



Published in final edited form as:

J Clin Epidemiol. 2010 January ; 63(1): 64–74. doi:10.1016/j.jclinepi.2009.03.001.

A most stubborn bias: No adjustment method fully resolves confounding by indication in observational studies

Jaclyn Lee Fong Bosco,

Boston University School of Medicine, Department of Medicine, Geriatrics Section, Boston, MA,
Boston University School of Public Health, Department of Epidemiology, Boston, MA

Rebecca A. Silliman,

Boston University School of Medicine, Department of Medicine, Geriatrics Section, Boston, MA,
Boston University School of Public Health, Department of Epidemiology, Boston, MA

Soe Soe Thwin,

Boston University School of Medicine, Department of Medicine, Geriatrics Section, Boston, MA

Ann M. Geiger,

Wake Forest University School of Medicine, Division of Public Health Sciences, Winston-Salem, NC

Diana S. M. Buist,

Group Health Center of Health Studies, Seattle, WA

Marianne N. Prout,

Boston University School of Public Health, Department of Epidemiology, Boston, MA

Marianne Ulcickas Yood,

Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT

Reina Haque,

Kaiser Permanente Southern California, Pasadena, CA

Feifei Wei, and

HealthPartners Research Foundation, Minneapolis, MN

Timothy L. Lash

Boston University School of Public Health, Department of Epidemiology, Boston, MA, Boston University School of Medicine, Department of Medicine, Boston, MA

Abstract

Objective—To evaluate the effectiveness of methods that control for confounding by indication, we compared breast cancer recurrence rates among women receiving adjuvant chemotherapy versus those who did not.

Study Design and Setting—In a medical record review-based study of breast cancer treatment in older women (n=1798) diagnosed 1990-1994, our crude analysis suggested adjuvant

Correspondent author: Jaclyn Lee Fong Bosco, Boston University Medical Center, 88 East Newton Street, Robinson 2, Boston, MA 02118, (t) 617-638-8998, (f) 617-638-8387. (e) jfong@bu.edu.

Preliminary results of this research were presented at the 40th annual meeting of the Society for Epidemiologic Research (June 2007).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

chemotherapy was positively associated with recurrence [hazard ratio (HR)=2.6 (95% confidence interval (CI)=1.9, 3.5)]. We expected a protective effect, so postulated that the crude association was confounded by indications for chemotherapy. We attempted to adjust for this confounding by restriction, multivariable regression, propensity scores [PS], and instrumental variable [IV] methods.

Results—After restricting to women at high-risk for recurrence (n=946), chemotherapy was not associated with recurrence [HR=1.1 (95% CI=0.7, 1.6)] using multivariable regression. PS adjustment yielded similar results [HR=1.3 (95% CI=0.8, 2.0)]. The IV-like method yielded a protective estimate [HR=0.9; (95% CI=0.2, 4.3)]; however imbalances of measured factors across levels of the IV suggested residual confounding.

Conclusion—Conventional methods do not control for unmeasured factors, which often remain important when addressing confounding by indication. PS and IV analysis methods can be useful under specific situations, but neither method adequately controlled confounding by indication in this study.

Keywords

confounding by indication; propensity score; instrumental variable; non-randomized studies; breast cancer; chemotherapy

What is new?

- **Key Findings:** (a) The implementation of propensity score adjustment does not guarantee comparability between the exposure groups; (b) A strong instrumental variable may be confounded; (c) Restricting to a more homogenous population remains an effective way to control for confounding.
- **What this adds to what was known:** The use of propensity scores and instrumental variable methods is not universally effective in all observational settings.
- **What is the implication, what should change now:** Researchers should understand the limitations and appropriateness of the propensity score and instrumental variable methods in relation to their data before implementing and interpreting results that may be as biased as results generated by conventional methods.

Introduction

Confounding by indication remains an often intractable threat to validity in observational studies.¹ While confounding is best controlled by a randomized design, randomization is not always feasible. For example, patients cannot be randomized to receive placebo when an efficacious therapy is available.² Furthermore, trials often exclude patients with pre-existing conditions,³ particularly older adults.⁴ Non-randomized designs must evaluate the effectiveness of therapies whose efficacy has been established in select groups by clinical trials, but not in broader populations that might react differently to the therapy. For these and other reasons,³ non-randomized studies of therapy effectiveness will remain important.¹ In addition, generalizing results from clinical trials with select patient populations may actually cause harm in the heterogeneous populations treated in clinical practice.⁵

As an example, clinical trials of adjuvant chemotherapy in women aged 40-59 years with early stage breast cancer demonstrate its efficacy with reductions in 5-year mortality between 20% and 40%,⁶ but it is uncertain whether these benefits extend to older women, who bear the

majority burden of breast cancer occurrence.⁷ Non-randomized studies of older women with early stage breast cancer suffered from differences in prognosis between women who received adjuvant chemotherapy and women who did not receive adjuvant chemotherapy,^{8,9} and thus are potentially biased by confounding by indication.

When the validity of a study is threatened by confounding by indication, it is not straightforward to determine which method of adjustment, if any, is most effective in obtaining a valid and precise estimate of effect. Conventional methods to adjust for confounding, such as restriction and multivariable regression, leave residual confounding due to unmeasured factors. Thus propensity score (PS) adjustment and the instrumental variable (IV) approach have become increasingly popular,¹⁰⁻¹³ with the intent to address this residual confounding by simulating a randomized environment. PS adjustment theoretically increases comparability between the comparison groups by creating pseudo randomization of measured confounders.¹⁴ The goal of the IV approach is to reduce confounding by indication through the use of a variable that is associated with the exposure, unrelated to the confounders, and has no direct association with the outcome other than through the exposure.¹⁵ However, several investigators have cautioned that these alternative methods are not universal solutions to the problem of confounding by indication.^{10,11,13,16-18}

Three observational studies used the SEER-Medicare¹⁹ linked dataset and found adjuvant chemotherapy decreased the rate of breast cancer-specific mortality²⁰ and all-cause mortality²⁰⁻²² in older women, with the greatest benefit seen in women with node positive, estrogen receptor negative tumors.^{20,21} Based on these results and those of clinical trials among middle-aged women,⁶ we expect adjuvant chemotherapy to be protective against breast cancer recurrence in older women. With this prior information in mind, we compared methods used to reduce confounding. We implemented restriction, multivariable regression, PS adjustment, and an IV-like method to estimate incidence rates of breast cancer recurrence in women who received adjuvant chemotherapy compared to women who did not, in the Breast Cancer Treatment Effectiveness in Older Women (BOW) cohort.^{8,9,23}

Methods

Study Population

The BOW cohort study was conducted at six integrated healthcare systems that are part of the 14-system consortium of the Cancer Research Network (CRN).²⁴ The overall goal of the CRN is to increase the effectiveness of preventive, curative and supportive interventions for major cancers through a program of collaborative research, and to determine the effectiveness of cancer control interventions that span the natural history of major cancers among diverse populations and health systems. The six systems were Group Health Cooperative, Seattle, WA; Meyers Primary Care Institute/Fallon Community Health Plan, Worcester, MA; Kaiser Permanente Southern California; Lovelace/Sandia Health System, New Mexico; HealthPartners, Minneapolis, MN; and Henry Ford Health System, Detroit, MI. The institutional review boards of each healthcare system and the Boston University Medical Center approved this study.

