotherapy
- >5% reduction in mortality over 10 years: Recommend adjuvant chemotherapy

Widespread use of local protocols based on explicit thresholds are relatively recent (within approximately the last 5 years). National guidelines have considered the idea of threshold but have left the precise level implicit with statements such as “Adjuvant anthracycline-taxane combination chemotherapy should be considered for all patients with breast cancer where the additional benefit outweighs risk” (7). Both local protocols and national guidelines note the need for shared decision making in this area and do not preclude treatment outside of the prescribed thresholds.

The preceding introduction described the current context of adjuvant therapy decision making in Scotland. Treatment decisions in the 2001-2015 period covered by the available data were not made under the existing guidelines or with the current PREDICT online tool prior to its inception in 2010. However, we believe the existing guidelines and online tool may be good proxy measures of the implicit thresholds and treatment benefit estimates of past practice. Some evidence exists to suggest similar methods have been in use over this period although in a less standardised manner. A survey of UK breast unit lead clinicians conducted in 2001/2002 to ascertain use of prognostic scoring systems indicated widespread but varied use of such methods. Among the 168 out of 218 units, 53% of units stated that a prognostic index was used as part of a formal treatment protocol. Further analysis of the treatment protocols of 22 of these units revealed a high degree of variation in how prognostic information was used to guide treatment decisions (8). A retrospective study of multi-disciplinary team (MDT) decisions conducted in the UK between 2007 and 2011 assessed concordance between a PREDICT protocol (as above: <3, 3-5,>5%) based treatment recommendation and actual treatment recommendation (6). In total, in 91 out of

109 cases the MDT recommendation was concordant with what would be suggested by the Predict protocol. A higher proportion of discordant cases were observed in the <3% predicted treatment benefit group. In this group, 8 out of 14 patients were recommended chemotherapy in contradiction to the protocol.

2 Econometric Methods

Regression Adjustment

Regression adjustment (RA) was implemented using Cox regression. Explanatory variables in the Cox regression included: chemotherapy use, PREDICT 10-year probability of mortality, age at diagnosis, number of positive lymph nodes, pathological tumour size, tumour histological grade, mode of detection, ER status, Her2 status, hormone therapy use, radiotherapy use, year of diagnosis, Scottish Index of Multiple Deprivation (SIMD) quintile, Charlson comorbidity status, log total inpatient bed days and log total outpatient visits (5 years prior to diagnosis). Interactions of other clinical prognostic factors with ER status were also included based on the reported PREDICT modelling process which fits separate models for ER+ and ER- groups. In the analysis with breast cancer death as the outcome PREDICT 10-year probability of breast cancer mortality was used in instead of PREDICT 10-year probability of mortality.

PSM Analysis

Probit regression was used to generate the propensity scores. Predicted probabilities of a chemotherapy use for each observation were obtained based on regression of the same set of explanatory variables as in the RA analysis on a binary indicator of chemotherapy use.

Matching of treated and non-treated observations was achieved by nearest-neighbour matching on the propensity scores. Three matching methods were investigated: [PSM 1] treated observations were matched 1:1 with non-treated observations with replacement (re-use of non-treated observations), [PSM 2] matching 1:1 without replacement, and [PSM 3] limited 1:1 matching with replacement to matching within calipers set to 0.25 standard deviations of the logit of the propensity score, as suggested by (9). In all cases matching was restricted to observations within the region of common support (only treated observations with propensity scores between the maximum and minimum scores in the non-treated observations were included). Furthermore, the sample was trimmed to observations ranging from a propensity score of 0.05 to 0.95 prior to matching.

To examine the quality of the matches achieved by each matching method, baseline covariate balance in the matched treated and non-treated samples was assessed by

comparing the means or proportions of baseline covariates.

Cox regression (PSM Cox) and Peto logrank method (10) (PSM LR) were used to estimate hazard ratios for chemotherapy in the matched samples. Proportional hazard was assumed based on reported results from the trial meta-analysis which found no evidence against this assumption (11, 12), it was assumed this would also hold in the RWD setting. When matching with replacement was used observations were weighted accordingly. Confidence intervals were calculated by a simple bootstrap of individual cases with 1000 iterations.

An additional sensitivity analysis was carried out with the proposed instrumental variable removed from the set of explanatory variables in the probit regression used to obtain the propensity scores. This regression model for propensity scores would be more appropriate if the instruments are judged to be valid (13). As the validity of the instruments cannot be tested empirically therefore the interpretation of results, and/or weighting given to each method's set of results, must depend on the readers judgement and prior beliefs.

Instrumental Variables Analysis

After considering a number of candidate instruments, two were selected to be used in treatment effect estimation: PREDICT benefit score (IV 1) and the PREDICT benefit score interaction with a post 2010 dummy variable (IV 2). The first specification is more efficient (greater statistical power) but with greater potential bias while the second is less efficient but has less potential for bias. See the section below (SA2) for instruments considered and reasons for selection of the instruments. IV1 relies on the benefit score being independent from expected survival conditional on the prognostic score and chemotherapy use. As benefit score is the difference in expected survival with and without chemotherapy this should be the case by construction, however it requires the prognostic model to be well calibrated. IV2 exploits the introduction of the PREDICT online tool and this assumes that the ability to access this information influenced the decisions made across the range of PREDICT benefit scores.

PREDICT scores have the potential to provide a valid instrumental variable but this relies on the prognostic model being perfectly calibrated so that all information about survival available from the input variables is captured by the prognostic score, to guarantee that the benefit score (which is a function of these same inputs) is independent from survival without treatment. An instrument that does not rely on this assumption can be created by interacting PREDICT benefit score with a post-2010 dummy variable. This exploits the change in the

relationship between benefit score and the adjuvant therapy decision that is expected to occur when this information is available to clinicians and patients compared to when it was not available. A possible limitation of the second approach is that this instrument is likely to be weaker, in the sense of having a smaller association with the probability of receiving chemotherapy.

The two-stage residual inclusion (2SRI) strategy was used in this study (14). This method can provide consistent estimates of the treatment effect in situations with non-linear first and second stage regressions (due to limited dependent variables). In this case the first stage was estimated by logistic regression and the second stage was estimated by Cox regression. Confidence intervals were calculated by a simple bootstrap of individual cases with 1000 iterations. A Wald test was used to assess the strength of the proposed instruments in the first stage regression. The assumptions of independence of the instrument from unmeasured confounding variables and that the instrument does not affect outcomes expect through the effect on treatment status cannot be empirically tested. A judgement on their validity in this case must be reached by consideration of theory and the understanding of the decision making process.

RDD Analysis

Before considering estimation of the RDD treatment effect an assessment of whether the assumptions were met in this scenario was made. For full details of assumptions assessed see the section below (SA2).

Assessment of the continuity assumption [A3] was made for both potential assignment variables (PREDICT v1.2 and v2) by a combination of visual inspection of histograms and covariate balance tests for observations +/-0.5% above and below each threshold (3% and 5% probability of chemotherapy benefit). Assumption [A1] was tested by visual inspection of binned scatterplots of chemotherapy use against the assignment variable in two ways; [1] scatterplots with simple binning by quantiles (20 quantiles) and [2] an automated procedure for selecting equally spaced bins using the weighted integrated mean square error (WIMSE) (15). In addition, the plots produced using method [2] also include polynomial regression fit by an automated procedure (15).

Instrumental Variables considerations

Table 2.1 Candidate instrumental variables

Candidate variable	Notes
Temporal:	

(all potentially implemented using binary, multivalue categorical or continuous variables related to date of treatment)	
HER2/trastuzumab	2006 June – SMC approval 2008/9 – roll-out/recorded in data
'3rd generation' chemotherapy (taxanes)	2005/6, 2007 Cochrane review
PREDICT online tool	2010, 2012 v2
Aromatase Inhibitors (AIs)	2005 trials report, 2007 roll-out
Sentinel Node Biopsy (SNB)	2003
No dissection of some negative node SNB (guideline?)	2011
Screening extension trial (47-73)	2007
Screening double reading	2011
Digital mammography	2010 gradual
First screening cohort exits programme – max population screening prevalence reached	2008
Regional:	
Territorial health board (HB) of residence	Can be interacted with the events above to use variation in speed of adoption
Specific selected health boards	Can be interacted with the events above to use variation in speed of adoption
Time diagnosis-to-surgery, time surgery-to-chemo/hormone/radiotherapy – HB level mean	Can be calculated by period or as moving average to give measure of demand pressure
Patient level variables:	

Urban/rural classification	Can influence convenience of chemotherapy Adjusting for deprivation is important
Age in relation to specific boundaries	70/75/80 – candidates for ages at which "old age as a contraindication for chemotherapy might be applied". Known bias for ages that are multiples of 5.
PREDICT scores - **Prognostic score:** probability of survival/mortality at X years without adjuvant treatment - **Benefit score**: percentage benefit from chemotherapy - Higher order transforms, interactions or discontinuities	- Use treatment benefit score conditioned on prognostic score. If model is perfectly calibrated this is guaranteed instrument because benefit score is independent from outcome by design, all effect of benefit score determinants on the outcome are captured in the prognostic score. - Interaction of the chemotherapy benefit score with a post 2010 dummy variable. This takes advantage of effect of the introduction of PREDICT on clinical practice. - Use the apparent discontinuity/non-linearity in chemo uptake in 1-2% benefit range rather than the expected discontinuity at 3%. Extra caution would be needed to protect against chance finding.

Rationale for selecting instruments:

Individual candidate temporal trend variables or event dummy variables were rejected as instruments because of the high potential for confounding by other contemporaneous events. There are too many changing factors in breast cancer incidence and treatment in

this time period to reliably estimate the effect of a single factor on adjuvant chemotherapy use. These include the introduction of new therapies such aromatase inhibitors and targeted therapy with Trastuzumab, changes in the national breast screening programme age range and introduction of digital mammography, changing surgical practices such as greater use of sentinel lymph node biopsy, and long term trends in the incidence of different molecular subtypes of breast cancers (ER+ and ER-) related to factors such as past use of hormone replacement therapy.

Regional variation was explored at the health board level (geographical units of National Health Service organization roughly equivalent to counties usually with a single large hospital delivering breast cancer care), however relatively little variation in adjuvant chemotherapy use was observed after controlling for other factors (prognostic factors and deprivation status). Therefore, regional variables related to provider preferences were not considered viable as instruments. Urban/rural status and age-based variables were rejected as being unlikely to satisfy the exclusion restriction.

PREDICT scores have the potential to provide a valid instrumental variable but this relies on the prognostic model being perfectly calibrated so that all information about survival available from the input variables is captured by the prognostic score, to guarantee that the benefit score (which is a function of these same inputs) is independent from survival without treatment. An instrument that does not rely on this assumption can be created by interacting PREDICT benefit score with a post-2010 dummy variable. This exploits the change in the relationship between benefit score and the adjuvant therapy decision that is expected to occur when this information is available to clinicians and patients compared to when it was not available. A possible limitation of the second approach is that this instrument is likely to be weaker, in the sense of having a smaller association with the probability of receiving chemotherapy.

Two specifications were selected to be used in treatment effect estimation: PREDICT benefit score and the PREDICT benefit score interaction with a post 2010 dummy variable. The first specification is more efficient (greater statistical power) but with greater potential bias while the second is less efficient but has less potential for bias.

Regression Discontinuity Assumptions Further Details

Two broad classes of RDD exist (16). The first class is when the threshold fully determines treatment allocation, i.e. 100% of patients are exposed to treatment above the threshold, and none below; this is called a sharp RDD. The second class is when the treatment rule is less strict, and treatment may occur on either side of the threshold. However, the probability of being treated sharply increases at the threshold. This is called a fuzzy RDD. In this study, as

is common when clinical guidelines and decision aids are in use, we have an opportunity to apply a fuzzy RDD.

In fuzzy RDD analysis there is an assignment variable A with a threshold a_0, the treatment status (T) of each individual (i) is partially determined by the level of A and there are treated and untreated individuals on both sides of the threshold. The probability of treatment is determined by a different function (i.e. is discontinuous) on either side of the threshold (16):

$$Pr(T_i = 1|A_i) = \begin{cases} p_1(A_i) if A_i \geq a_0 \\ p_0(A_i) if A_i < a_0 \end{cases}, p_0(A_i) \neq p_1(A_i)$$

When the assumptions for RDD are met the identified causal effect is typically interpreted as a "local average treatment effect" (LATE) (17). It is sometimes also called the complier average treatment effect or complier average causal effect (CATE/CACE) (18).

The assumptions required for a RDD to yield an unbiased estimate relate to the assignment mechanism, the relationship between assignment and outcomes and the relationship between the treatment threshold and potential confounding factors. Following the exposition provided by Geneletti et al (19) these can be explained based on the concept of conditional independence. The context is as above; there is an assignment variable A with a threshold a_0, patients are recommended treatment, or more likely to be recommended treatment, if the level of A is greater than a_0.

[1] Firstly, the assignment to treatment must be influenced by the level of the assignment variable being over the threshold; or equivalently, a threshold indicator variable Z (1 if $\geq a_0$, 0 otherwise) is not independent of treatment status variable T:

$$Pr(T, Z) \neq p(T)p(Z)$$

This assumption is tested by visual assessment of the discontinuity and estimation of the first-stage regression.

[2] A second assumption is that the level of threshold is independent of the characteristics of the individual (potential confounders C) at any level of the assignment variable. Conditional on the assignment variable, Z is independent from C:

$$Pr(Z, C|A) = Pr(Z|A)p(C|A)$$

The assumption would not hold if different thresholds were applied (e.g. to men and women) and this is not accounted for in the analysis (e.g. by separating the analysis). This assumption is verified by external knowledge of the thresholds.

[3] **Continuity assumption (exchangeability assumption)**. To identify the discontinuity in outcome as being causally related to the indicator variable it is also necessary to assume continuity of the expected outcome conditional on the assignment variable (in both treated and non-treated cases). This can be stated as:

$$E(Y|Z, A = a, T, C) is continuous in (at a_0) for T = 0,1.$$

The assumption would not be valid in cases where individuals can manipulate their assignment scores (without error) so as to be placed on the side of the threshold that they wish. In this scenario, the outcome is no longer independent from the threshold indicator variable conditional on the assignment variable because the observed assignment variable scores are not the true scores. Visual inspection of histograms and formal testing for a discontinuity in the density function of the assignment variable have been suggested to investigate this assumption (20). Another means of checking this assumption is to perform tests of covariate balance between groups of observations above and below, and close to, the threshold.