Detailed data collection methods have been described previously.²³ Briefly, our cohort included women age 65 or older diagnosed with early stage (I to IIB) breast cancer between 1990 and 1994 at one of these six integrated health systems. Women with bilateral cancer or other malignancies except non-melanoma skin cancer were excluded if their diagnosis was within five years before, or 30 days after, their initial breast cancer diagnosis. Our exposure of interest was adjuvant chemotherapy, therefore women who received only a biopsy (n=22), neoadjuvant chemotherapy (n=3), or had implausible chemotherapy start and stop dates

recorded (n=13) were excluded from this analysis. We will refer to this population as the “unrestricted cohort”.

Data collection

Demographic and tumor characteristics, breast cancer treatments, recurrence, and comorbid conditions were collected via medical record reviews conducted up to 10 years post diagnosis. Details of the medical record review are described by Thwin *et al.*²⁶

Analytic variables

Adjuvant chemotherapy—Women who received adjuvant chemotherapy were considered the index group. Among women who received adjuvant chemotherapy, the median length of time to last adjuvant chemotherapy course was 183 days after diagnosis. Type of chemotherapy, start and stop dates, number of courses, and completion were also collected. Women who were not referred, not recommended, refused, or did not receive adjuvant chemotherapy comprised the reference group. Women with no mention of chemotherapy in the medical records were assigned to the reference group.

Follow-up time—We defined the start of follow-up as the date of last adjuvant chemotherapy course (index) or 183 days after diagnosis (reference), and follow-up continued until the diagnosis of breast cancer recurrence, death from any cause, disenrollment from the healthcare system, or the completion of 10 years of follow-up, whichever came first.

Breast cancer recurrence—Breast cancer recurrence was defined as a tumor pathologically or clinically diagnosed during the follow-up period. Tumors that occurred in the same breast as the original tumor or in any lymph node or distant site were classified as a recurrence. Women with recurrence (n=16) or death (n=6) that occurred before the last date of chemotherapy course or before 183 days after diagnosis were excluded from the analyses.

Patient Characteristics—Demographic, tumor, and breast cancer treatment characteristics were considered potential confounders in the association between adjuvant chemotherapy and recurrence. Women were categorized by age at diagnosis (65–69; 70–74; 75–79; ≥80 years old), race/ethnicity (non-Hispanic White; Hispanic and/or Other Race), tumor size (<1; 1 to <2; 2 to <3; ≥3 centimeters [cm]), node positivity (negative [no presence of breast cancer in lymph nodes]; 1–3 positive nodes; ≥4 positive nodes; not determined), histologic grade (well differentiated; intermediate or moderately differentiated; poorly differentiated, undifferentiated, or anaplastic; not determined or stated), primary therapy (breast conserving surgery [BCS] only; BCS plus radiation therapy; mastectomy), estrogen receptor (ER) expression (positive; negative; other), progesterone receptor (PR) expression (positive; negative; other), tamoxifen (prescribed; not prescribed), and baseline Charlson Comorbidity Index score (0; 1; ≥2).²⁷ Women who did not have an axillary lymph node dissection were similar to women who were node negative and the two groups were combined. Women who were recorded as “other” for ER or PR expression were combined with ER positive expression and PR positive expression, respectively. Women who were prescribed tamoxifen or another hormonal agent (n=2) were classified as having received tamoxifen.

Data analysis

Descriptive statistics for demographic, tumor, and treatment characteristics were calculated using univariate statistics. These characteristics were also evaluated as potential confounders of the association between adjuvant chemotherapy and breast cancer recurrence using contingency table analyses.

We compared several methods in their ability to obtain valid and precise results, using the prior from trials of younger breast cancer patients as a guide for the expected direction of the effect. Figure 1 illustrates the analytic samples used for each of the analytic methods described below. All analyses were performed using SAS statistical software version 9.1 (SAS Institute, Cary, North Carolina).

Unadjusted analysis—Using Cox proportional hazards regression on the unrestricted cohort, we estimated the hazard ratio associating receipt of adjuvant chemotherapy versus not receiving adjuvant chemotherapy.

Restricted analysis—Within the unrestricted cohort, we identified a restricted subset of women as high-risk for recurrence using the St. Gallen²⁵ criteria from the calendar time of diagnosis (1992). These criteria combine tumor size, node positivity, histologic grade, and ER and PR expression to identify women who are considered at high-risk for recurrence. A woman was classified as high-risk if she was node positive, or node negative with one of the following three tumor characteristics: (1) poorly differentiated, grade III histology; (2) ER negative and ≥ 1 cm diameter; or (3) ER positive and >2 cm diameter. Using this restricted cohort to reduce confounding, we conducted Cox proportional hazards regression to estimate the association between adjuvant chemotherapy and breast cancer recurrence.

Restriction and multivariable regression—Using the restricted cohort, we adjusted for demographic characteristics (age group, race/ethnicity, healthcare system, baseline Charlson Comorbidity Index score²⁷), tumor characteristics (tumor size, node positivity, histologic grade, ER expression, and PR expression), and treatment characteristics (primary therapy, tamoxifen prescription) to estimate the hazard ratio of breast cancer recurrence comparing those who received chemotherapy with those who did not.

Propensity score method—A propensity score is a summary confounder score that is modeled using the exposure as the dependent variable.^{14,28,29} Using logistic regression with the restricted cohort, we modeled the probability of receiving adjuvant chemotherapy as a function of the variables included in the multivariable adjusted model. To increase comparability between our index and reference groups, we trimmed the data to include only women with overlapping scores between the index and reference groups. With the trimmed dataset, we used Cox proportional hazards regression to model the association between adjuvant chemotherapy and recurrence, using three PS adjustment approaches. First we divided the trimmed sample into PS quintiles. We adjusted for PS quintiles and used the lowest quintile as the reference. Second, we adjusted for the continuous PS measure in the Cox proportional hazards model. Last, we used a doubly robust adjustment, in which we adjusted for the continuous PS and the variables used to predict the probabilities of receiving adjuvant chemotherapy.

As recommended by Sturmer *et al*,¹⁰ we evaluated the distribution of patient characteristics within the PS quintiles among women who received chemotherapy and women who did not. To assess whether our trimmed dataset differed from the restricted cohort, we performed multivariable adjustment on the trimmed dataset and compared these results to the results of the restricted cohort.

Instrumental variable-like method—The instrumental variable method has been used in analyses when confounding by indication is suspected.^{18,30-33} Specifically, the use of our IV-like approach was intended to control for the confounding by unmeasured indications for chemotherapy. Using an approach similar to Brookhart *et al*'s preference-based IV method,^{13,33} in the restricted cohort, we used each patient's surgeon's chronologically preceding patient's receipt of adjuvant chemotherapy (preceding patient within our dataset) as the IV

within strata of stage and ER expression to estimate the effect of receipt of adjuvant chemotherapy on time to breast cancer recurrence. We used a surgeon's preceding patient's receipt of adjuvant chemotherapy as a surrogate for a medical oncologist's preceding patient's prescription of adjuvant chemotherapy because we did not have information on each patient's medical oncologist (in addition, some patients did not see a medical oncologist). We assigned the IV by stratifying the dataset by surgeon. Within each surgeon, the data were sorted by the patient's date of diagnosis in chronological order. Patients of surgeons who only treated one participant in our dataset were excluded from the IV-like analysis. For surgeons with greater than one patient, the chronologically preceding patient's receipt of chemotherapy was assigned. The chronologically first patient for each surgeon was excluded so that each patient would have an IV defined.

IV-like estimation requires a two-step process. The first step used logistic regression to estimate the probability of receiving adjuvant chemotherapy given the preceding patient's receipt of adjuvant chemotherapy, and included patient characteristics (demographic, tumor, and treatment) in the model. The second step predicted time to recurrence from the probabilities calculated in the first step, using Cox proportional hazards regression and adjusting for patient characteristics.

Using a patient's surgeon's preceding patient's receipt of adjuvant chemotherapy as the IV, we relied on three key assumptions about the properties of the IV: (1) surgeon's previous patient's receipt of adjuvant chemotherapy was independent of the unmeasured risk factors in the current patient (IV not associated with confounders); (2) surgeon's previous patient's receipt of adjuvant chemotherapy was independent of the outcome in the current patient (IV had no direct effect on outcome); and (3) surgeon's previous patient's receipt of adjuvant chemotherapy varies within surgeons (IV associated with exposure).

Following methods outlined by Brookhart and Schneeweiss, we assessed the validity and interpretation of our estimate from our IV-like approach.¹³ The strength of the IV was estimated by performing simple linear regression with the IV as the independent variable and receipt of chemotherapy as the dependent variable in the model. We assessed the strength of our IV by comparing it with the strength reported by Brookhart and Schneeweiss.¹³

We used measured patient characteristics as proxies for unmeasured variables. To evaluate whether our IV assumptions were violated, we calculated the prevalence differences of patient characteristics between the levels of the IV and the prevalence differences of patient characteristics between the two levels of receiving chemotherapy. We assessed the imbalance of these characteristics by calculating prevalence difference ratios between the IV relative to receipt of chemotherapy. Prevalence difference ratios less than the null value of 1 indicated that the patient characteristics were more balanced across the levels of the IV than across the levels of the exposure. The prevalence difference ratios were compared with the strength of the IV. If the prevalence difference ratios were less than the strength of the IV, then the estimate for the association between adjuvant chemotherapy and recurrence using the IV-like method would result in a less biased estimate than using conventional methods.¹² Then we looked at the prevalence differences across the IV. For each characteristic, if the prevalence difference across the IV was not close to zero (no difference), then the IV remained confounded by that characteristic, and residual confounding could not be ruled out.

The widths of the 95% confidence intervals around the hazard ratios for each analytic method were calculated as the ratio of the upper limit to the lower limit. Larger widths were interpreted as having less precision.

We repeated each analytic method to assess whether the rate of recurrence varied by type of chemotherapy regimen.

Results

Frequencies for demographic and tumor characteristics for the unrestricted cohort who received primary therapy (n=1798), the cohort restricted to women at high-risk for recurrence (n=946), the propensity score analytic sample (n=723), and the instrumental variable analytic sample (n=539) are presented in Table 1 by receipt of chemotherapy. For women classified as high-risk, 20% experienced a breast cancer recurrence. In the unrestricted, restricted, PS, and IV samples, a higher proportion of women who received adjuvant chemotherapy were in the youngest age category (65-69 years), had a baseline Charlson score of 0, and were node positive, while a lower proportion were ER positive compared with those who did not receive adjuvant chemotherapy. These differences in distributions illustrate the potential for confounding by indication. Adjustment for tumor characteristics had the largest impact on the effect estimates. Node positivity had that highest magnitude of confounding of 1.7, followed by histology, tumor size, and ER status (each have a magnitude of confounding of about 1.3).

In the unrestricted cohort, receipt of adjuvant chemotherapy was crudely associated with recurrence (hazard ratio [HR]=2.6; 95% confidence interval [CI]=1.9, 3.5). After restricting the cohort to women at high-risk for recurrence, the hazard ratio relating recurrence to receipt of chemotherapy (HR=1.8 [95% CI=1.3, 2.5]) seemed to be confounded by indications for receipt of chemotherapy, presuming the prior based on clinical trials demonstrating a protective effect holds true in this population.⁶ We observed a modest increased hazard rate of breast cancer recurrence in women who received adjuvant chemotherapy compared with those who did not after multivariable regression (HR= 1.1; 95% CI=0.7, 1.6).

The propensity score distributions among women who received chemotherapy versus those who did not showed no substantial overlap (Figure 2), even after trimming the extreme probabilities of receiving (“All Exposed”) and not receiving chemotherapy (“All Unexposed”). Our PS trimmed sample consisted of 723 women at high-risk for recurrence. The crude estimate for the PS analytic sample was HR=1.7 (95% CI=1.2, 2.5). The PS quintile adjustment method yielded a slightly higher hazard ratio (HR=1.3; 95% CI=0.8, 2.0) than the multivariable regression method. Both the continuous and the doubly robust PS adjustment methods yielded a HR=1.1 (95% CI=0.7, 1.7). The multivariable adjusted association in the PS trimmed sample was similar to what we observed using the multivariable method on the restricted cohort (HR=1.1; 95% CI=0.7, 1.7).

For the IV-like method, to ensure that an instrument was assigned for each patient, 253 women were excluded because they were the only patient, in our dataset, seen by their surgeon or because they were the chronologically first patient, in our dataset, for their surgeon. The final analytic sample included 539 high-risk women. The crude estimate for the IV analytic sample was HR=2.1 (95% CI=0.1, 3.8). The IV adjusted estimate was HR=0.9 (95% CI=0.2, 4.3), but confounding was not completely controlled. Although all of our prevalence difference ratios were less than the strength of the IV of 23.7%, residual associations between the IV and several measured characteristics —such as histology, tumor size, and node positivity— remained (Table 2). Our prevalence difference ratios were both above and below the null, indicating that for some characteristics (age, comorbidity, tamoxifen prescription, ER expression, and PR expression) the IV was more balanced across levels of the characteristic than the observed exposure, but for others (race, tumor size, node positivity, histology, and primary therapy) the IV was less balanced than the observed exposure. For example, the imbalance in tumor size <1 cm was an absolute difference of 2.99 between those who received adjuvant chemotherapy and those who did not. The imbalance was reduced to 0.67 for the IV prevalence difference, resulting in a prevalence difference ratio of 0.22. Some of these characteristics are important prognostic markers for recurrence risk, so these residual associations portend the potential for residual confounding by indication. Figure 3 depicts the estimates and standard errors for the

association between adjuvant chemotherapy and breast cancer recurrence for the unadjusted and adjusted methods.