Another case in which the continuity assumption may not hold is when the assignment variable is a categorical variable with a small number of discrete values. In such cases a discontinuity may exist that could be incorrectly attributed as a treatment effect. This can be determined based on knowledge of the assignment variable and visual inspection of histograms of the assignment variable.

[4] It is assumed that outcome Y is independent from the threshold indicator Z conditional on the treatment status, assignment variable and confounding variables:

$$Pr(Y, Z|T, A, C) = Pr(Y|T, A, C)Pr(Z|T, A, C)$$

This assumption guarantees that the effect of Z on the outcome Y is due entirely to the effect of Z on T and the resulting effect of the change in T on Y. This assumption would not be met if the threshold also influences treatment decisions for other therapies that have effects on the same outcome.

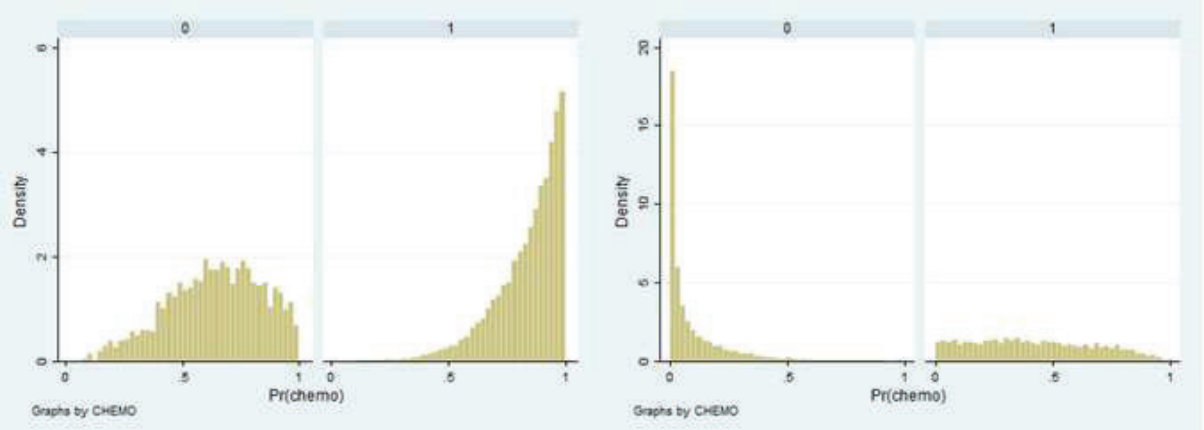
To check this assumption 'placebo' analysis could be undertaken replacing the treatment variable with other treatments that are assumed not to be influenced by the threshold.

[5] In the case of the fuzzy design only, in which treatment is not fully determined by Z, it is necessary to assume that the assignment rule is not reversed for some cases, i.e. Patients are treated only if below threshold rather than only if above threshold. This is unlikely to be important in the scenarios considered in health care. Guidelines may not be followed but are unlikely to be followed in reverse.

The assumptions of threshold being set independently of potential confounders [A2], that the threshold only influences the outcome through the effect on the treatment of interest [A4], and that the assignment rule is never reversed [A5] were considered met based on knowledge of the clinical scenario.

3 PSM Histograms of Propensity Scores, Balance Tests and Full results

Figure 3.1 Propensity score histograms



Left: trial represented population, Right: Full cohort population, 0: no chemotherapy, 1: chemotherapy

Table 3.1 Balance PSM V1 – 1:1 Nearest Neighbour

Variable	Trial Represented				Full cohort			
	Mean untreated	Mean treated	P NH: m1=m2	SMD	Mean untreated	Mean treated	P NH: m1=m2	SMD
Age	56.06	55.55	<0.001	0.063	57.32	55.91	<0.001	0.142
PREDICT 10y mort.	0.38	0.38	0.031	0.035	0.38	0.36	<0.001	0.085
Screen detected	0.31	0.31	0.669	0.007	0.28	0.28	0.435	-0.011
No. Positive nodes	2.55	2.5	0.508	0.011	2.07	2.14	0.208	-0.017
Tumour size	26.18	26.01	0.538	0.01	26.09	25.14	<0.001	0.058
grade 2	0.38	0.39	0.048	-0.033	0.35	0.39	<0.001	-0.088
grade 3	0.59	0.56	0.001	0.054	0.62	0.57	<0.001	0.09
ER +	0.73	0.72	0.171	0.023	0.74	0.76	0.002	-0.042
log inpatient days	0.26	0.24	0.191	0.022	0.36	0.3	<0.001	0.069
log outpatient apts	1.08	1.06	0.19	0.022	1.16	1.09	<0.001	0.072
Charlson (0/1)	0	0	0	0	0.05	0.05	0.01	0.035
SIMD	3.07	3.03	0.119	0.026	2.98	3.06	<0.001	-0.05
hormone therapy	0.62	0.63	0.414	-0.013	0.63	0.65	0.001	-0.043
radiotherapy	0.67	0.69	0.001	-0.057	0.65	0.7	<0.001	-0.117

P NH m1 = m2: p-value for test of null hypothesis mean 1 = mean 2

SMD: Standardised mean difference

Table 3.2 Balance PSM V2 – 1:1 Nearest Neighbour no replacement

Variable	Trial Represented				Full cohort			
	Mean untreated	Mean treated	P NH: m1=m2	SMD	Mean untreated	Mean treated	P NH: m1=m2	SMD
Age	59.62	60.09	0.031	-0.064	59.05	55.91	<0.001	0.308
PREDICT 10y mort.	0.31	0.31	0.24	-0.035	0.27	0.36	<0.001	-0.402

Screen detected	0.45	0.46	0.534	-0.018	0.38	0.28	<0.001	0.207
No. Positive nodes	1.66	1.61	0.543	0.018	0.89	2.14	<0.001	-0.359
Tumour size	21.85	22.62	0.042	-0.06	20.61	25.14	<0.001	-0.339
grade 2	0.46	0.5	0.025	-0.066	0.59	0.39	<0.001	0.401
grade 3	0.37	0.38	0.299	-0.031	0.33	0.57	<0.001	-0.476
ER +	0.81	0.8	0.207	0.037	0.88	0.76	<0.001	0.325
log inpatient days	0.35	0.35	0.806	-0.007	0.36	0.3	<0.001	0.071
log outpatient apts	1.11	1.12	0.742	-0.01	1.1	1.09	0.316	0.014
Charlson (0/1)	0	0	0	0	0.06	0.05	0.001	0.047
SIMD	3	2.95	0.262	0.033	3.11	3.06	0.002	0.042
hormone therapy	0.75	0.72	0.053	0.057	0.8	0.65	<0.001	0.33
radiotherapy	0.67	0.62	<0.001	0.105	0.69	0.7	0.006	-0.037

Table 3.3 Balance PSM V3 – 1:1 Nearest Neighbour no replacement within calipers

Variable	Trial Represented				Full cohort			
	Mean untreated	Mean treated	P NH: m1=m2	SMD	Mean untreated	Mean treated	P NH: m1=m2	SMD
Age	58.97	59.02	0.819	-0.007	59.03	58.8	0.294	0.022
PREDICT 10y mort.	0.32	0.33	0.66	-0.014	0.33	0.32	0.011	0.053
Screen detected	0.42	0.43	0.452	-0.024	0.33	0.36	0.002	-0.065
No. Positive nodes	1.78	1.86	0.497	-0.022	1.45	1.37	0.247	0.024
Tumour size	22.93	23.34	0.348	-0.03	22.85	23.2	0.212	-0.026
grade 2	0.48	0.45	0.112	0.051	0.44	0.46	0.017	-0.05
grade 3	0.42	0.42	0.896	0.004	0.51	0.48	0.001	0.071
ER +	0.78	0.79	0.635	-0.015	0.81	0.83	0.031	-0.045
log inpatient days	0.3	0.32	0.33	-0.031	0.38	0.38	0.758	0.006
log outpatient apts	1.1	1.13	0.337	-0.031	1.13	1.13	0.998	0
Charlson (0/1)	0	0	0	0	0.06	0.05	0.196	0.027
SIMD	3.02	2.96	0.22	0.04	3.05	3.04	0.871	0.003
hormone therapy	0.72	0.71	0.721	0.012	0.73	0.74	0.436	-0.016
radiotherapy	0.67	0.64	0.083	0.056	0.69	0.67	0.108	0.034

Table 3.4 PSM Cox regression results

			All cause mortality			Breast cancer mortality		
Group	**Matching method**	**N treated**	**deaths**	**HR**	**95% CI:**	**deaths**	**HR**	**95% CI:**
Trial	PSM 1	14756	2358	0.67	(0.55,	3182	0.63	(0.54,

Represented					0.85)			0.77)
	PSM 2	4586	475	0.75	(0.62, 0.91)	759	0.72	(0.63, 0.84)
	PSM 3	3830	431	0.69	(0.59, 0.86)	648	0.66	(0.59, 0.8)
Full Cohort	PSM 1	21840	3463	0.71	(0.57, 0.76)	4950	0.63	(0.54, 0.68)
	PSM 2	21840	2772	1.19	(1.08, 1.26)	4169	0.94	(0.88, 0.99)
	PSM 3	9178	1267	0.71	(0.64, 0.8)	1905	0.67	(0.63, 0.76)

HR: Hazard ratio for chemotherapy CI: confidence interval N/A: non-applicable

Table 3.5 PSM Cox regression results – sensitivity analysis without instrumental variable

			All cause mortality			**Breast cancer mortality**		
Group	**Matching method**	**N treated**	**deaths**	**HR**	**95% CI:**	**deaths**	**HR**	**95% CI:**
Trial Represented	PSM 1	14948	3376	0.56	(0.54, 0.77)	2573	0.57	(0636, 0.86)
	PSM 2	4616	475	0.74	(0.64, 0.87)	487	0.78	(0.66, 0.96)
	PSM 3	3888	431	0.69	(0.6, 0.83)	442	0.72	(0.63, 0.92)
Full Cohort	PSM 1	22342	5159	0.6	(0.54, 0.68)	3648	0.66	(0.56, 0.75)
	PSM 2	22342	4240	0.93	(0.88, 1)	2819	1.19	(1.1, 1.29)
	PSM 3	9372	1930	0.71	(0.65, 0.78)	1262	0.77	(0.67, 0.84)

4 Instrumental variables full results

Table 4.1 IV regressions 1st and 2nd stage results by sample

	Trial represented		Full cohort	
Specification	**Coefficient / HR (P Value)**	**95% CI**	**Coefficient / HR (P Value)**	**95% CI**
All-cause mortality				
Instrument 1:				
PREDICT benefit score [1st stage]	0.295 (P<0.001)	(0.222, 0.369)	0.492 (P<0.001)	(0.435, 0.549)
Chemotherapy [2nd stage]	0.77	(0.62, 1)	0.81	(0.73, 0.92)
Instrument 2:				
PREDICT benefit score*post-2010 [first stage]	0.141 (P<0.001)	(0.088, 0.195)	0.094 (P<0.001)	(0.059, 0.129)
Chemotherapy [2nd stage]	0.73	(0.58, 0.96)	0.8	(0.72, 0.91)
Breast Cancer Mortality				
Instrument 1:				
PREDICT benefit score [1st stage]	0.295 (P<0.001)	(0.222, 0.369)	0.457 (P<0.001)	(0.398, 0.515)
Chemotherapy [2nd stage]	0.91	(0.7, 1.2)	0.88	(0.77, 1.03)
Instrument 2:				
PREDICT benefit score*post-2010 [first stage]	0.142 (P<0.001)	(0.088, 0.195)	0.094 (P<0.001)	(0.059, 0.129)
Chemotherapy [2nd stage]	0.79	(0.62, 1.02)	0.82	(0.72, 0.96)
Other explanatory variables omitted from table				

5 Details of assessment of RDD assumptions/requirements

Figure 5 shows histograms over the full range and restricted to the region of the thresholds for PREDICT v1.2. Figure 6 shows the same for PREDICT version 2. The continuity assumption appears to not be met for PREDICT v1.2 due to the high density of observation immediately following the 3% threshold and the high density of observations immediately preceding the 5% threshold. The existence of many peaks and troughs across the whole distribution, along with details of the construction of the score, suggests that this is probably due to the dominance of the categorical variables in the algorithm which generates the scores. The continuity assumption appears to be met for PREDICT v2 because the histogram is relatively smooth in the region of the thresholds. PREDICT version 1.2 was eliminated as a candidate assignment variable based on this assessment.

Figure 5.1 Histograms, PREDICT version 1.2 chemotherapy benefit, full range and region of thresholds

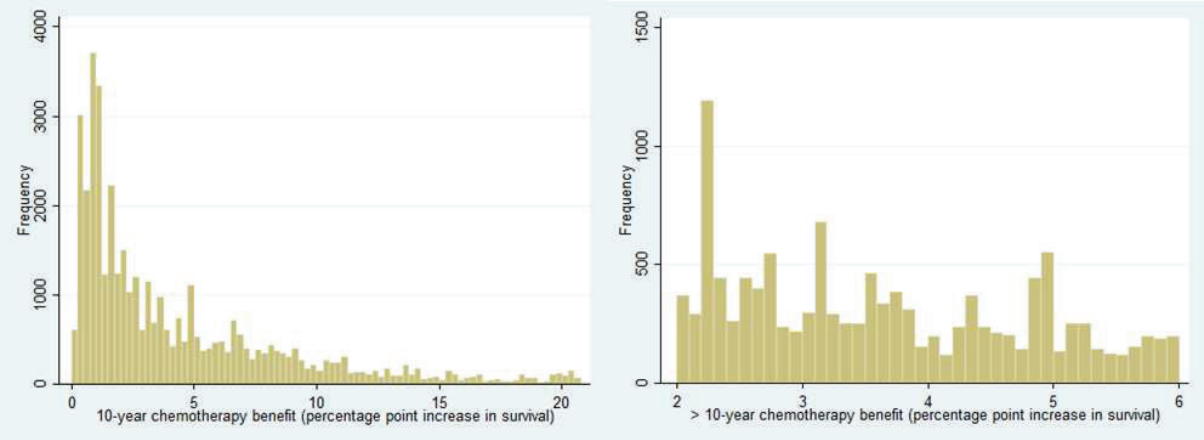
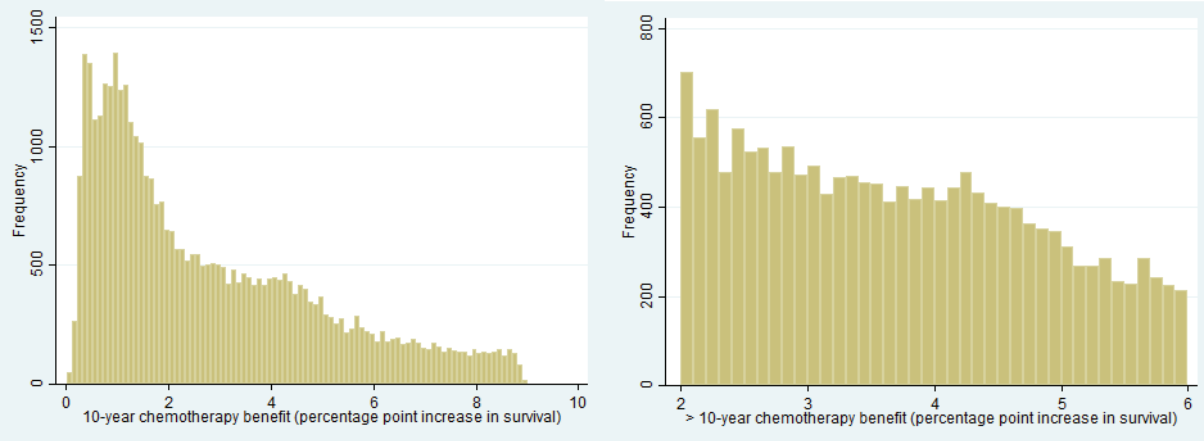


Figure 5.2 Histograms, PREDICT version 2 chemotherapy benefit, full range and region of thresholds



Baseline covariate balance around the 3% threshold is assessed in table 11 and around the 5% threshold in table 12. Many baseline covariates are inputs to the PREDICT model and therefore highly correlated with the assignment variable because PREDICT chemotherapy benefit scores are a complex function of these inputs. PREDICT 10-year survival scores and the prognostic input variables are significantly different on either side of threshold. This is as expected and does not invalidate the RDD method. Two important baseline covariates that are not inputs to the PREDICT model are Charlson comorbidity and inpatient bed days, these are approximately balanced on either side of the threshold. This is supportive of the continuity assumption.

Table 5.1 Baseline covariate balance around 3% threshold

Variable	**Below (-0.5%)**	**Above (+0.5%)**	**Below (-0.5%)**	**Above (+0.5%)**	
	N	**N**	**Mean or %**	**Mean or %**	**Statistical test**
Prognostic – in PREDICT					
Age	2,538	2,307	61.1	60.6	
Screen detected (%)	2,536	2,305	24.3	19.8	
Tumour size	2,538	2,307	23.4	25.1	
Number of nodes positive	2,538	2,307	0.7	1	
Grade 1 (%)	2,524	2,293	1.9	1.2	
2			45	36	
3			53.1	62.8	
ER status + (%)	2,538	2,307	95.4	86.8	
PREDICT 10-year survival	2,538	2,307	66.1	62.9	T test P < 0.0001
Prognostic – Not in PREDICT					

Charlson comorbidity (%)	2,538	2,307	6.7	6.2	Pearson chi-sq. P = 0.410
Inpatient bed days, all 5 years prior	2,538	2,307	2.7	2.5	T test P = 0.4497

Table 5.2 Baseline covariate balance around 5% threshold

Variable	**Below (-0.5%)**	**Above (+0.5%)**	**Below (-0.5%)**	**Above (+0.5%)**	
	N	**N**	**Mean or %**	**Mean or %**	**Statistical test**
Prognostic – in PREDICT					
Age	1,856	1,363	61.9	62.5	
Screen detected (%)	1,847	1,346	15	14.4	
Tumour size	1,856	1,363	27.9	30.9	
Number of nodes positive	1,856	1,363	1.8	2.7	
Grade 1 (%)	1,849	1,360	0.5	0.2	
2			22.6	22.6	
3			76.9	77.2	
ER status + (%)	1,856	1,363	48.7	52.8	
PREDICT 10-year survival	1,856	1,363	53.5	48.3	T test P < 0.0001
Prognostic – Not in PREDICT					
Charlson comorbidity (%)	1,856	1,363	6.3	7	Pearson chi-sq. P = 0.5667

Inpatient bed days, all 5 years prior	1,856	1,363	3.5	3	T test P = 0.3130

Figure 5.3 - Binned scatterplot (simple) PREDICT chemotherapy benefit score and chemotherapy use, trial represented

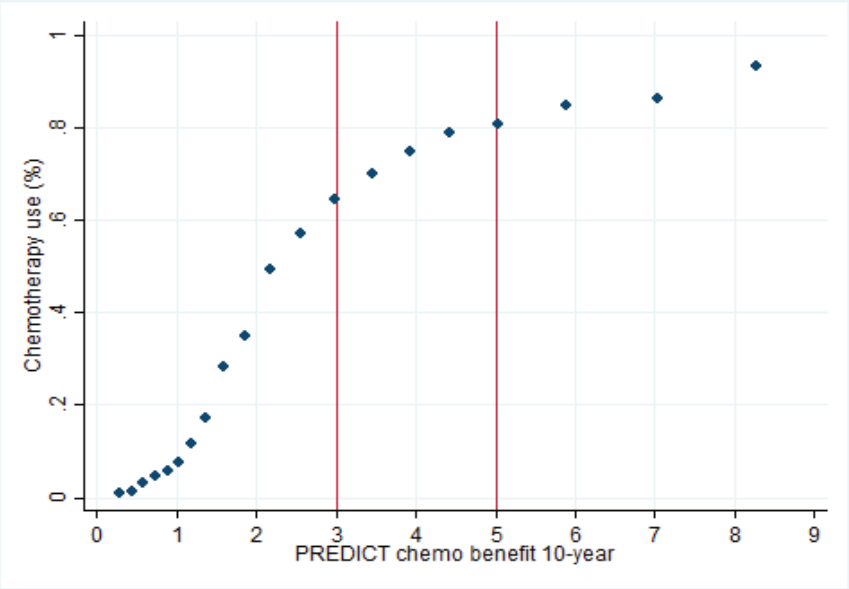
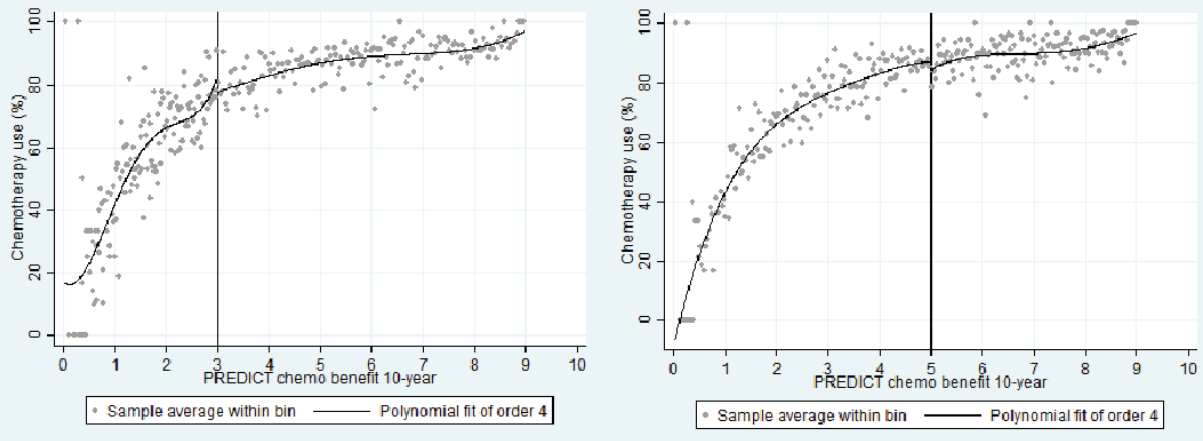


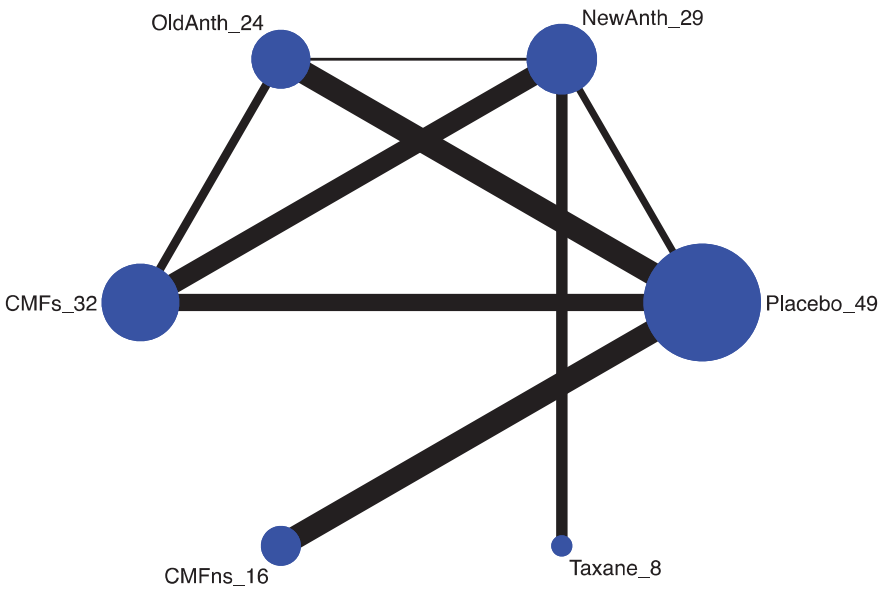
Figure 5.4 - Binned scatterplot (WIMSE) - PREDICT chemotherapy benefit and chemotherapy use, trial represented, 3% and 5% thresholds



6 – Detailed of randomised trial meta-analysis

Network meta-analysis of the EBCTCG- included trials

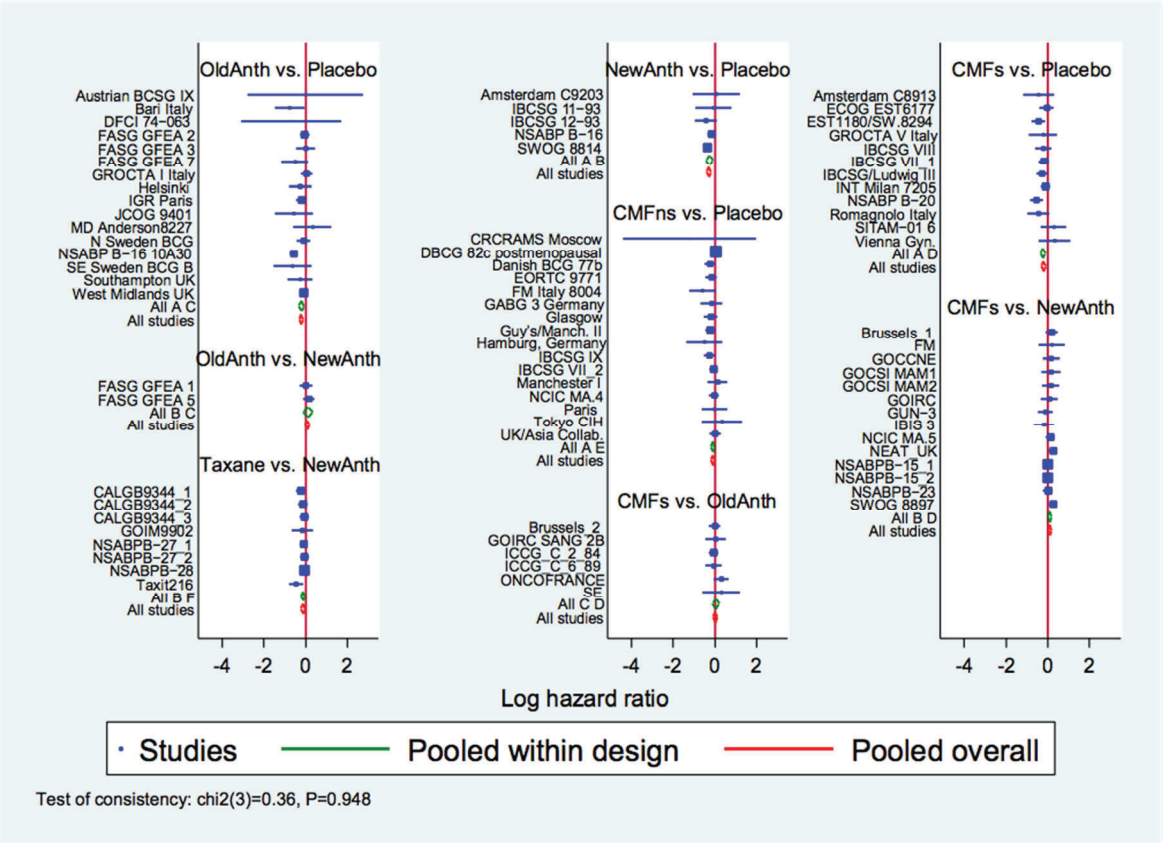
Figure 6.1 Network of the EBCTCG trials



OldAnth = “Anthracycline dose/cycle < A60 or E90,”
NewAnth = “Anthracycline dose/cycle exactly A60 or E90”
CMFs = “Standard CMF (or near-standard CMF) regimens”
CMFns = “Other CMF regimens”
Numbers indicate total number of comparisons
Source for individual trial details: (12)

Breast cancer mortality

Figure 6.2 Forest plot for EBCTCG trials with breast cancer mortality as outcome



Note: Outcomes are treated as binary (Deaths/Women), i.e. making the assumption that the follow-up time is the same in both treatment arms for each trial. Based on the information reported in the original meta-analysis, this was not the case.