Among the women who received chemotherapy, 67% received a cyclophosphamide-methotextrate-flourouracil (CMF) regimen, 28% received an adriamycin-based regimen, and 4.8% were classified as having another regimen. Due to small numbers in chemotherapy subgroups, we could only examine the effect of CMF chemotherapy regimen on the rate of recurrence. Using the unrestricted, restricted, and PS methods, the results did not change appreciably, except that there were wider intervals around the estimates. Using the instrumental variable method the association between CMF and recurrence became slightly more protective (HR=0.6; 95% CI = 0.1, 3.8).

Discussion

The association between receipt of adjuvant chemotherapy and recurrence risk in older women with breast cancer provides a useful example of the manner in which confounding by indication can complicate non-randomized studies of treatments in general populations. When considering treatment recommendations to reduce breast cancer recurrence, oncologists treating geriatric patients take into account tumor prognostic factors and additional factors such as life expectancy, physical function, and quality of life.³⁴ With minimal trial-based information available to inform clinical guidelines, which currently offer no guidance for treating older women with cancer,³⁵ non-randomized studies are vitally important. However, non-randomized studies are only reliable when confounding by indication is handled adequately. When treatment with adjuvant chemotherapy among older patients is based on clinical judgment, controlling for prognostic factors alone leaves residual confounding by indication.

Although not intended to control for unmeasured confounding,^{10-12,14} propensity score adjustment has been implemented in studies for this reason;^{21,22} however consistent with other reports, our results suggest propensity scores do not provide any better control for unmeasured confounding than multivariable regression.^{10,11} Even after controlling for known prognostic factors, we obtained effect estimates in the causal direction, which would not be correct given the prior on the expected direction of effect, which is based on results from clinical trials in younger women.⁶ Our results using the IV-like approach yielded a slightly protective estimate of the association; however the imbalance of measured factors across levels of the IV indicated that our estimate remained confounded. Thus, no method of adjustment completely resolved this bias.

Selection bias and misclassification are unlikely explanations for our results. The potential for selection bias due to barriers to care was reduced by using an unselected sample of Medicare-insured women with complete data on treatment from integrated healthcare systems.²³ The inter-rater reliability of medical record abstraction was $\geq 90\%$ overall;³⁶ with 90% sensitivity and 96% specificity for breast cancer recurrence classification and 90% sensitivity and perfect specificity for receipt of chemotherapy.³⁶

Another possibility may be that the protective effect of adjuvant chemotherapy seen in younger women does not apply to early stage breast cancer in older women. The meta-analysis of 194 randomized controlled trials from 1985–2000 stratified by age yielded a similar finding as our IV-like result for women 70 years of age or older (recurrence rate ratio = 0.88) for 15-years of follow-up.⁶ Yet, only 3.6% of the 95,403 women participating in the polychemotherapy trials were in this age category.⁶ Thus, this meta-analysis finding should be interpreted with caution because geriatric women were underrepresented.^{3,4} It is likely that the women who enrolled

in these trials were healthier⁴ than the general elder patient population living with breast cancer.
37

We explored whether the effect of chemotherapy on recurrence varied by type of regimen. When we repeated the analyses restricting to chemotherapy exposure to CMF regimen, other than less precision for the estimates, the hazard ratios were nearly the same, except the instrumental variable estimate became slightly more protective. We could not perform subgroup analyses for the adriamycin-based regimen due to small numbers. However, in younger patients, for whom the data are adequate to assess the differences between adriamycin-based and non-adriamycin-based chemotherapy, the difference in recurrence between these types of chemotherapy is ~3% at 5 years after diagnosis.⁶

We explored potential explanations for our PS and IV findings. Our PS quintile adjustment suggested a stronger association among women who received adjuvant chemotherapy and recurrence than the other PS methods. Subjects were not evenly distributed between the quintiles, which was due to the inability of the PS quintile adjustment to discriminate scores between subjects with the same probability of exposure. Thus the majority of subjects fell into the lowest quintile (Q1), which may explain why the continuous and doubly robust PS adjustments yielded better control.

We assessed whether our PS findings could be explained by differences in patient characteristics between the PS trimmed sample and restricted cohort by comparing multivariable regression results of the two analytic samples. We found nearly identical results, indicating that the distributions were similar. Additionally, our PS adjustment results were nearly equivalent to those yielded by the multivariable method. Propensity scores are thought to be superior to multivariable regression models because they theoretically allow control for multiple measured confounders and increase comparability between the index and reference groups.^{10-12,14} However, in a review by Sturmer and colleagues, they found only 13% of 69 studies had multivariable adjusted results >20% different than results from adjusting for PS.¹⁰ Moreover, we found that even after trimming our dataset to exclude non-overlapping propensity scores, the distribution of the propensity scores among those who received chemotherapy (index) versus those who did not (reference) still lacked comparability. This finding suggests residual confounding, which we could not examine using conventional methods. As expected, the propensity score method did not rectify the confounding by indication in our study; it persisted in the cohort of high-risk patients even after adjusting for measured prognostic factors that are considered when prescribing adjuvant chemotherapy.

We compared our IV-like method to Brookhart and Schneeweiss' example of a preference-based IV method to provide a better understanding of the validity of our IV result. They studied approximately 50,000 subjects,³³ whereas after applying the exclusions required to implement our IV-like method, our analytic sample was 539 women. The IV acts as if we had randomized the exposure and like randomization, substantial departures in the data from the presumed balance of measured and unmeasured confounders is more likely in smaller studies.

More probable explanations may be violations of the IV assumptions, which Hernan and Robins have emphasized are unverifiable.¹⁷ We initially questioned the strength of our IV because an IV that is weakly associated with exposure can bias the estimate more than not adjusting at all.^{17,18,38,39} However, the strength of our IV was equivalent to the strength of the IV used by Brookhart and Schneeweiss (23%)¹³ and similar to the strength of other preference-based IVs that they have encountered (MA Brookhart, unpublished data, 2008).

We then evaluated whether our IV was independent of unmeasured risk factors. We assessed the plausibility of confounding by unmeasured factors by comparing the prevalence differences

of measured factors across levels of the IV. Imbalances remained among measured characteristics, suggesting that there may be clustering of patient risk factors within certain surgeons. Therefore, we cannot rule out that associations between important unmeasured factors and the IV may exist. The imbalance of measured patient characteristics across levels of the IV indicated that the IV is confounded. The IV estimate for the association between receipt of chemotherapy and breast cancer recurrence controlled for more confounding than the other methods, but did not completely resolve the bias.