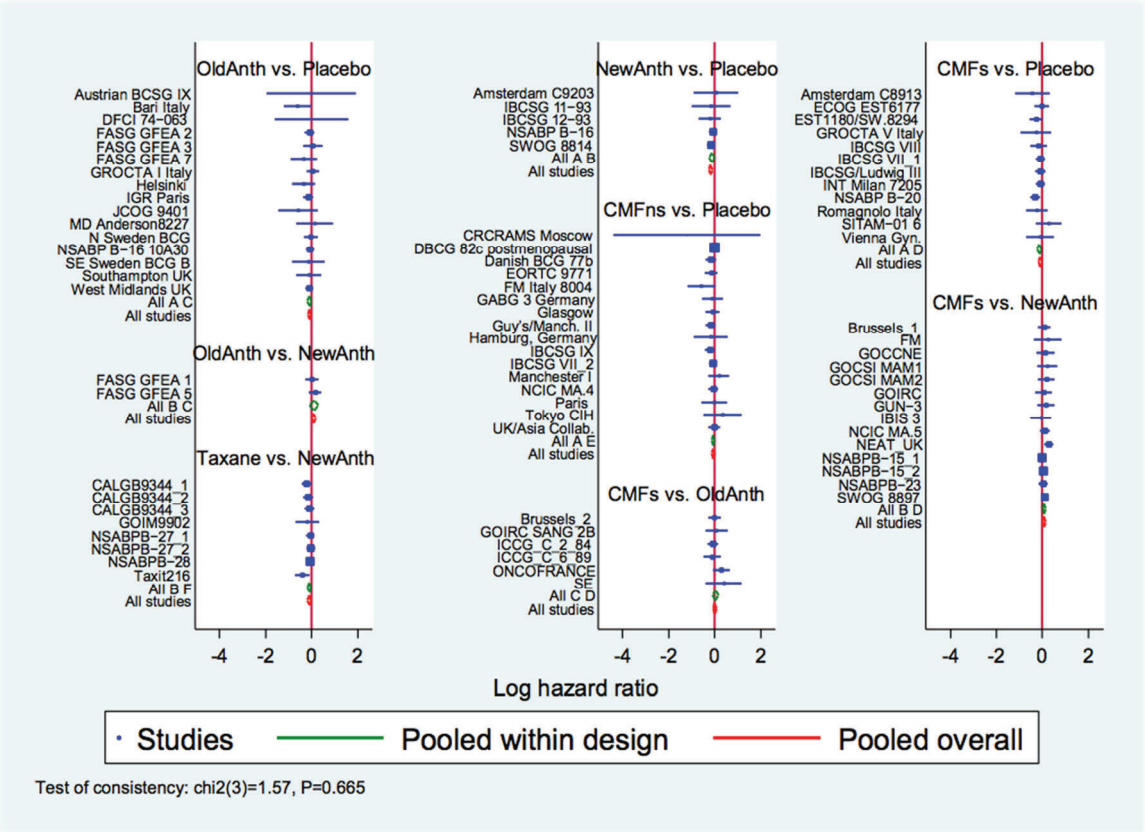
Using placebo the reference and including both direct and indirect evidence, we obtained the following effect estimates for each treatment:

Table 6.1 Hazard ratio in breast cancer mortality for different breast cancer treatments vs placebo

	HR	CI
NewAnth	0.75	0.68; 0.82
OldAnth	0.81	0.74; 0.88
CMF standard	0.82	0.75; 0.89
CMF non-standard	0.91	0.83; 0.99
Taxane	0.66	0.58; 0.76

All-cause mortality

Figure 6.3 Forest plot for EBCTCG trials with all-cause mortality as outcome



Using placebo the reference and including both direct and indirect evidence, we obtained the following effect estimates for each treatment:

Table 6.2 Hazard ratio in all-cause mortality for different breast cancer treatments vs placebo

	HR	CI
NewAnth	0.84	0.78; 0.91
OldAnth	0.90	0.84; 0.97
CMF standard	0.91	0.85; 0.97
CMF non-standard	0.94	0.88; 1.00
Taxane	0.76	0.69; 0.84

Forest plots for individual RWE methods combined with randomised data

Figures 6.4-6.8 TR, breast cancer mortality

Figure 6.4 RA

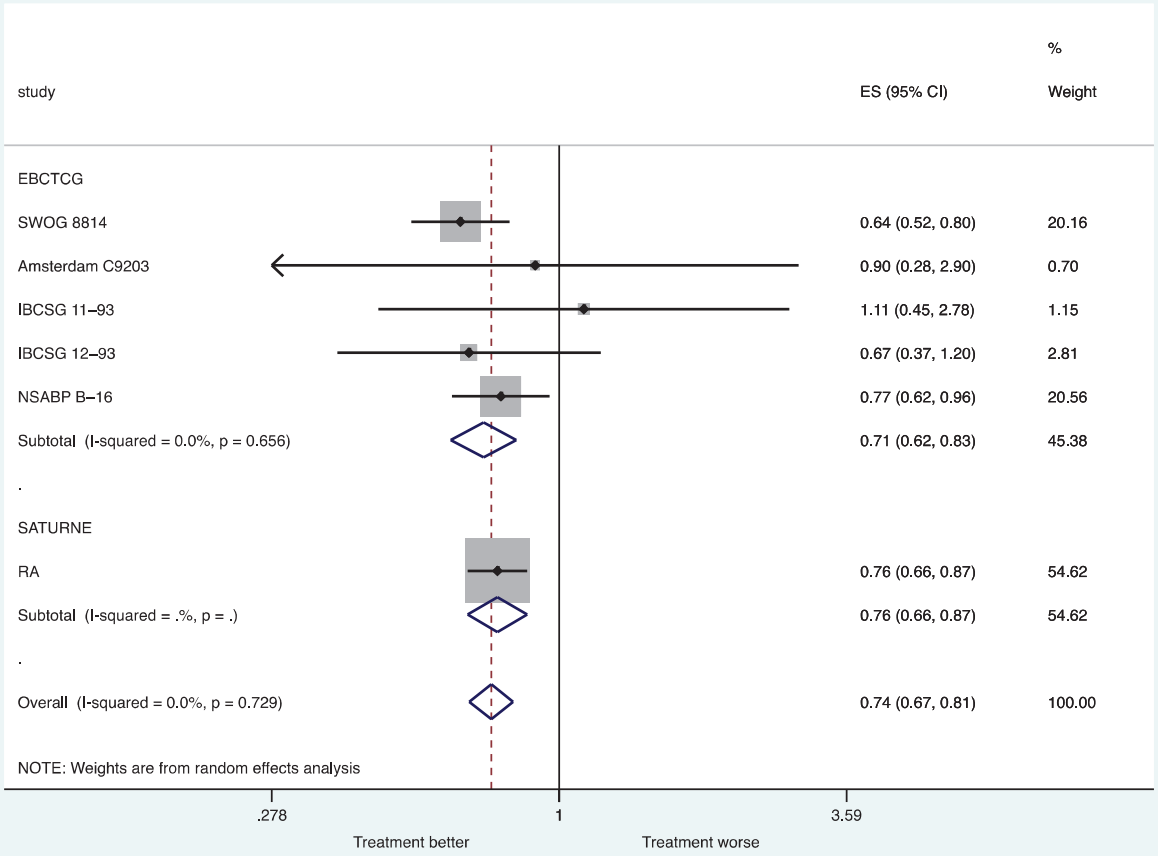


Figure 6.5 PSM Cox

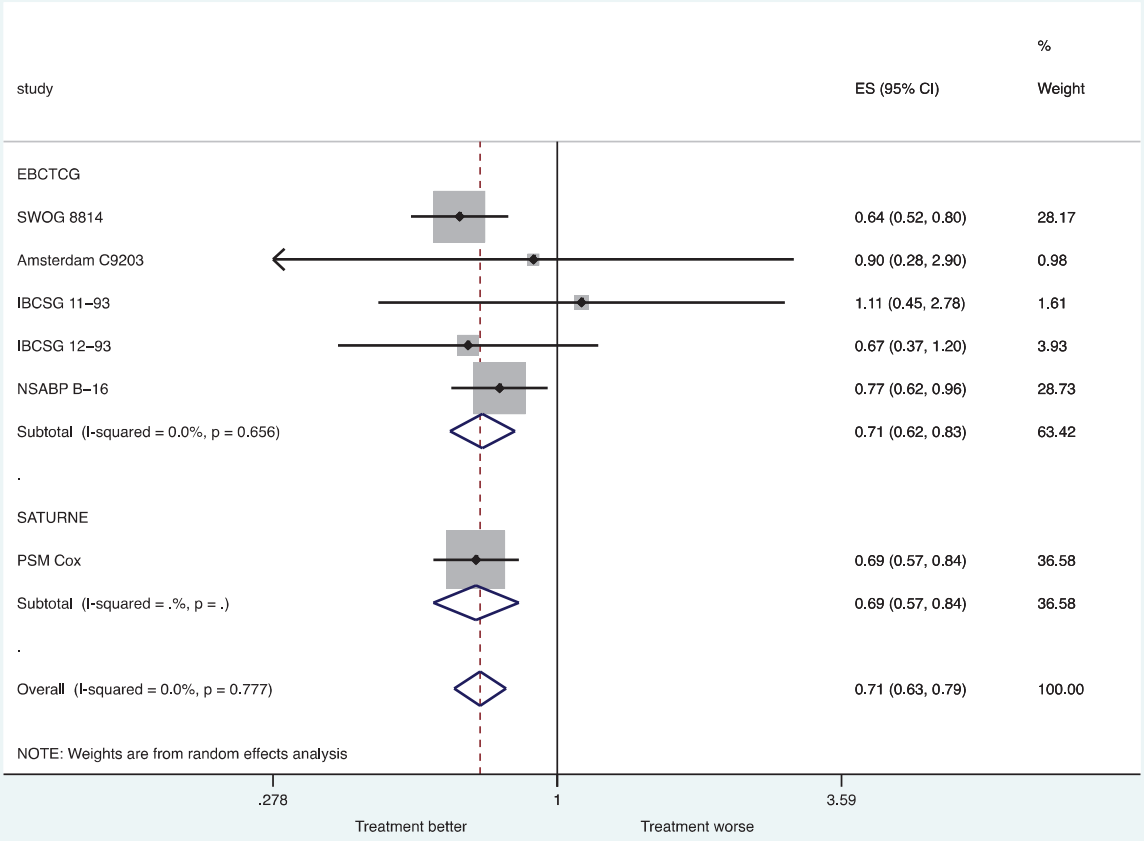


Figure 6.6 PSM LR

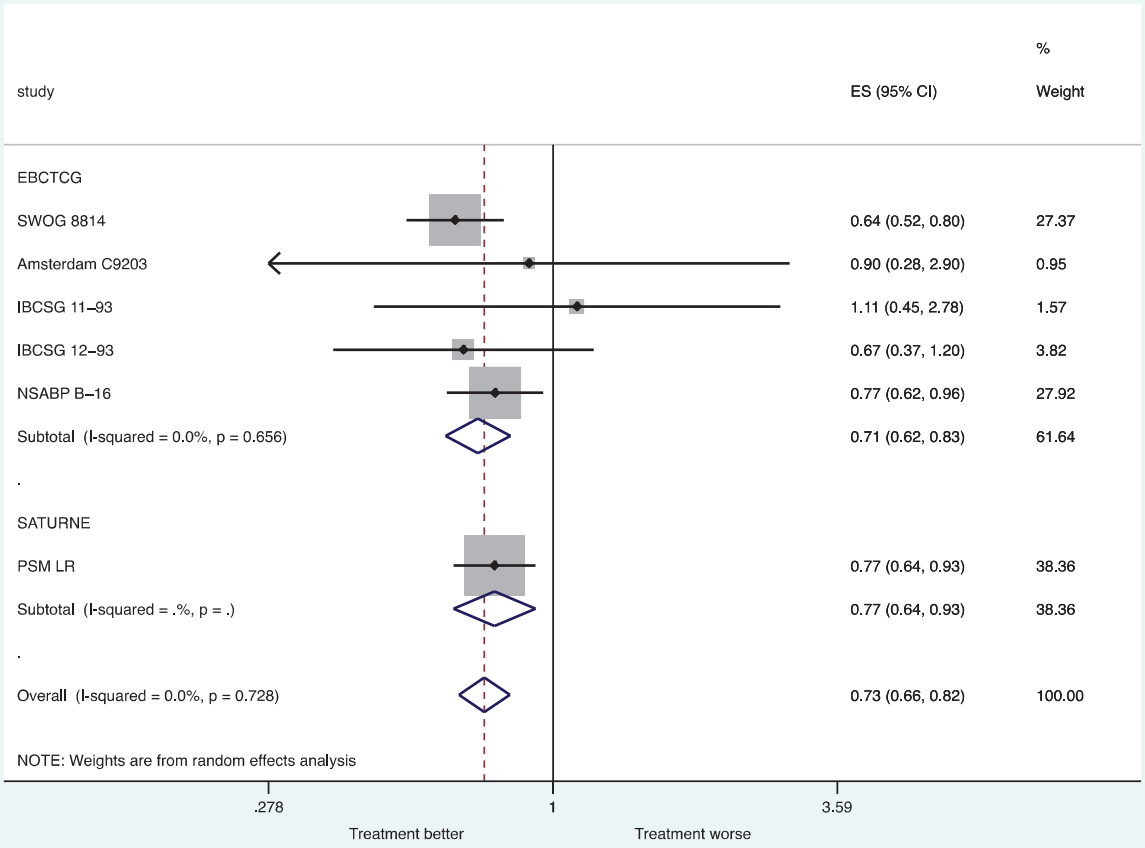


Figure 6.7 IV1

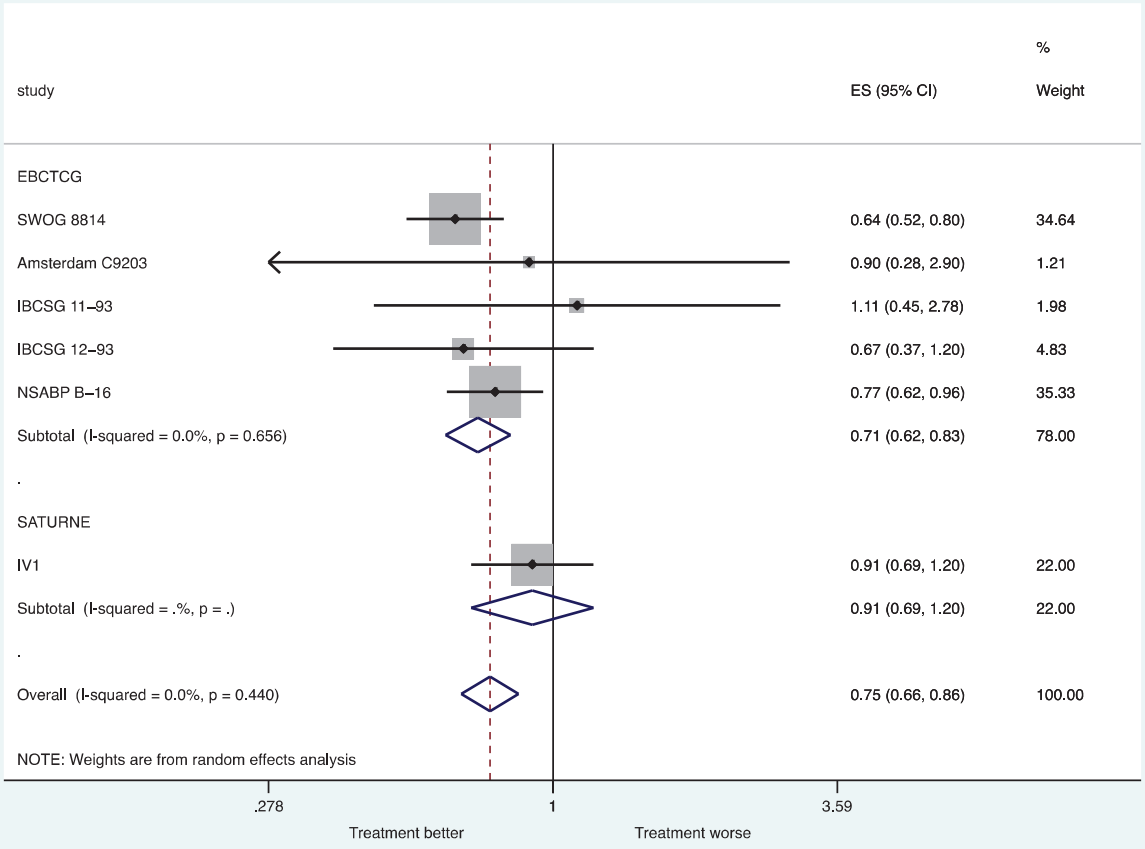
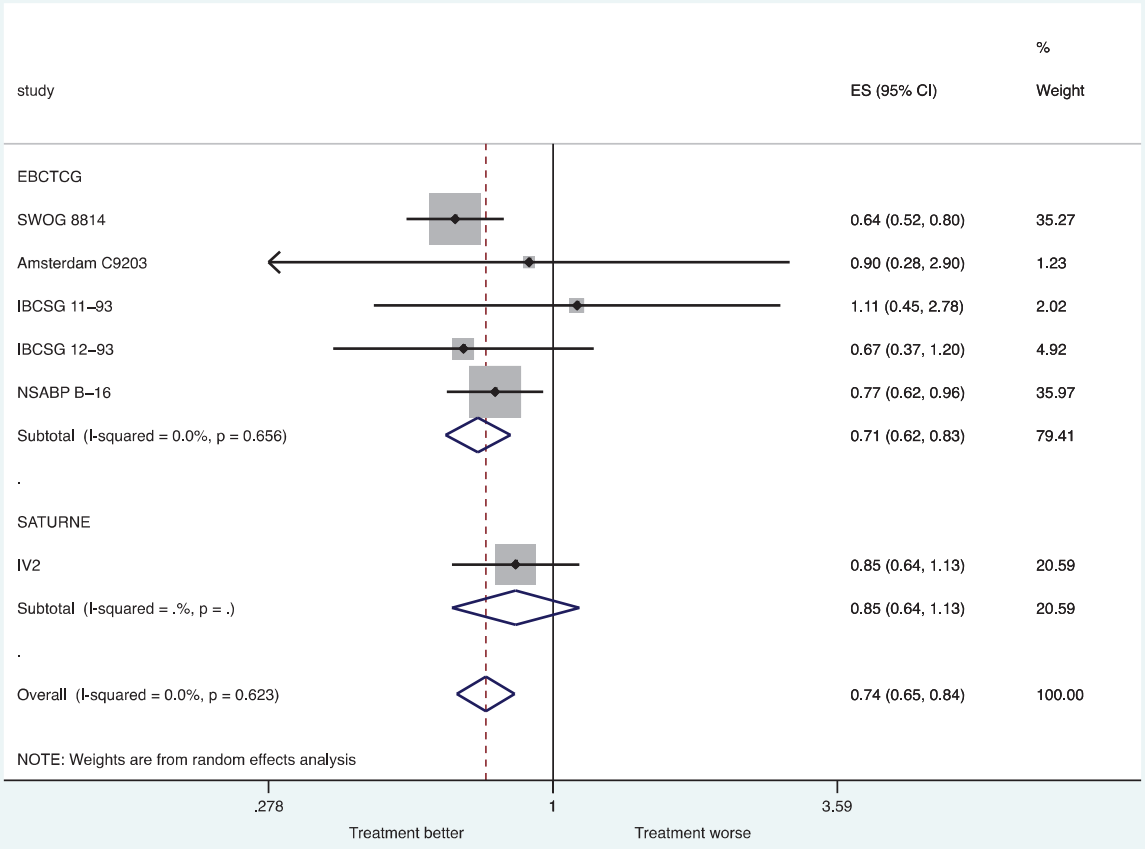