The intervals around the estimates were wider using the PS method and IV method than conventional methods. The widths of the intervals (ratio of upper to lower limits) around the unrestricted, restricted, PS continuous and doubly robust, and PS quintile estimates were 1.8, 1.9, 2.4, and 2.5, respectively. The width of the interval around the IV estimate was substantially larger at 22. This demonstrates that our IV-like method was less statistically efficient than the conventional methods and, therefore, larger samples may be needed for IV methods to be feasible.

Alternative methods have been suggested to reduce confounding in observational studies, yet we found that conventional methods such as restriction and multivariable regression were as effective as the propensity score method. Our IV-like method was the only approach that yielded a protective association. However, we must be cautious in its interpretation because of the residual confounding in the distribution of measured factors across levels of the IV. The use of these alternative analytic methods to control for confounding by indication is not universal across all observational settings.^{10,13,16}

Non-randomized studies of therapy effectiveness will remain important contributions to our scientific knowledge base. Such studies will, however, remain susceptible to confounding by indication, despite advancing methods to control this seemingly intractable bias. Understanding the limitations and appropriateness of the propensity score and instrumental variable methods is an essential step before implementing and interpreting results that may be as biased as results generated by conventional methods.

Acknowledgments

The authors thank Dr. Alan Brookhart for his thoughtful review and guidance regarding the instrumental variable analysis. We would also like to thank Dr. Terry Field, site-principal investigator at Meyers Primary Care Institute: Fallon Community Health Plan/Fallon Foundation/University of Massachusetts Medical School.

Source of Financial support: Supported by Public Health Service Grant R01 CA093772 (Breast Cancer Treatment Effectiveness in Older Women, Rebecca A. Silliman, PI) from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

References

1. Giordano SH, Kuo YF, Duan Z, Hortobagyi GN, Freeman J, Goodwin JS. Limits of observational data in determining outcomes from cancer therapy. *Cancer* 2008;112:2456–2466. [PubMed: 18428196]
2. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med* 1994;331:394–398. [PubMed: 8028622]
3. Sørensen HT, Lash TL, Rothman KJ. Beyond randomized controlled trials: A critical comparison of trials with nonrandomized studies. *Hepatology* 2006;44:1075–1082. [PubMed: 17058242]
4. Avorn J. In defense of pharmacoepidemiology - embracing the yin and yang of drug research. *N Engl J Med* 2007;357:2219–2221. [PubMed: 18046025]
5. Gross CP, Steiner CA, Bass EB, Powe NR. Relation between prepublication release of clinical trial results and the practice of carotid endarterectomy. *JAMA* 2000;284:2886–2893. [PubMed: 11147985]
6. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* 2005;365:1687–1717. [PubMed: 15894097]

7. Lash TL, Silliman RA. Re: prevalence of cancer. *J Natl Cancer Inst* 1998;90:399–400. [PubMed: 9498492]
8. Geiger AM, Thwin SS, Lash TL, Buist DSM, Prout MN, Wei F, Field TS, Ulcickas Yood M, Frost FJ, Enger SM, Silliman RA. Recurrences and second primary breast cancers in older women with initial early-stage disease. *Cancer* 2007;109:966–974. [PubMed: 17243096]
9. Ulcickas Yood M, Owusu C, Buist DSM, Geiger AM, Field TS, Thwin SS, Lash TL, Prout MN, Frost FT, Enger SM, Silliman RA. Mortality impact of less-than-standard therapy in older breast cancer patients. *J Am Coll Surg* 2008;206:66–75. [PubMed: 18155570]
10. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006;59:437–447. [PubMed: 16632131]
11. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology* 2006;98:253–259. [PubMed: 16611199]
12. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;158:280–287. [PubMed: 12882951]
13. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat* 2007;3:Article 14
14. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
15. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;29:722–729. [PubMed: 10922351]
16. Terza JV, Bradford WD, Dismuke CE. The use of linear instrumental variables methods in health services research and health economics: a cautionary note. *Health Services Research* 2008;43:1102–1120. [PubMed: 18546544]
17. Hernan M, Robins JM. Instruments for causal inference. *Epidemiology* 2006;17:360–372. [PubMed: 16755261]
18. Martens E, Pestman W, de Boer A, Belitser S, Klungel O. Instrumental variables: application and limitations. *Epidemiology* 2006;17:260–267. [PubMed: 16617274]
19. Warren JL, Klabunde CN, Schrag D, Bach P. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care* 2002;40 (suppl):IV3–IV18.
20. Giordano SH, Duan Z, Kuo YF, Hortobagyi GN, Goodwin JS. Use and outcomes of adjuvant chemotherapy in older women with breast cancer. *J Clin Oncol* 2006;24:2750–2756. [PubMed: 16782915]
21. Elkin EB, Hurria A, Mitra N, Schrag D, Panageas KS. Adjuvant chemotherapy and survival in older women with hormone receptor-negative breast cancer: assessing outcome in a population-based, observational cohort. *J Clin Oncol* 2006;24:2757–2764. [PubMed: 16782916]
22. Du XL, Jones DV, Zhang D. Effectiveness of adjuvant chemotherapy for node-positive operable breast cancer in older women. *J Gerontol A Biol Sci Med Sci* 2005;60:1137–1144. [PubMed: 16183952]
23. Enger SM, Thwin SS, Buist DSM, Field T, Frost F, Geiger AM, Lash TL, Prout M, Yood MU, Wei F, Silliman RA. Breast cancer treatment of older women in integrated health care settings. *J Clin Oncol* 2006;24:4377–4383. [PubMed: 16983106]
24. Wagner EH, Greene SM, Hart G, Field TS, Fletcher S, Geiger AM, Herrinton LJ, Hornbrook MC, Johnson CC, Mouchawar J, Rolnick SJ, Stevens VJ, Taplin SH, Tolsma D, Vogt TM. Building a research consortium of large health systems: the cancer research network. *J Natl Cancer Inst Monogr* 2005;2005:3–11. [PubMed: 16287880]
25. Glick JH, Gelber RD, Goldhirsch A, Senn HJ. Meeting highlights: adjuvant therapy for primary breast cancer. *J Natl Cancer Inst* 1992;84:1479–1485. [PubMed: 1433332]