Figure 6.8 IV2



Figures 6.9-6.13 Full cohort, breast cancer mortality

Figure 6.9 RA

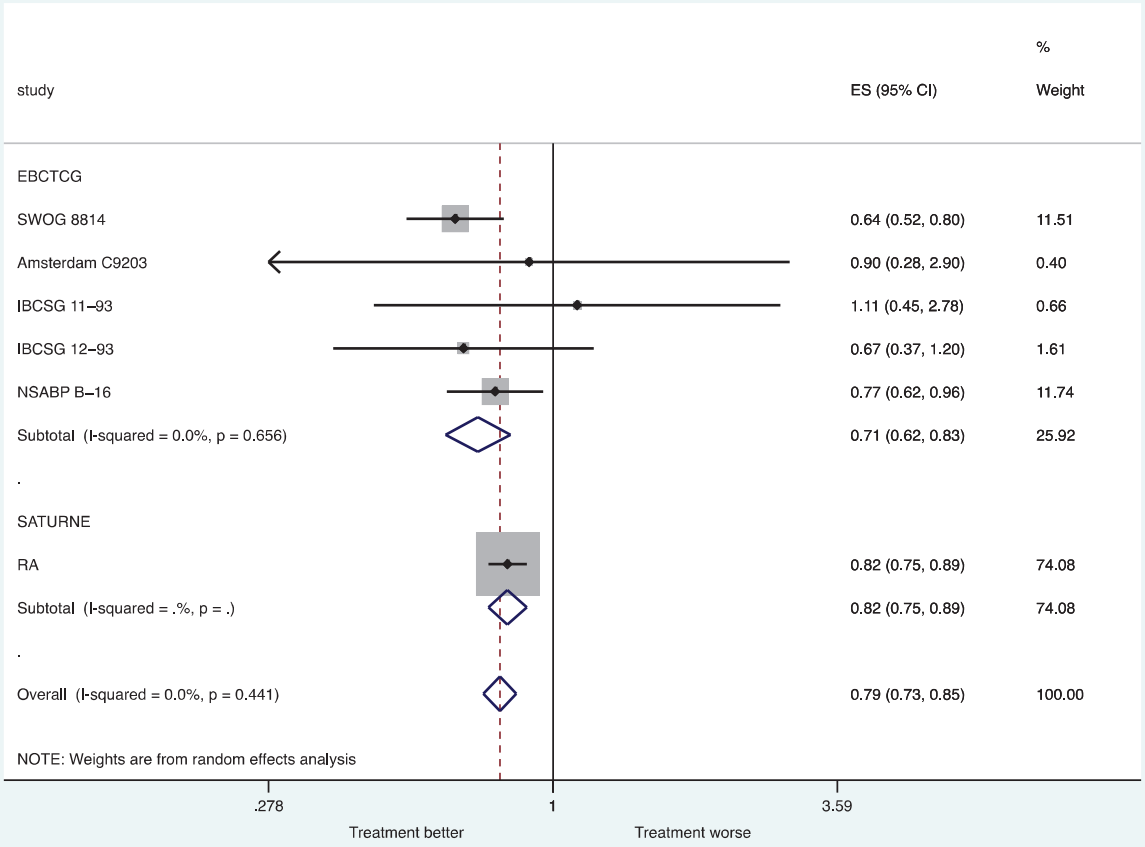


Figure 6.10 PSM Cox

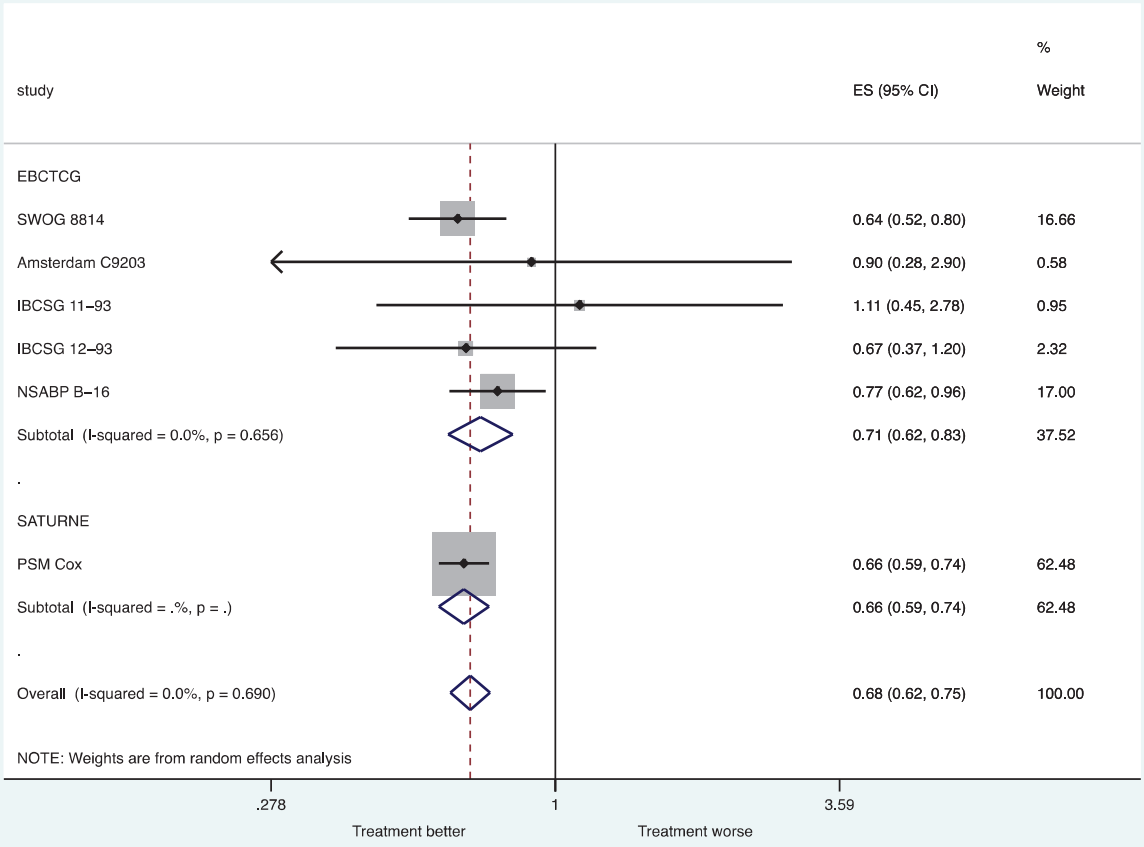


Figure 6.11 PSM LR

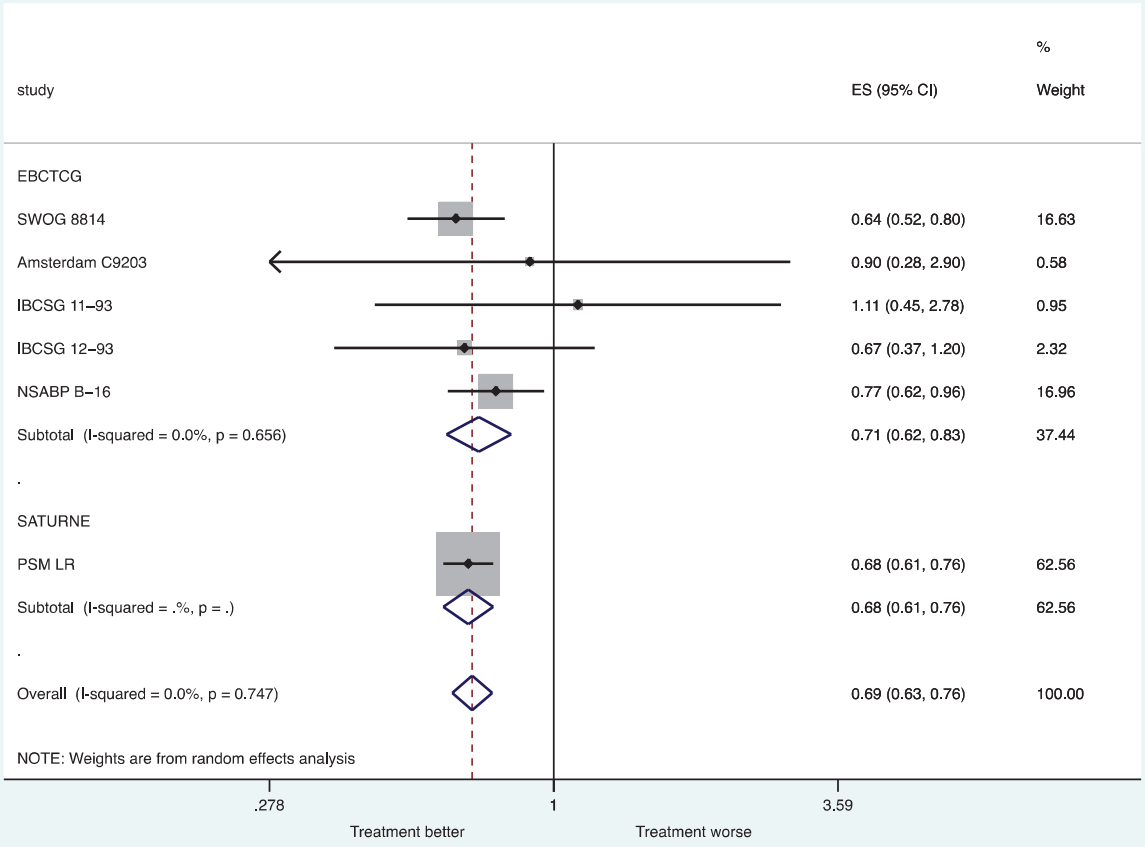


Figure 6.12 IV1

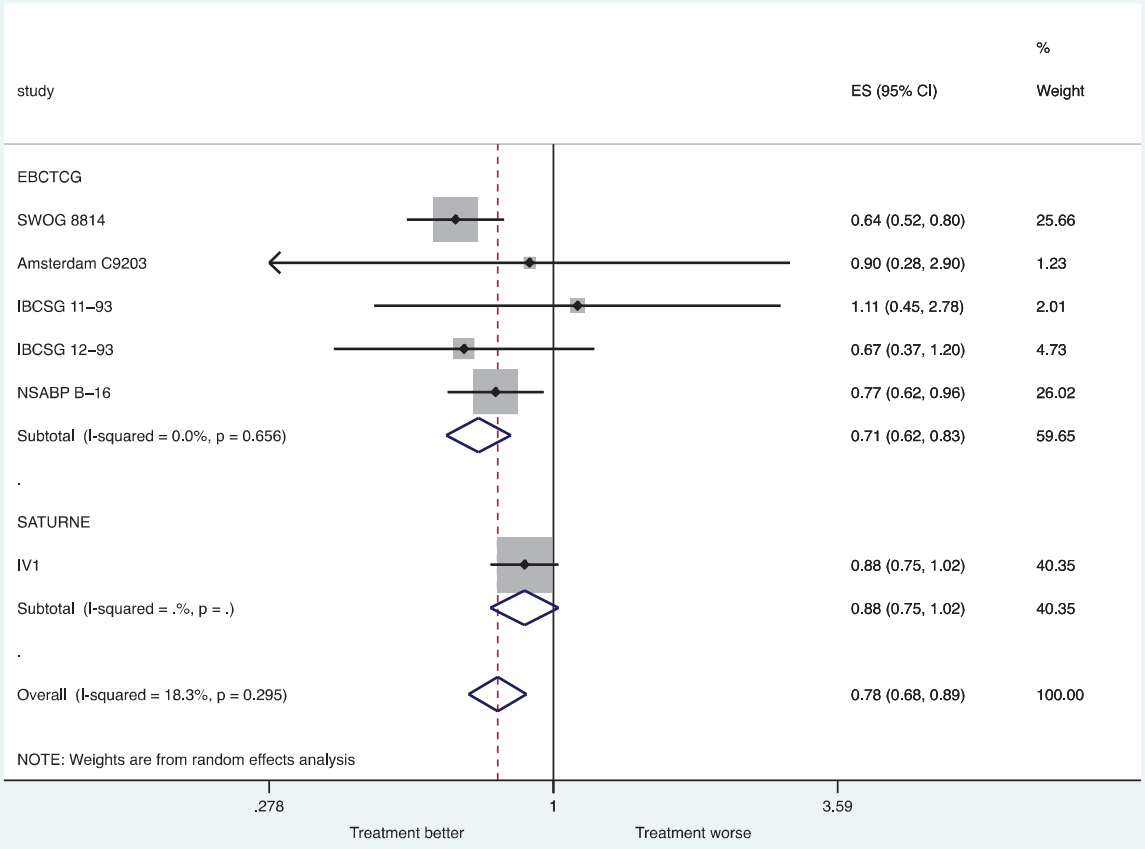
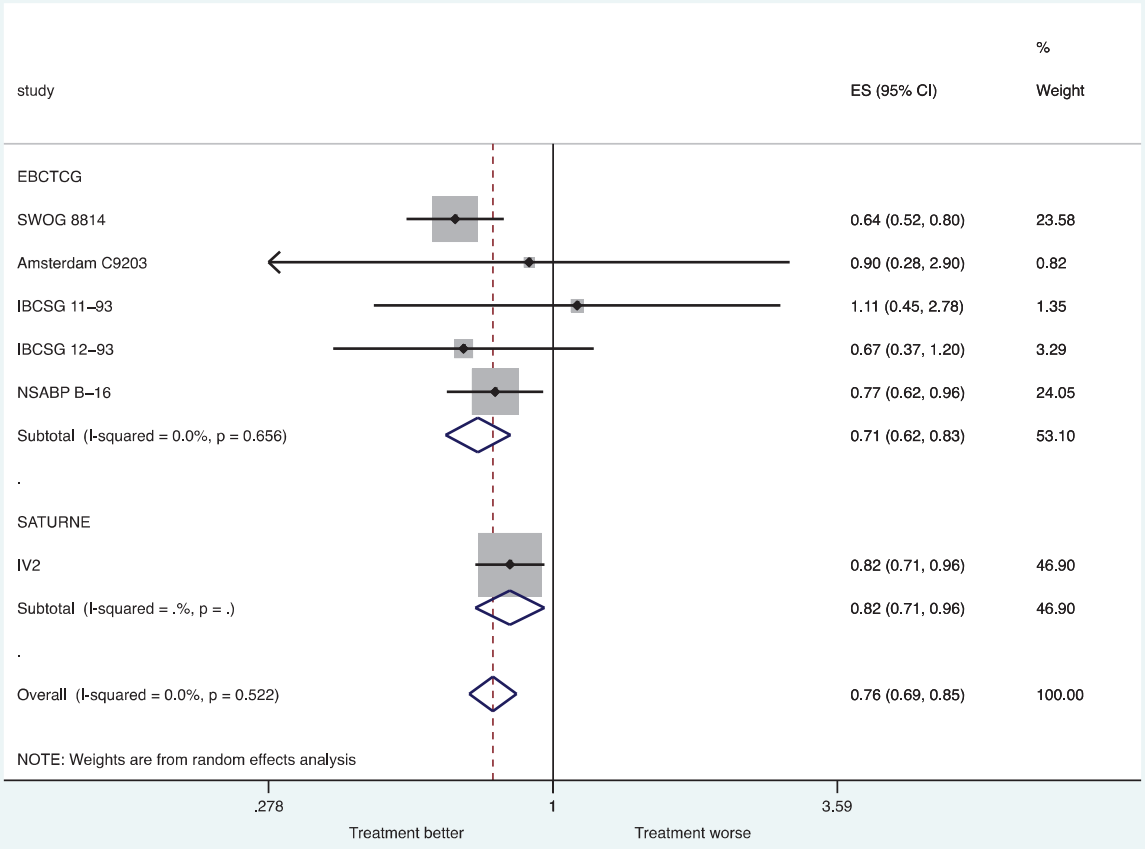


Figure 6.13 IV2



Figures 6.14-6.18 TR, all-cause mortality

Figure 6.14 RA

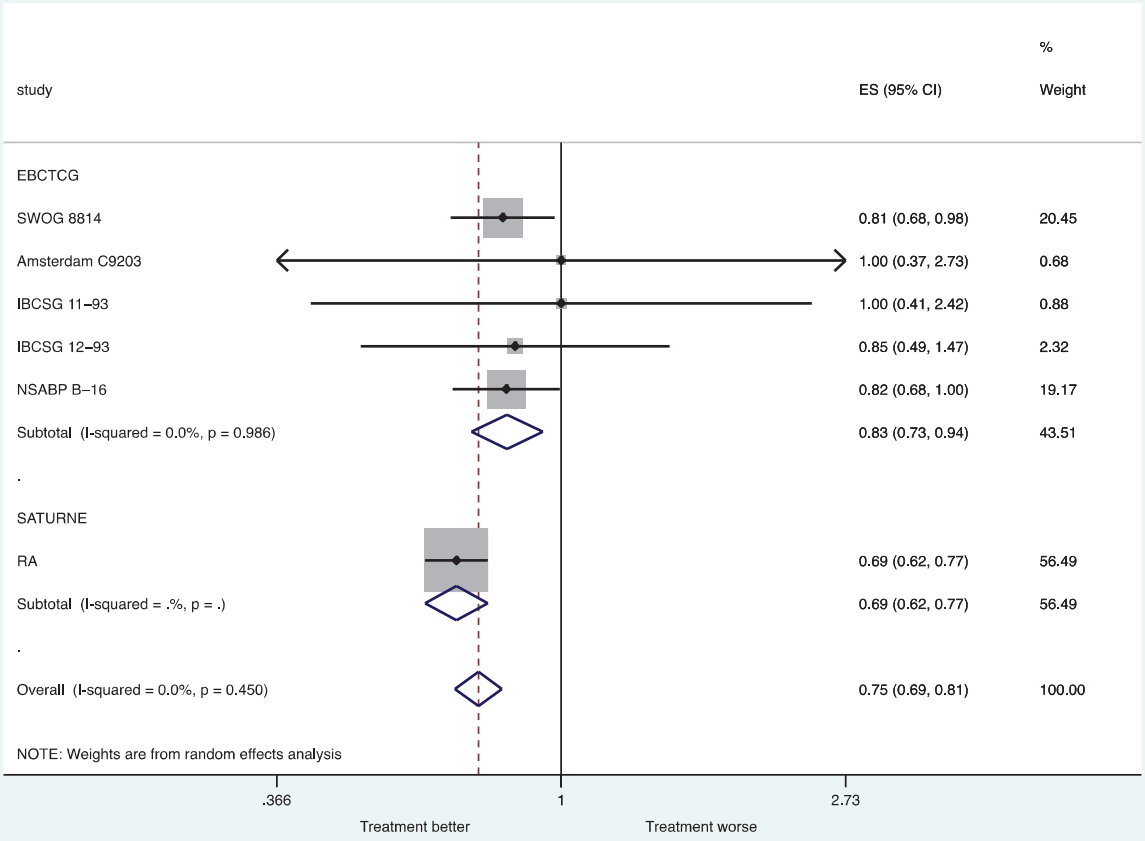


Figure 6.15 PSM Cox

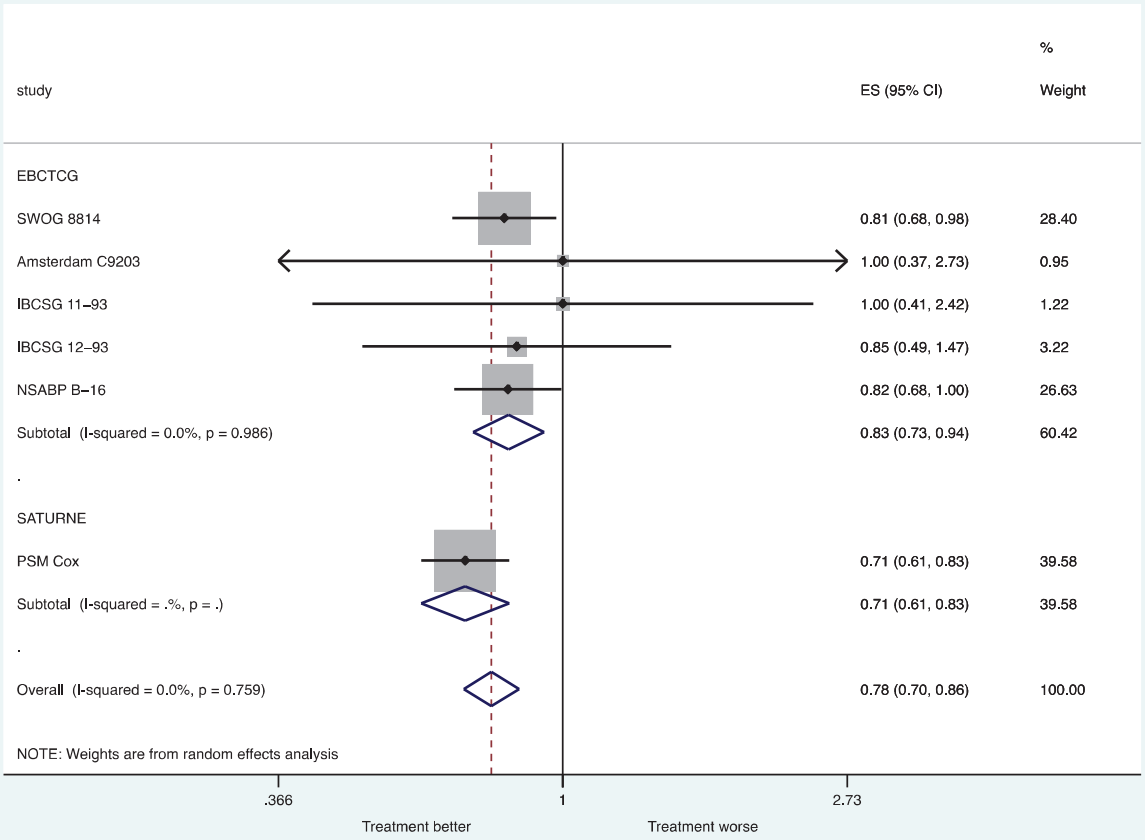


Figure 6.16 PSM LR

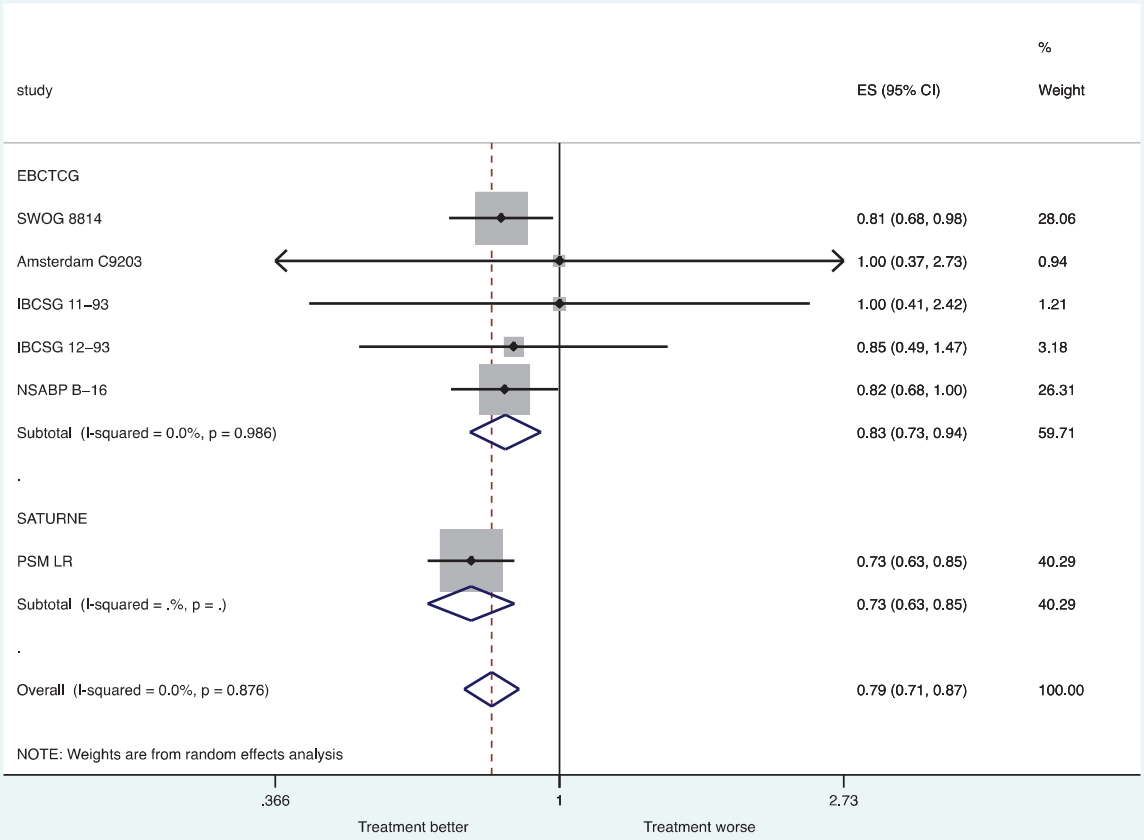


Figure 6.17 IV1

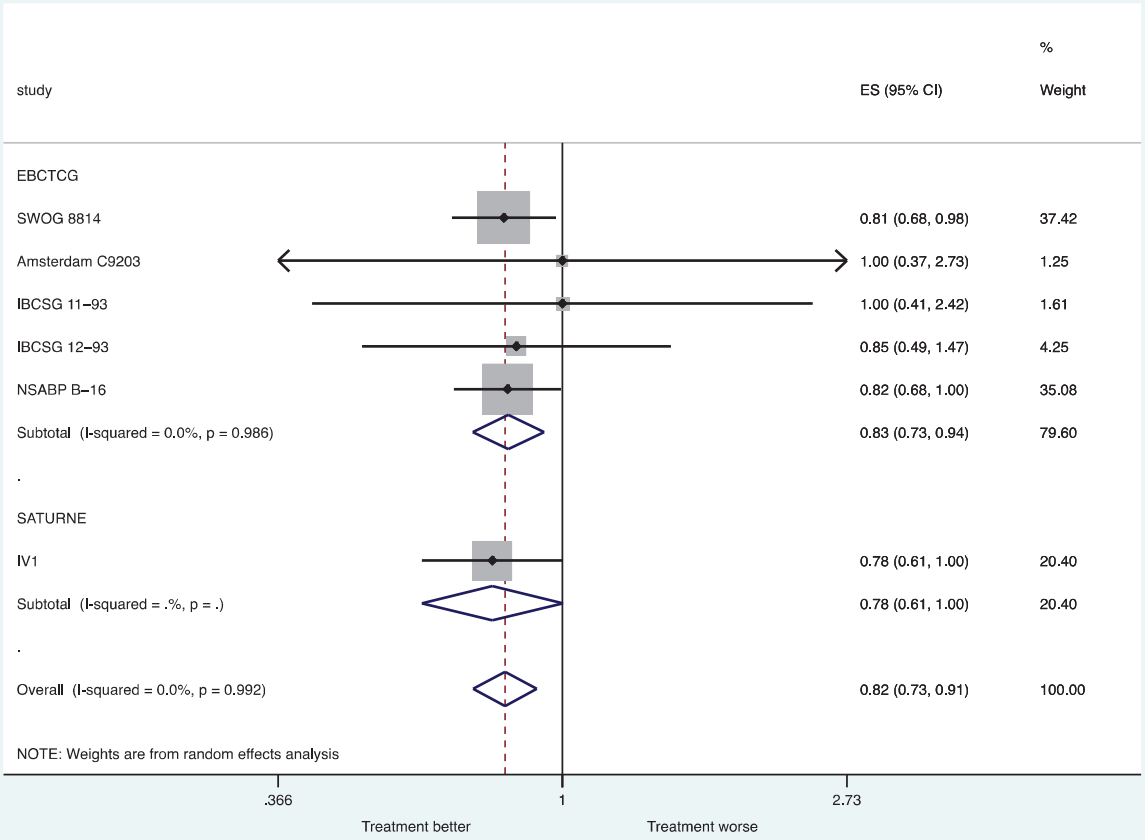
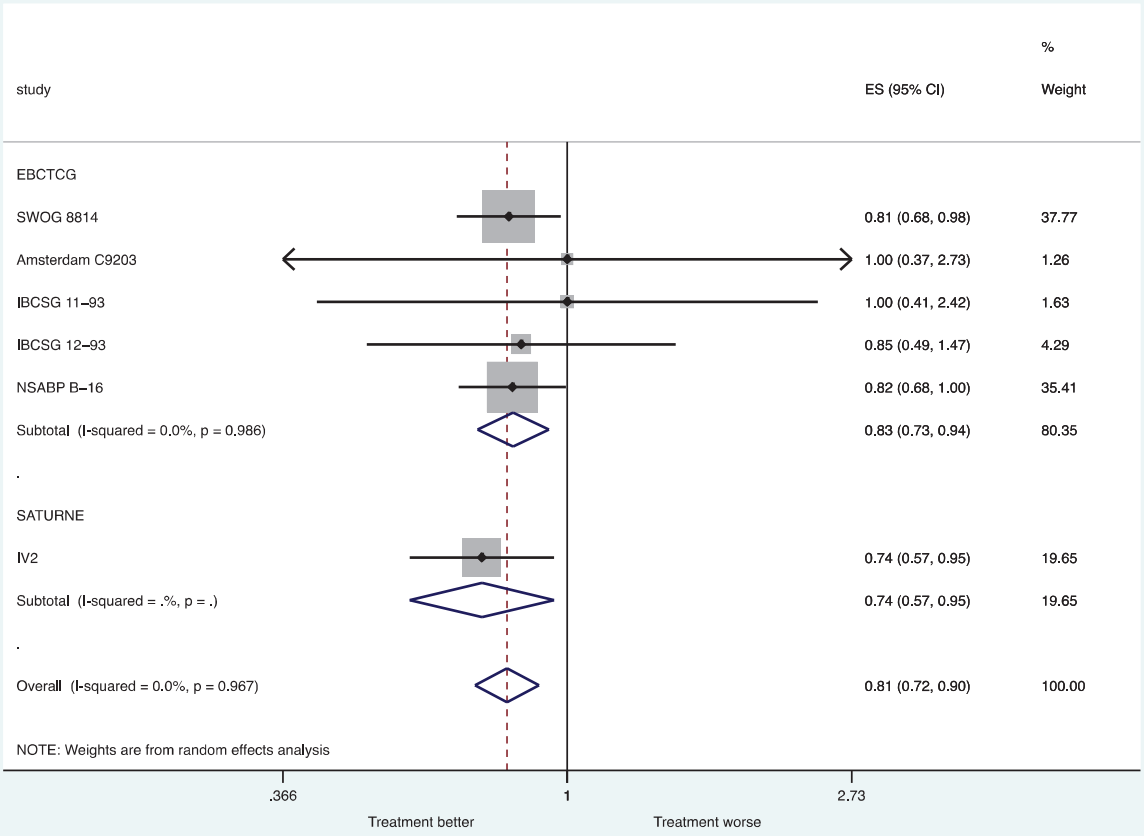