26. Thwin S, Clough-Gorr K, McCarty M, Lash T, Alford S, Buist D, Enger S, Field T, Frost F, Wei F, Silliman R. Automated inter-rater reliability assessment and electronic data collection in a multi-center breast cancer study. *BMC Medical Research Methodology* 2007;7:23. [PubMed: 17577410]
27. Charlson M, Pompei P, Ales K, MacKenzie C. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373–383. [PubMed: 3558716]
28. Sturmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol* 2005;161:891–898. [PubMed: 15840622]
29. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127:757–763. [PubMed: 9382394]
30. Kahn KL, Tisnado DM, Adams JL, Liu H, Chen WP, Hu FA, Mangione CM, Hays RD, Damberg CL. Does ambulatory process of care predict health-related quality of life outcomes for patients with chronic disease? *Health Services Research* 2007;42:63–83. [PubMed: 17355582]
31. Schmoor C, Caputo A, Schumacher M. Evidence from nonrandomized studies: a case study on the estimation of causal effects. *Am J Epidemiol* 2008;167:1120–1129. [PubMed: 18334500]
32. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. 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* 2007;297:278–285. [PubMed: 17227979]
33. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* 2006;17:268–275. [PubMed: 16617275]
34. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. *N Engl J Med* 1998;338:1516–1520. [PubMed: 9593791]
35. NIH consensus conference. Treatment of early-stage breast cancer. *JAMA* 1991;265:391–395. [PubMed: 1984541]
36. Lash TL, Fox MP, Thwin SS, Geiger AM, Buist DSM, Wei F, Field TS, Yood MU, Frost FJ, Quinn VP, Prout MN, Silliman RA. Using probabilistic corrections to account for abstractor agreement in medical record reviews. *Am J Epidemiol* 2007;165:1454–1461. [PubMed: 17406006]
37. McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 1999;319:312–315. [PubMed: 10426754]
38. Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica* 1997;65:557–586.
39. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *J Am Stat Assoc* 1995;90:443–450.

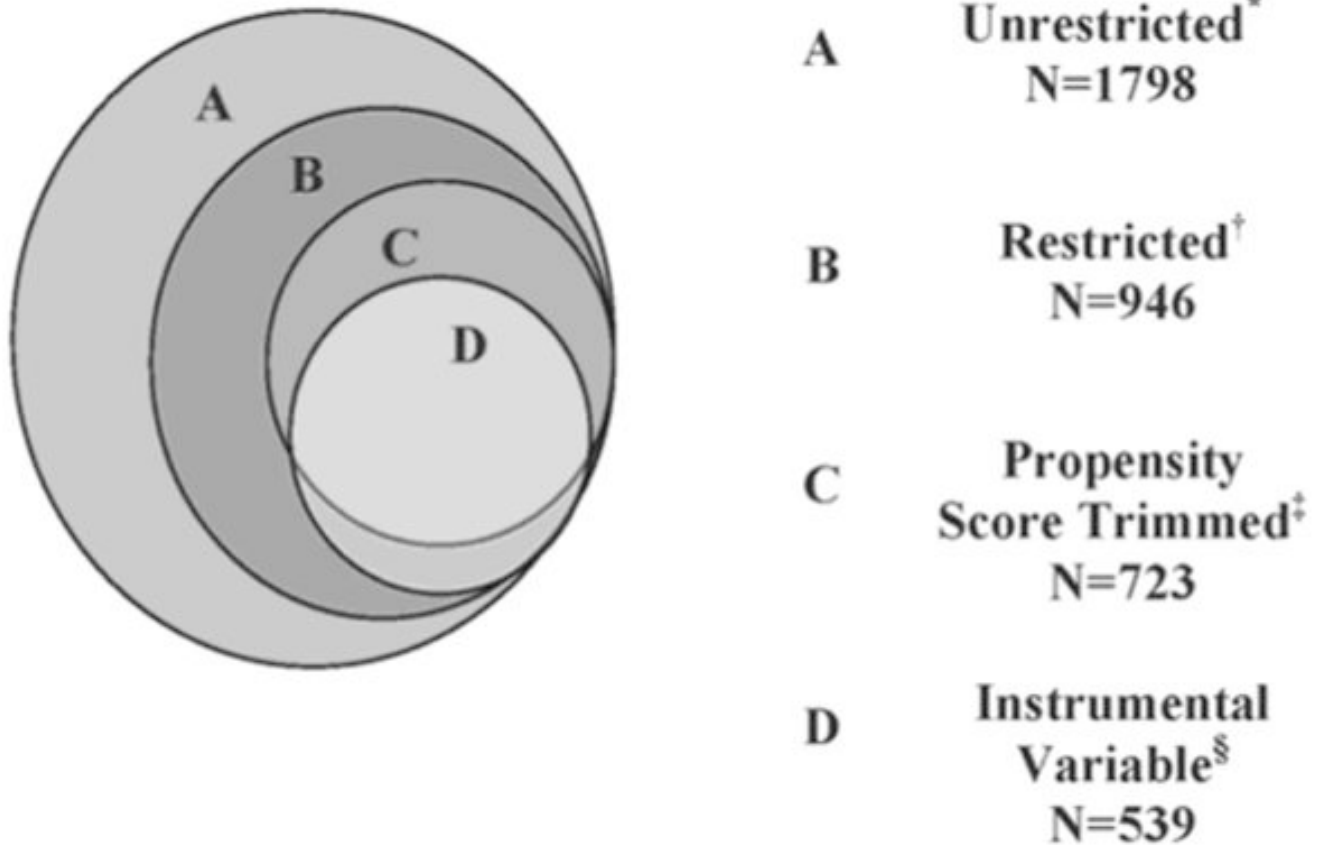


Figure 1. Venn diagram of analytic sample sizes for each adjustment method used to control for confounding by indication in a study of older women with breast cancer

* BOW cohort excluding women with biopsy only (n=22), neoadjuvant chemotherapy (N=3); inconsistent chemotherapy dates (n=13), and women who had a recurrence (n=16) or died (n=6) before the start of adjuvant chemotherapy or 183 days after date of diagnosis.

† Restricted to women who are classified as high-risk for recurrence by the 1992 St. Gallen Criteria.²⁵

‡ Trimmed the sample to exclude women who did not have overlapping propensity scores.

§ Excludes 253 women because they were the only patient seen by their surgeon or because they were the chronologically first patient for their surgeon, in our dataset.

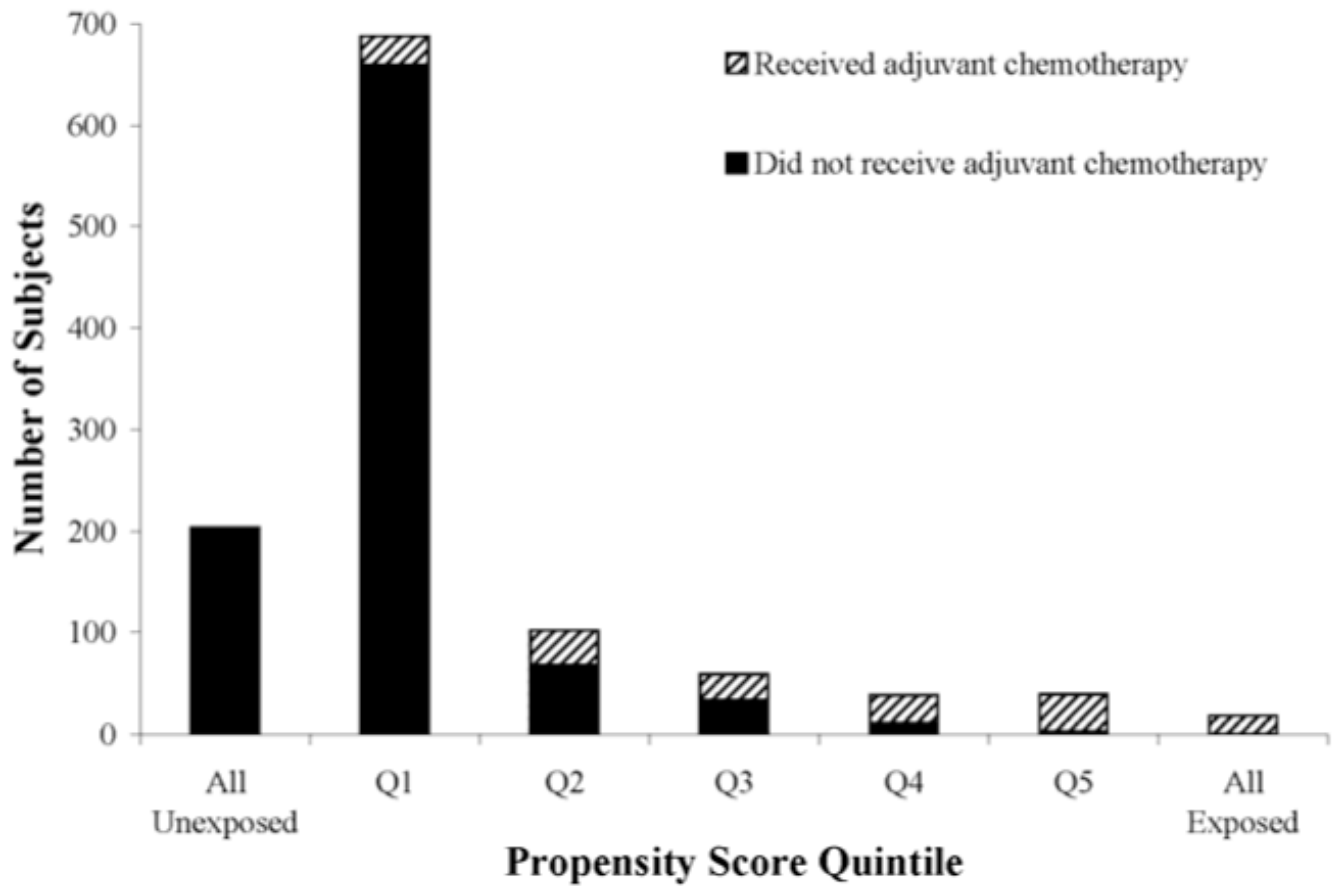


Figure 2. Propensity score distribution for adjuvant chemotherapy in older women with breast cancer by quintile. The propensity score analytic sample trimmed the “All exposed” and “All unexposed” categories

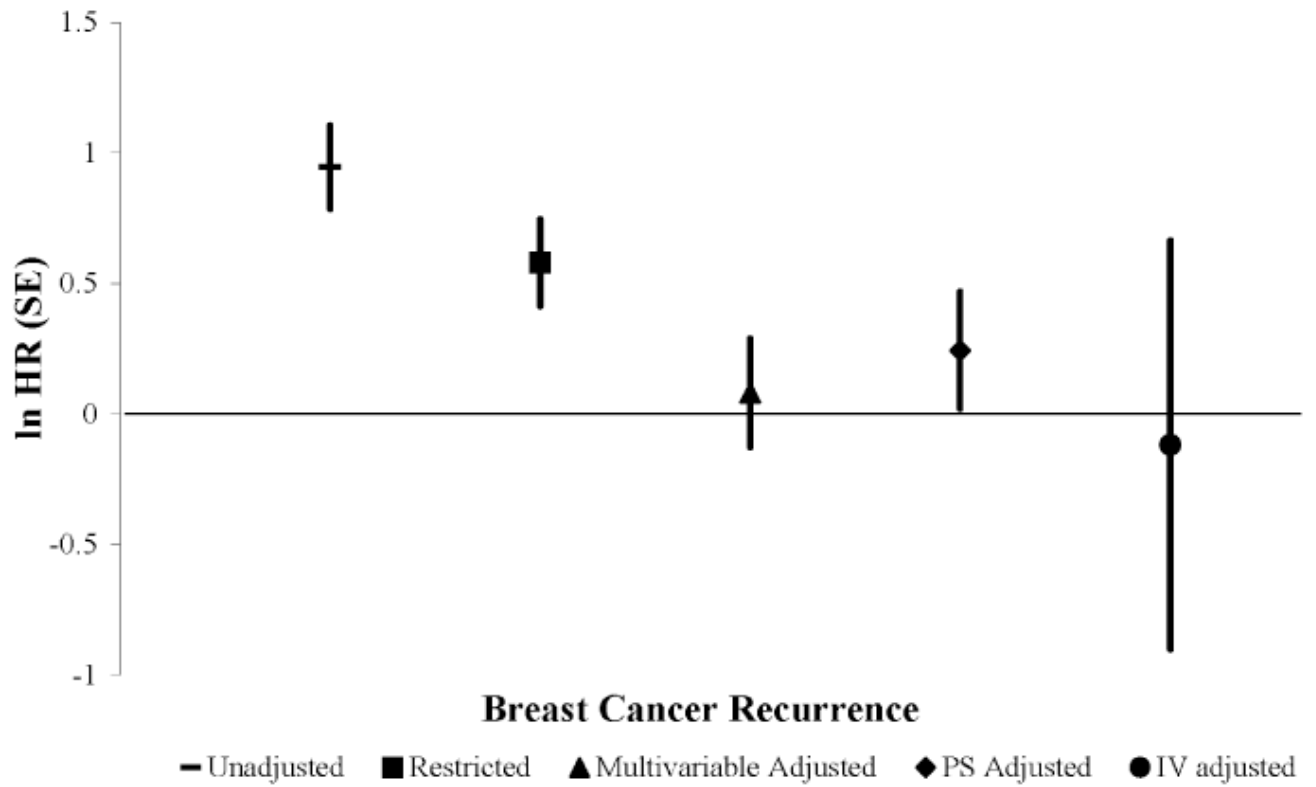


Figure 3. Estimates and standard errors for the association between adjuvant chemotherapy and rate of breast cancer recurrence in older women

Table 1
Demographic, tumor, and treatment characteristics in the subjects for the unrestricted cohort, propensity score analytic sample, and instrumental variable analytic sample