Figure 6.18 IV2



Figures 6.19-6.13 full cohort, all-cause mortality

Figure 6.19 RA

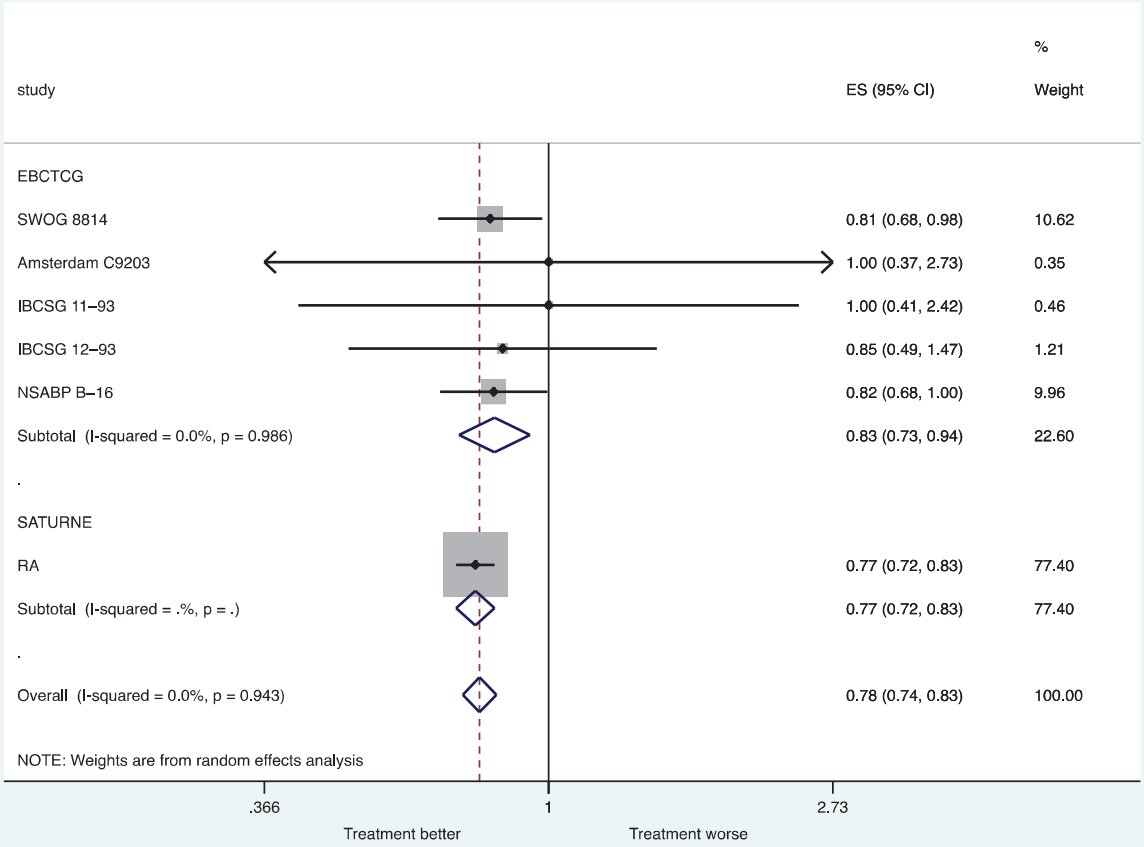


Figure 6.20 PSM Cox

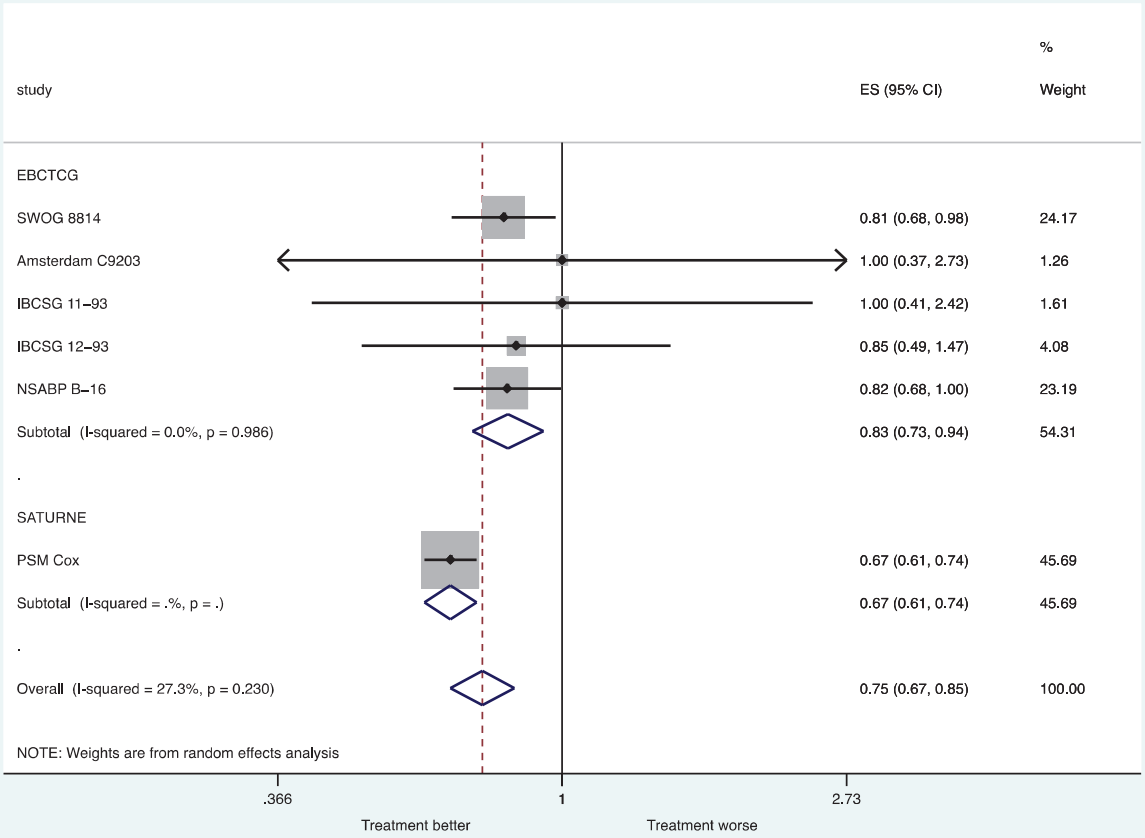


Figure 6.21 PSM LR

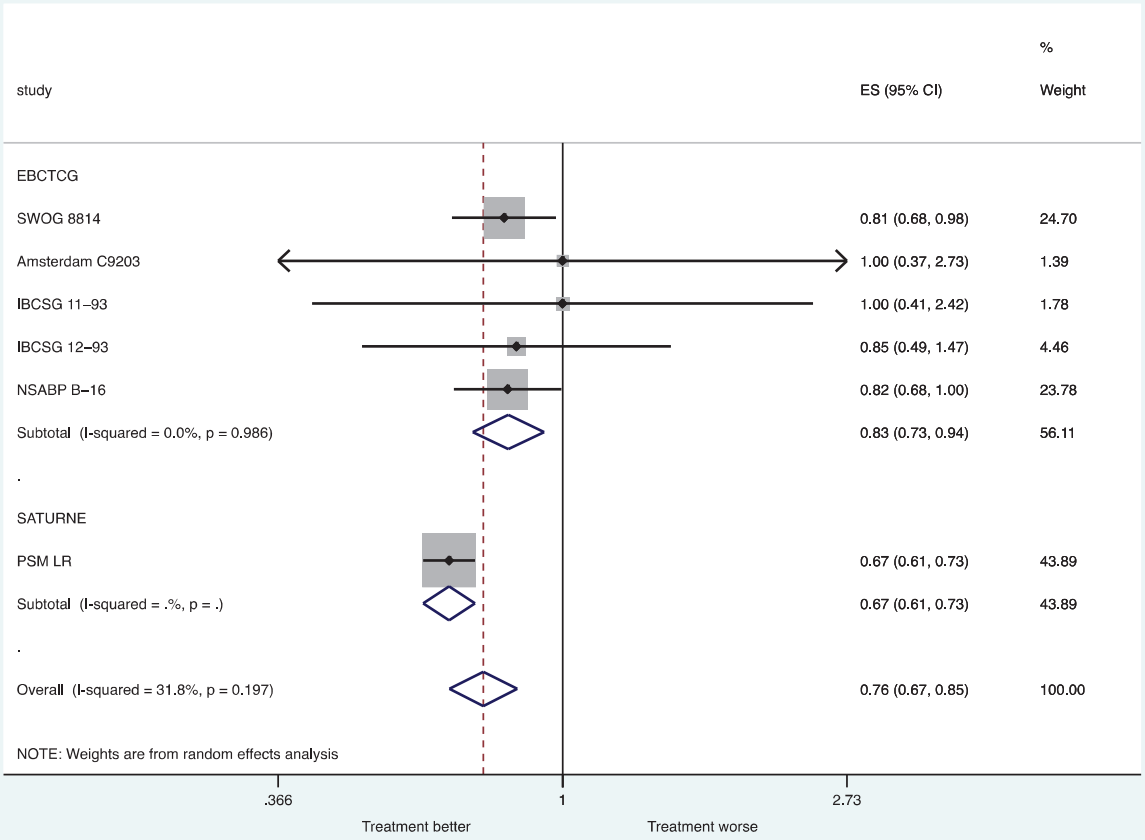


Figure 6.22 IV1

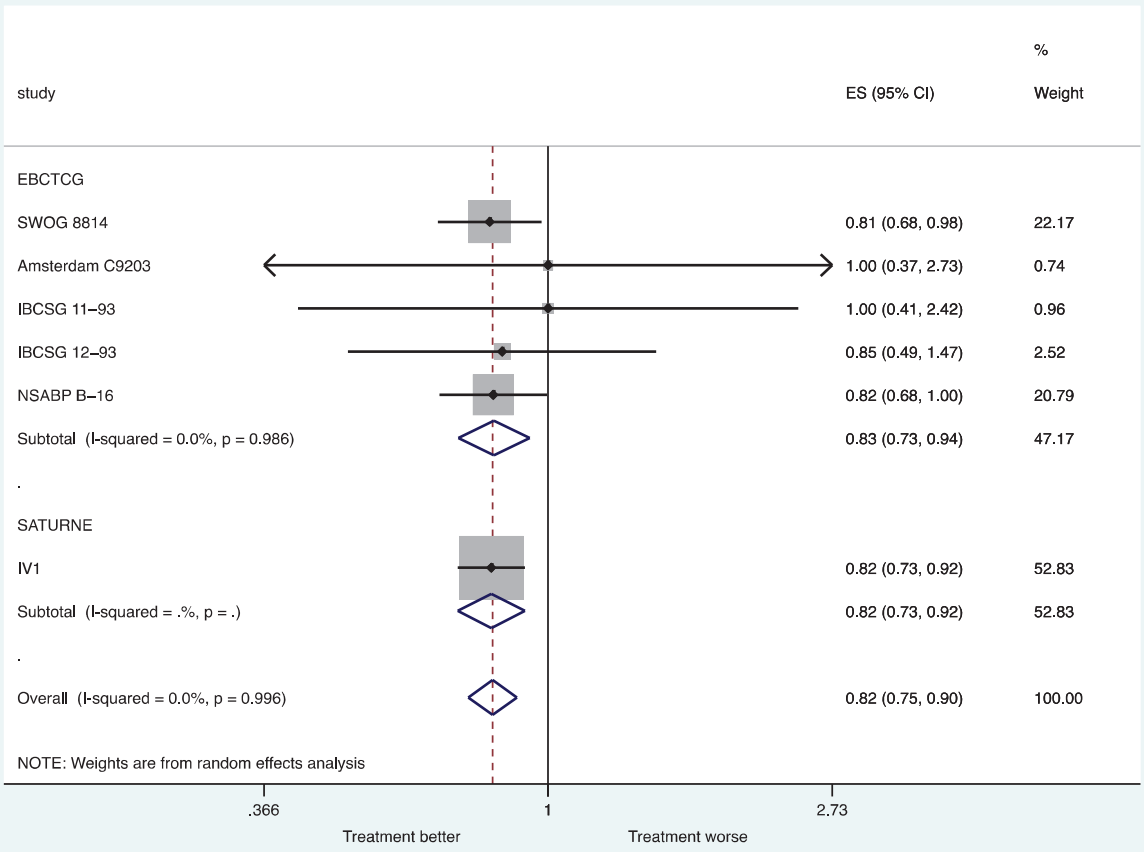
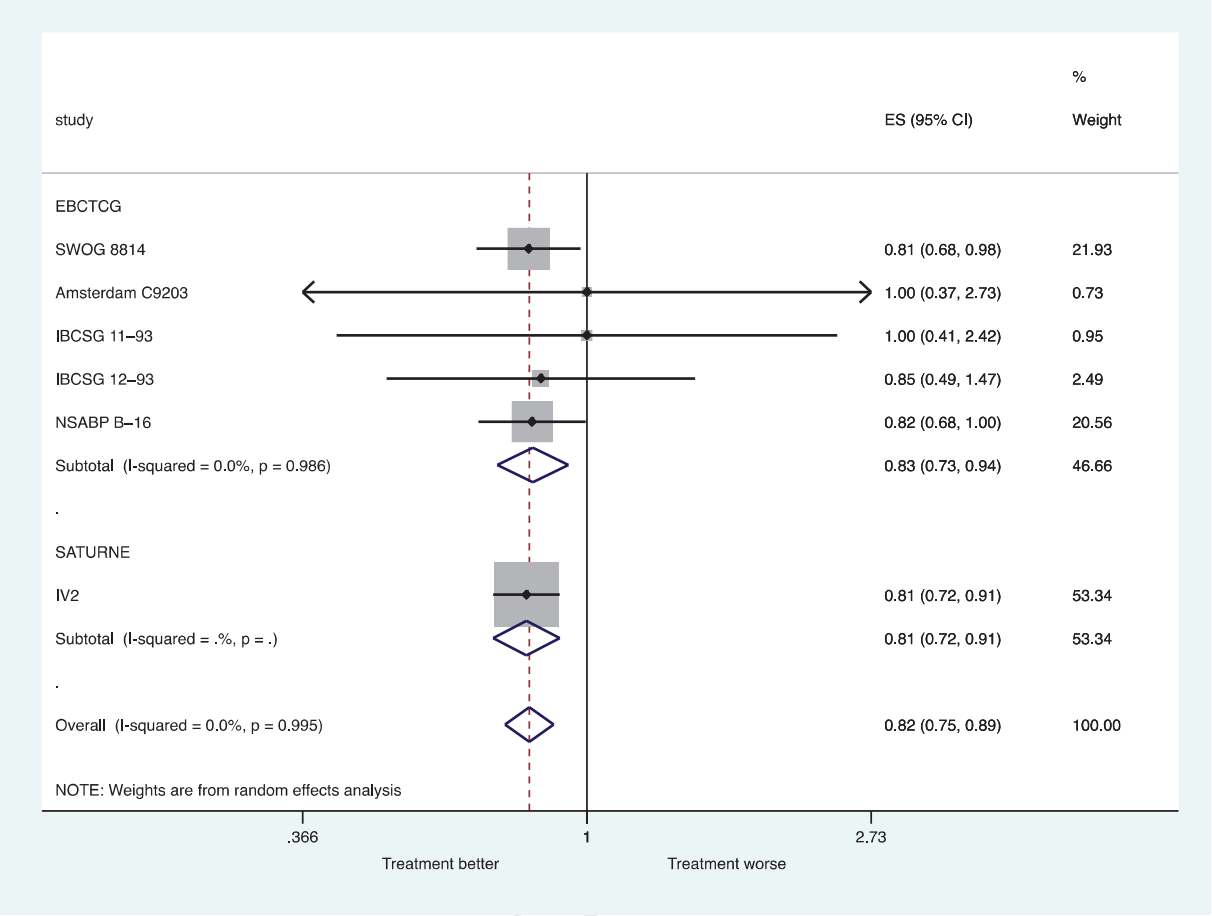


Figure 6.23 IV2



Appendix References

1. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a Combined Comorbidity Index. Journal of Clinical Epidemiology. 1994;47(11):1245-51.
2. Wishart GC, Bajdik CD, Dicks E, Provenzano E, Schmidt MK, Sherman M, et al. PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. Br J Cancer. 2012;107(5):800-7.
3. Candido Dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res. 2017;19(1):58.
4. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res. 2010;12(1):R1.
5. Wishart GC, Rakha E, Green A, Ellis I, Ali HR, Provenzano E, et al. Inclusion of KI67 significantly improves performance of the PREDICT prognostication and prediction model for early breast cancer. Bmc Cancer. 2014;14.
6. Down SK, Lucas O, Benson JR, Wishart GC. Effect of PREDICT on chemotherapy/trastuzumab recommendations in HER2-positive patients with early-stage breast cancer. Oncol Lett. 2014;8(6):2757-61.
7. Scottish Intercollegiate Guidelines Network (SIGN). SIGN 134: Treatment of primary breast cancer. Edinburgh; 2013.

8. Williams C, Brunskill S, Altman D, Briggs A, Campbell H, Clarke M, et al. Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy. Health Technol Asses. 2006;10(34):1-+.
9. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150-61.
10. Berry G, Kitchin RM, Mock PA. A comparison of two simple hazard ratio estimators based on the logrank test. Stat Med. 1991;10(5):749-55.
11. Group EBCTC. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. New Engl J Med. 1988;319(26):1681-92.
12. Early Breast Cancer Trialists' Collaborative Group. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100 000 women in 123 randomised trials. The Lancet. 2012;379(9814):432-44.
13. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiology and drug safety. 2011;20(3):317-20.
14. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J Health Econ. 2008;27(3):531-43.
15. Calonico S, Cattaneo MD, Titiunik R. Optimal Data-Driven Regression Discontinuity Plots. J Am Stat Assoc. 2016;110(512):1753-69.
16. Wooldridge J. Econometric analysis of cross section and panel data: MIT Press; 2010.
17. Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. Econometrica. 1994;62(2):467-75.
18. Moscoe E, Bor J, Barnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. J Clin Epidemiol. 2015;68(2):122-33.
19. Geneletti S, O'Keeffe AG, Sharples LD, Richardson S, Baio G. Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. Stat Med. 2015;34(15):2334-52.
20. McCrary J. Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics. 2008;142(2):698-714.

CRediT Author Statement

Ewan Gray: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Project Administration. **Joachim Marti**: Conceptualization, Methodology, Writing – Review and Editing. **David H Brewster**: Conceptualization, Methodology, Writing – Review and Editing . **Jeremy C Wyatt**: Conceptualization, Methodology, Writing – Review and Editing. **Romain Piaget-Rossel**: Methodology, Software, Formal Analysis, Writing – Review and Editing. **Peter S Hall**: Conceptualization, Methodology, Writing – Review and Editing, Supervision.

What is new?

- Regression adjustment, propensity score matching and instrumental variables were feasible methods for estimating the effectiveness of adjuvant chemotherapy in early stage breast cancer while regression discontinuity design was not.
- Estimates of treatment effectiveness were similar between Real world evidence (RWE) methods and a meta-analysis of randomised trials for breast cancer mortality but not for all-cause mortality.
- RWE should be interpreted cautiously, in the context of the available RCT evidence and with consideration of alternative methods that can be implemented using observational data.

Author Declaration of interests: “Feasibility and results of four real-world evidence methods for estimating the effectiveness of adjuvant chemotherapy in early stage breast cancer”

Ewan Gray, Joachim Marti, David H Brewster, Jeremy C Wyatt, Romain Piaget-Rossel and Peter S Hall

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.