Characteristic	Unrestricted Cohort* (n=1798)		Restricted Cohort† (n=946)		Propensity Score Cohort‡ (n=723)		Instrumental Variable Cohort§ (n=539)	
	Chemo N%	No Chemo N%	Chemo N%	No Chemo N%	Chemo N%	No Chemo N%	Chemo N%	No Chemo N%
Age categories								
65-69	10663	50631	9963	25632	8560	24442	6066	14332
70-74	4728	48430	4730	23730	4431	19934	2527	14532
75-79	116.6	29118	95.7	14819	96.4	9116	44.4	8118
80+	31.8	35122	31.9	14819	32.1	488.3	22.2	7918
Race / Ethnicity								
Non-Hispanic White	14386	132581	13787	65183	12186	47481	7987	37985
Hispanic or Other Race	2414	30219	2113	13317	2014	10418	1213	6515
Charlson Comorbidity Index								
0	13480	109367	12680	53167	11279	42874	7684	29967
1	3219	45428	3120	21627	2820	14625	1415	12428
2+	10.6	855.2	10.6	425.3	10.7	81.4	11.1	255.6
Tumor Size								
<1 cm	116.6	36522	106.3	658.2	96.4	528.9	55.5	388.5
1 - <2 cm	5231	69342	4931	25432	4331	19634	3134	13029
2 - <3 cm	4728	36823	4428	27635	4028	19634	1921	16236
3+ cm	5734	20613	5535	19425	4935	13824	3640	11826
Node positivity								
None	4326	130880	4025	47260	4028	29050	2123	25557
1 to 3 nodes	6438	26216	6038	25532	5539	23140	3538	15535
4+ nodes	6036	623.8	5837	627.9	4633	6110	3538	388.5
Not determined	10.6	33621	10.6	11014	10.7	284.8	11.1	6414
Histologic Grade								
Well differentiated	148.4	27917	138.2	668.4	139.2	559.5	99.9	4510
Intermediate/moderate	4929	61538	4730	24030	4431	18432	2527	14131
Poorly differentiated/undifferentiated/anaplastic	8953	32620	8554	31640	7261	24141	5358	16938
Not determined/ stated	159.0	41225	138.2	16721	128.5	10218	44.4	9321
Estrogen Receptor Expression								
Positive	7143	122775	6843	56471	6848	40469	4752	35679
Negative	9054	19212	8554	16421	6949	14024	4044	5512
Indeterminate/Other	63.6	21313	53.2	617.7	42.8	386.5	44.4	378.3
Progesterone Receptor Expression								
Positive	5734	98060	5434	45057	5338	32155	3336	27461
Negative	10261	38524	9761	25132	8258	20635	5358	12528
Indeterminate/Other	84.8	25716	74.4	8811	64.3	559.5	55.5	4911
Primary Therapy								
Breast Conserving Surgery only (BCS)	31.8	21113	21.3	759.5	21.4	172.9	11.1	4911
BCS plus Radiation Therapy	4024	59036	4025	20926	3827	17129	2325	12628
Mastectomy	12474	83151	11673	50564	10172	39468	6774	27361
Tamoxifen prescribed								
Yes	10261	108066	9761	62479	9467	45478	6369	37383
No	6539	55734	6139	16521	4733	12822	2831	7517
Chemotherapy regimen								
CMF**	11267	--	10969	--	9870	--	6066	--
Adriamycin-based	4728	--	4227	--	3626	--	2730	--
Other	84.8	--	74.4	--	75.0	--	44.4	--
Chemotherapy completed								
CMF	9657	--	9359	--	8460	--	5358	--

Characteristic	Unrestricted Cohort* (n=1798)		Restricted Cohort† (n=946)		Propensity Score Cohort‡ (n=723)		Instrumental Variable Cohort§ (n=539)	
	Chemo N%	No Chemo N%	Chemo N%	No Chemo N%	Chemo N%	No Chemo N%	Chemo N%	No Chemo N%
Adriamycin- based	4426	--	3925	--	3424	--	2426	--
Other	63.6	--	63.8	--	64.3	--	44.4	--
Not complete	2113		2013		1712		1011	
Diagnosed with Recurrence								
Yes	4929	19112	4931	13817	4230	10718	3134	7817
No	11871	144188	10969	65183	9970	47582	6066	37083

* Excludes: 22 biopsy only; 3 neoadjuvant chemotherapy; 13 inconsistent chemotherapy dates; 16 recurrences and 6 deaths before the start of adjuvant chemotherapy or 183 days after date of diagnosis.

† From 1992 St. Gallen Criteria

‡ Propensity score analytic sample after trimming the non-overlapping scores between the index and reference groups.

§ Instrumental variable analytic sample after excluding patients of surgeons who only treated one patient and the chronologically first patient.

** CMF = Cyclophosphamide-methotrexate-flourouracil

Table 2
Assessment of imbalance of measured patient characteristics across levels of instrumental variable and exposure (adjuvant chemotherapy) and prevalence difference ratios

Patient characteristics	Prevalence Difference across levels of instrument [*]	Prevalence Difference across levels of exposure [†]	Prevalence Difference Ratio [‡]
Age categories			
65-69	18.16	34.01	0.53
70-74	-2.25	-4.90	0.46
75-79	-8.39	-13.68	0.61
80+	-7.52	-15.43	0.49
Race / Ethnicity			
Non-Hispanic White	3.15	2.21	1.43
Hispanic or Other Race	-2.28	-1.32	1.73
Charlson Comorbidity Index			
0	9.66	19.78	0.49
1	-6.89	-12.30	0.56
2+	-2.77	-4.48	0.62
Tumor Size			
< 1 cm	-0.67	-2.99	0.22
1 to < 2 cm	-6.1	5.05	1.21
2 to < 3 cm	4.07	-15.28	0.27
3+ cm	2.69	13.22	0.20
Node Positivity			
None or Not determined	-15.22	-33.84	0.45
1 to 3 nodes	5.01	3.86	1.30
4+ nodes	10.21	29.98	0.34
Histologic Grade			
Well differentiated	-1.62	-0.15	10.80
Intermediate/moderate	18.96	-4.00	4.74
Poorly differentiated/undifferentiated/ anaplastic	-6.34	20.52	0.31
Not determined/ stated	-11.01	-16.36	0.67
Estrogen Receptor Expression			
Positive	-24.3	-31.68	0.72
Negative	24.3	31.68	0.77
Progesterone Receptor Expression			
Positive	-22.16	-30.34	0.73
Negative	22.16	30.34	0.73
Primary Therapy			
BCS only	-8.01	-9.84	0.81
BCS plus Radiation Therapy	-6.38	-2.86	2.23
Mastectomy	14.39	12.69	1.13
Tamoxifen Prescribed			
Yes	-7.03	-14.03	0.57
No	7.03	14.03	0.57

* Prevalence of being assigned the index condition for the instrument minus the prevalence of being assigned the reference condition for the instrument.

† Prevalence of having the index condition (receiving adjuvant chemotherapy) minus the prevalence of not having the reference condition (not receiving adjuvant chemotherapy).

‡ Prevalence difference for the instrument divided by the prevalence difference for the exposure.